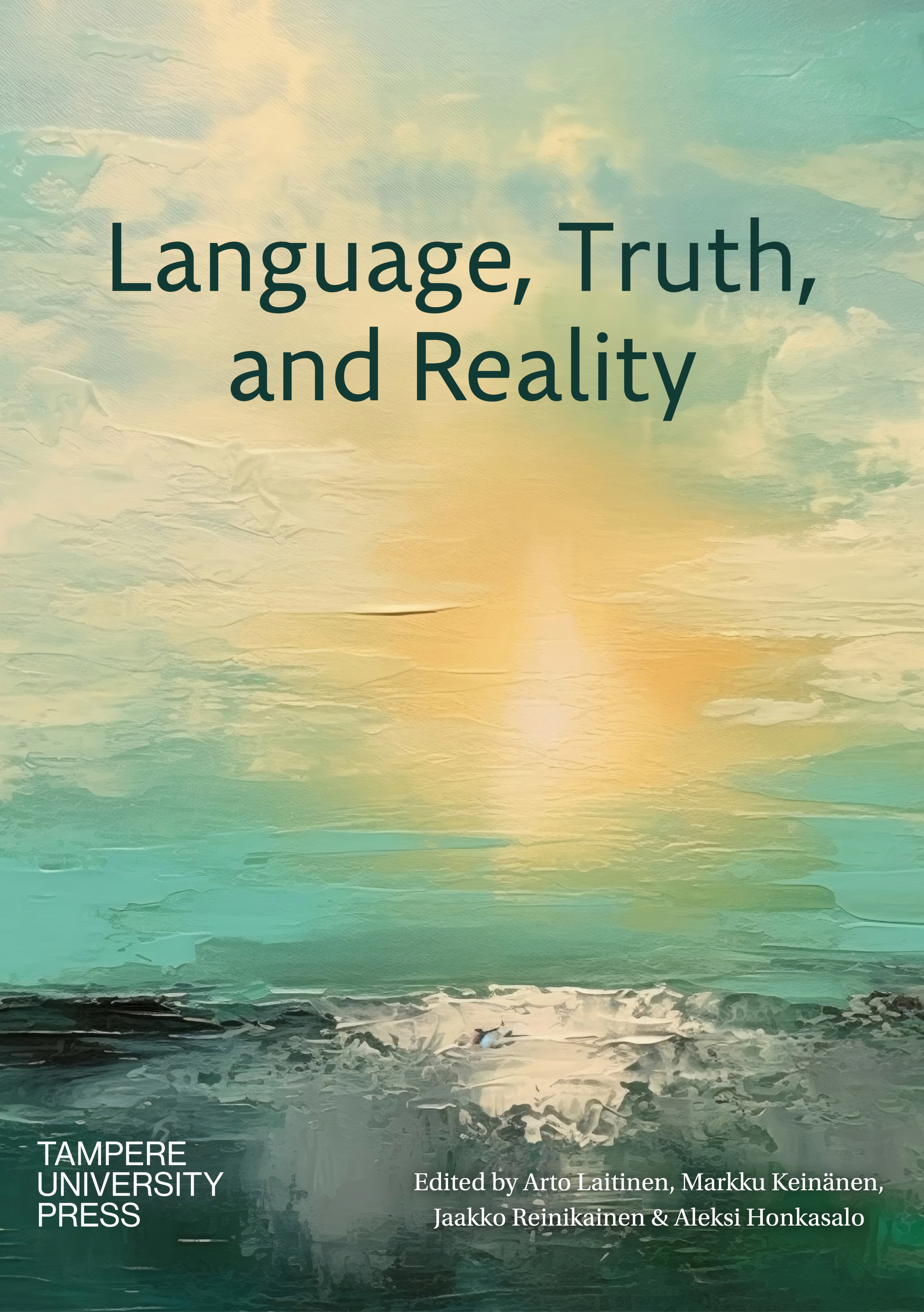
Language, Truth, and Reality



Edited by Arto Laitinen, Markku Keinänen, Jaakko Reinikainen & Aleksi Honkasalo

TAMPERE
UNIVERSITY
PRESS

Language, Truth, and Reality

Language, Truth, and Reality

Philosophical essays in honour of Panu Raatikainen

Edited by
Arto Laitinen, Markku Keinänen, Jaakko Reinikainen & Aleksi Honkasalo

TAMPERE
UNIVERSITY
PRESS

Tampere University Press
https://edition.fi/tup
https://isni.org/isni/0000000507168743

VERTAISARVIOITU
KOLLEGIALT GRANSKAD
PEER-REVIEWED
www.tsv.fi/tunnus

Typography by
Markus Itkonen

Layout by
Eija Kylmäniemi

Cover design by
Jonna Lahti

ISBN 978-952-359-072-4 (PDF)
ISBN 978-952-359-073-1 (print, paperback)

Kustantaja: Tampere University Press, Tampere, Suomi
Painopaikka: Libri Plureos GmbH, Hamburg, Germany

Suggested citation:
Laitinen, A., Keinänen, M., Reinikainen, J. & Honkasalo, A. (eds.) (2025). *Language, Truth, and Reality*. Tampere University Press. https://doi.org/10.61201/h5fbdk64

Contents

Part I
Philosophy and Truth

Part II
Language

Part III
Reality

List of figures

1
Language, truth, and reality
An introduction

Arto Laitinen, Markku Keinänen, Jaakko Reinikainen & Aleksi Honkasalo

Let us start with a quote from Timothy Williamson (Chapter 2):

> “I have known Panu Raatikainen and his work since the last millennium. He has always struck me as a force for both sanity and clarity, two quite different virtues. The first time I heard him give a paper, when I was still a professor at Edinburgh, he deftly used the basic distinction between theories and languages to diagnose a fatal confusion in some famous arguments about the philosophy of language, cutting through the technicalities to expose the underlying philosophical error. Again and again, Panu deflates pretentious overblown claims and looks beneath slick formulations to see what they conceal. He does so in debates where formal logical considerations loom large, and mathematical prowess can easily be mistaken for philosophical insight; because he understands the mathematics so well, he is able to show what it does *not* imply about the problem at issue. Such work is vital to keeping philosophy honest and on the right track. Long may Panu continue to contribute in his distinctive way!”

These words by Timothy Williamson provide a fitting motto for this collection. The collection brings together leading philosophers who have encountered Panu Raatikainen in various academic occasions, as well as past and present colleagues. Its purpose is to celebrate Panu’s 60th birthday with a mosaic of papers discussing

Laitinen, A., Keinänen, M., Reinikainen, J. & Honkasalo, A. (2025). Language, truth, and reality: An introduction. In A. Laitinen, M. Keinänen, J. Reinikainen & A. Honkasalo (eds.) *Language, Truth, and Reality*. Tampere University Press. (pp. 9–16). https://doi.org/10.61201/h5fbdk64

various topics handpicked from his long and prosperous career in philosophy.

Panu Anssi Kalevi Raatikainen, native to Kajaani, defended his doctoral dissertation *Complexity, Information, and Incompleteness* at the University of Helsinki in 1998. In 2014 he got a tenure-track position in the University of Tampere, which in 2021 grew into full professorship. He is a Docent in Helsinki (2001) and in Tampere (2013) and was a Visiting Researcher in the Universities of St. Andrews (2004), New York (2005), and London (2006). He has held numerous academic administrative, leadership, and supervisory positions, engaged regularly in teaching and peer-review work, and supervised MAs and PhDs. His research output exceeds over a hundred publications. Beyond academia, Panu is a well-known, ardent commentator and public discussant in the Finnish affairs of science, politics, and culture.

Panu's academic work in philosophy ranges from the philosophy of mind and language to truth and science, logic and mathematics. As aptly described above, his signature contribution in any area is bringing argumentative transparency to foggy and complex debates. A perfect example is Panu's long-lasting advocacy of semantic externalism, and in particular the causal-historical theory of reference inaugurated by Saul Kripke. Here, he has contributed several influential articles over the years, specialised in pointing out the distorted or mistaken views that often underlie the criticism of Kripke. Another, completely different area where Panu enjoys indubitable expertise is the literature on Gödel's incompleteness theorems, where he is the author of the respective entry in the famous Stanford Encyclopedia of Philosophy. A third, yet distinct debate he has left his mark on is the philosophy of causation, in particular mental causation, where he has critiqued the so-called exclusion argument advanced by Jaegwon Kim. Other areas that he has been involved in include the question of free will, various issues in the philosophy of science, and the history of analytic philosophy.

In addition to his internationally recognized contributions to philosophy, Panu has also published significant amount of work in Finnish. These include work aimed for academic philosophers, as well as work intended for wider audiences, such as *Ihmistieteet ja Filosofia* (*Philosophy and the Human Sciences*), which has been widely used in Finland as an undergraduate textbook for various social sciences and humanities.

Panu's most recent work in the philosophy of mind continues to highlight his aptness for critical thinking. For example, in a paper presented both at the 7th Parma Workshop on Semantics and Pragmatics (2024) and at the annual Colloquium of the Philosophical Society of Finland (2025), he presents a new, fatal problem for David Chalmers' influential two-dimensional framework for meaning and reference.

As Panu's work has in many ways touched on the nature of Truth, Language, and Reality, we were happy to notice that no book with that title has been published—our search only returned book sections. We were pleased and honoured to get contributions from eminent philosophers, from Panu's supervisees, and past and present partners in crime.

**

Let us next take a short look at the chapters of the volume. Part One contains essays on truth and on philosophy (Chapters Two to Seven).

Timothy Williamson takes a sober bird's-eye view at the various senses of naturalism currently circulating in analytic philosophy in Chapter Two, 'On "naturalism"'. Williamson finds that it is often too easy to both overstate and understate the relevance of natural and non-natural sciences to one's metaphilosophical views, showcasing this with select examples.

Anssi Korhonen examines *what analytic philosophy is* and *what it has been*, in Chapter Three, "In search of analytic philosophy." He starts from two points: first, analytic philosophy is a genuine and distinctly recognizable *philosophical* and *historical* phenomenon, whose identity differs from that of, say, phenomenology and the phenomenological tradition. Second, analytic philosophy is a key tradition in twentieth century Western thought that has existed for at least one hundred and twenty years, maybe over a hundred and forty years, defined ostensively by Sluga (1997, 17fn.):

> "Following common practice, I take analytic philosophy here as originating in the work of Frege, Russell, Moore, and Wittgenstein, as encompassing the logical empiricism of the Vienna Circle, English ordinary language philosophy of the post-war period, American mainstream philosophy of recent decades, as well as other worldwide affiliates and descendants."

Korhonen compares Panu Raatikainen's definition of analytic philosophy to those of others and ultimately concludes that *no* feasible analytic definition can be given. All such proposals, including Raatikainen's sophisticated revisionary definition, inevitably misrepresent the nature of the phenomenon under consideration. Instead, the problem should be tackled *historically*: researchers should stick to 'analytic philosophy' as a lexical item, thus obtaining a pre-theoretical and agreed-upon notion of *analytic tradition* as a starting-point.

Juhani Yli-Vakkuri and Zachary Goodsell engage with Alfred Tarski's theory of truth in Chapter Four, "A categorical theory of truth". In his classic 1933 paper "On the concept of truth in formalized languages", Tarski proposes his famous Convention T as a 'criterion of adequacy' for a definition of truth. There is widespread dissatisfaction with Convention T in the literature, beginning with Tarski's own paper, as well as with the very idea of a "criterion of adequacy"—as opposed to a criterion of correctness—for definitions of truth. A criterion of correctness would take the form of (in Tarski's terms) a categorical theory of truth for the object language, from which the definition could be derived. (The term categorical is understood here in a proof theoretical sense rather than the now more common model-theoretical sense.) Yli-Vakkuri and Goodsell propose that Convention T should be discarded and replaced by the categorical theory of truth they show how to construct, from which Tarski's definition of truth can be derived. Thus, Tarski's truth definitions are vindicated, while his (by his own lights) unsatisfying Convention T can be forgotten.

In Chapter Five, "Putnam's transcendental arguments", Sami Pihlström tackles a long-running debate between Panu Raatikainen, Hilary Putnam, Ilkka Niiniluoto and Pihlström himself over language and metaphysical realism. The exposition provides both a useful overview of the topic and argues that Putnam's later position can usefully be compared to Kant's transcendental idealism.

In Chapter Six, Markus Lammenranta defends a form of Academic skepticism that denies the possibility of knowledge about the external world. The standard argument for it relies on internalism and infallibilism, doctrines that were widely accepted in the history of epistemology until the late 20th century. Contemporary epistemologists typically deny at least one of them, because together they lead to skepticism. Skepticism is thought to be bad because it conflicts with common sense, our ordinary epistemic practices, and linguistic data. Lammenranta argues that this is not so, that Academic skepticism gives in fact a better explanation of our intuitions and linguistic data than dogmatic epistemology. Finally, following the steps of Arcesilaus, Carneades and Hume, he shows how Academic skepticism can give a good response to the Stoics' Apraxia objection that skepticism makes rational action and good life impossible. On the contrary, it is skepticism that makes a good and flourishing life possible.

In Chapter Seven, Inkeri Koskinen examines whether the argument from inductive risk is just research ethics. The argument from inductive risk (AIR) is one of the most influential arguments against the value-free ideal of science. The value-free ideal (VFI) states that "while non-epistemic values can legitimately influence the "external aspects" of science, such as the choice of research projects, only epistemic values – that is, values that promote the attainment of truth – have a legitimate role in the central stages of scientific research, especially in the assessment of evidence and the justification of findings." Researchers, however, have the "responsibility to consider the predictable, non-epistemic consequences of any errors they make in their research: a scientist, as a scientist, has no special license to recklessly or negligently risk others.". And AIR concludes: "Researchers face inductive risks throughout the research process. Therefore, non-epistemic values must also influence the internal stages of the process." Overall, Koskinen defends the view that the argument from inductive risk is at heart a (research) ethical one, but that it shows that value-freedom is untenable as an ideal. It starts by introducing the value-free ideal and the argument from inductive risk and then argues that ideals ought to be such that they can guide action. Finally, Koskinen argues that the argument from inductive risk does not just point out some constraints to our ability to follow the ideal of value-freedom but shows that it is undesirable as an ideal.

Part Two contains essays on *language* (Chapters Eight to Fourteen).

Michael Devitt argues in Chapter Eight, "Quantifier phrases with referential meanings?" that not only definite but also indefinite descriptions are semantically ambiguous, allowing for both conventional attributive and referential uses. This goes against Gricean strategies that seek to eliminate such ambiguity by an appeal to general pragmatic principles. Devitt also argues that examples provided by Mario

Gómez-Torrente for referential uses of quantifier phrases do not in fact fit the bill, although some other quantifier phrases do.

Genoveva Martí starts Chapter Nine, “No-content explanations” by noting that typically, explanations of semantic and cognitive phenomena are given by appeal to content. She argues that we can find in the philosophical literature some good explanations of semantic and cognitive phenomena that are not content-oriented. These include Wettstein on cognitive value; Donnellan on empty names and Perry's first papers on indexicals. The only reason to not accept their satisfactoriness is the insistence in clinging to the assumption of the primacy of content. Martí argues, however, that if propositional content is conceived heuristically, as a convenient tool, it may have a useful theoretical role. If content is relieved from its position as the unique tool with explanatory power, content may have, after all, a legitimate theoretical role in contributing, partially, to *some* explanations.

Pasi Valtonen examines, in Chapter Ten, the meaning of absurdity in the context of logical inferentialism and Carnap’s problem. According to Panu Raatikainen, logical inferentialism cannot solve Carnap’s problem, unlike its model-theoretic rival. In their reply, Julien Murzi and Ole Thomassen Hjortland show that intuitionistic inferentialists like Dummett and Prawitz can handle the problem but remain sceptical about a classical inferentialist solution. Valtonen reveals some problems with their solution due to Prawitz’s and Dummett’s view of absurdity. He offers a Tennant-style paraconsistent view of absurdity. It not only solves the exposed problems in the intuitionistic solution but also contributes to the classical inferentialist solution.

Jaakko Reinikainen defends in Chapter Eleven, “Questions of reference”, a piece of conventional wisdom—that descriptivism fails—with conventional arguments—namely, from incompleteness and redundancy—against a recent case made by Jens Kipper and Zeynep Soysal. He draws centrally from recent work by Panu Raatikainen, (2020) “Theories of Reference: What Was the Question?” with the overarching aim to show that many of Kipper and Soysal’s arguments can be met with answers already provided by Raatikainen.

Aleksi Honkasalo examines, in Chapter Twelve, the relationship between Carnapian explication, and modern conceptual engineering. It is now commonly recognised that conceptual engineering has its roots in Carnapian explication, in which vague prescientific concepts are refined into exact scientific concepts. However, whereas modern conceptual engineering is almost universally understood as a normative endeavour—instead of asking what concepts are, it ask what concepts should be—for Carnap language has “no morals”, and thus one is free to choose their language as one sees fit. Carnap’s liberal approach towards language could suggest that normativity is what differentiates modern conceptual engineering from Carnapian explication. Against this, Honkasalo suggests that there is room for normativity in Carnapian explication. First, he argues that weak means-to-end normativity is essential for understanding both the explication and conceptual engineering. Secondly, he argues that, if the ends of explication are worth pursuing, explication can be seen as a strongly normative practice.

In Chapter 13, "Theories of reference: what really is the question?" Jaakko Kuorikoski focuses on Panu Raatikainen's view that the main question in theories of reference is: *In virtue of what* does a referring expression refer to whatever it *in fact refers to?* He argues that the two italicized points are in need of clarification. "What is the nature of the 'in virtue of' relation and what is the nature of the putative 'fact' of referring? What kind of an explanation is the theory of reference supposed to provide and what kind of a phenomenon is it that we are trying to explain?" Kuorikoski argues for a naturalist view that a theory of reference is, in fact, "a highly stylized model of data in a verbal form. A data model is a representation of data, which highlights some selected systematic features of the data in a cognitively salient manner. In the case of theories of reference, the primary data are the semantic intuitions, understood very liberally."

In Chapter 14, concluding Part Two of the book, Mikko Yrjönsuuri takes a fresh look at the topic of universal language, focussing on Ockham's theory and its relevance for contemporary thought. How ideal and universal is Ockham's mental language? Is there a definite answer? "Scholars did find interesting similarities in Ockham when Chomsky and Fodor had success in claiming universality in human and mental languages. But as the success of the latter waned, it was realized that Ockham's idea wasn't quite the same either. It may be so for all history of philosophy. Questions change. Thus, every generation must find its own answers to what exactly the past philosophers were trying to do. It is best to find them in a way that is helpful to one's own contemporaries rather than trying to uncover some eternal philosophical truths."

Part Three of the collection contains essays about realism and aspects of reality (Chapters 15 to 22).

In Chapter 15, Jani Hakkarainen discusses the rehabilitation of ontology and metaphysics in the 20th century, the origin of which is commonly traced back to Quine's "On what there is". In the article, Quine presents metaphysics primarily as an ontology. At the same time, he gives "ontology" a slightly new meaning: the task of ontology is to account for the various entities we assume to exist when we take certain propositions to be true. Hakkarainen notes, however, that Quine was not the first 20th century philosopher to rehabilitate ontology as a legitimate field of philosophy. In outline, in addition to the Quinean conception of ontology, Hakkarainen presents five different senses of "ontology", without claiming that these six senses constitute an all-encompassing list, "everything".

Ilkka Niiniluoto challenges Hasok Chang's Pragmatic realism in Chapter 16, "Ten queries about Hasok Chang's pragmatic realism". He shares Panu Raatikainen's (2004, 2014) defence of critical realism. Niiniluoto poses no less than ten challenges to a Neo-Pragmatic account of truth and realism, as presented in Hasok Chang's *Realism for realistic people: a new pragmatist philosophy of science* (2022).

In Chapter 17, Arto Laitinen and David P. Schweikard examine realism in social ontology. Realism is trending in recent scholarship in social ontology, but as Raatikainen (2014) among others has shown, realism means many different things.

In social ontology, even social constructionism, formerly taken to be a decidedly anti-realist view, is now prominently regarded as a form of realism. But what exactly does this mean? In what sense are social constructionists in social ontology realists in this domain? And, more broadly, what kind of realism would be plausible to adopt in social ontology? In discussing these questions, the chapter pursues three aims: First, it locates the question of realism by distinguishing between substantive questions about the reality of social phenomena and the meta-debate about defining realism in this domain. Second, it clarifies what it means to adopt realism in social ontology by providing a basic map of realisms and anti-realisms. Building on received taxonomies and terminology, Laitinen and Schweikard characterize cognitivism, success theory, mind-independence, and non-reductionism as realisms, and the opposed views of non-cognitivism, error theory, mind-dependence, and reductionism as anti-realisms. Third, with respect to these distinctions, they argue that only success theory provides a plausible candidate for realism in social ontology. This is because non-reductionism, although appropriate for characterizing realism in some local debates, and mind-independence, although regarded as the hallmark of realism in general metaphysics, are too maximal or demanding as definitions of realism in social ontology. Cognitivism, Laitinen and Schweikard argue, is too minimal.

In Chapter 18, “A nominalist theory of natural kinds and kind essences”**,** Markku Keinänen formulates an eliminativist nominalist theory of natural kinds, which is nonetheless compatible with central epistemic and explanatory functions of natural kinds and natural kind classifications. According to the developed eliminativist nominalist view of natural kinds, there are no natural kinds. Since there are no natural kinds, there are no natural kind essences or de re necessary properties of natural kinds. There is nonetheless true general talk about the members of natural kinds and classifications of objects with the help of natural kind terms, which track mind-independent divisions. The nominalist theory stresses the epistemic and explanatory functions of natural kinds and natural kind classifications. By contrast, the metaphysically heavy functions of collecting the necessary properties of the members of the kind and determining the identity conditions of objects, which realists about natural kinds tend give to natural kinds, are taken care of by the nominalist basic ontologies. Because of its flexibility, this nominalist view of natural kinds interlocks well with the new theory of reference Panu Raatikainen (2020, 2021) defends.

Renne Pesonen discusses the relationship between free will and intentional explanations in Chapter 19, “On the irrelevance of freedom to the causal relevance of will”. Many compatibilists believe not only that the freedom of the will is compatible with determinism but also that the notion of free will is indispensable for agency and intentional explanation. However, assuming that “will” can be given a psychological or other functional interpretation, concerns about freedom turn out to be mostly irrelevant for the agency or causal efficacy of the will. Arguments from the causal closure of the physical against the causal relevance of the will can be countered by the standard anti-reductionist analysis of levels of explanation: Will (or some of its psychological cognates) need not be free in order to be real and causally relevant. Questions concerning freedom are either metaphysical or moral, but they are routinely confused with the separable question concerning the causal relevance of the will for intentional explanation.

Teemu Toppinen and Vilma Venesmaa examine mental and normative causation in Chapter 20. Panu Raatikainen offers an account of mental causation drawing on an interventionist approach to causation—developed, especially, in the context of philosophy of science – and on the idea that causal claims would carry an (often) implicit reference to contrast classes. Toppinen and Venesmaa defend the conditional claim that, if the interventionist account of mental causation of the kind that Raatikainen proposes is correct, then normative properties have causal power, even given a non-naturalist or a quasi-realist understanding of such properties. They note that the truth of the conditional might be taken to be problematic for the style of account of causation that Raatikainen favours, since it is often believed that normative properties should *not* turn out to have causal power given a non-naturalist or a quasi-realist construal of such properties. But the objective of the chapter is not to argue for this conclusion, only to argue for the truth of the conditional claim.

In Chapter 21 "Mental causation, folk psychology, and rational action explanation", Tomi Kokkonen evaluates the currently standard solution to the problem of mental causation pioneered, among others, by Panu Raatikainen. While being sympathetic to this way to tackle the problem, Kokkonen argues that since folk-psychological explanations are inherently ambiguous, there is no solution to *the* problem of mental causation. Rather, a clarification of the issue leads into a more multi-layered explication of mental causation events.

In the final Chapter, 22 "Could Raatikainen have written otherwise?", Valtteri Arstila examines the problem of free will. The question of free will is one of philosophy's classical problems, one which professor Raatikainen has addressed on two points. First, he has levied influential criticism against the so-called "causal exclusion argument" originally made by Jaegwon Kim. Second, he has identified key problems in the psychological experiments conducted by Benjamin Libet in the 1970s that purported to find a 'readiness potential' in subjects' brains that allegedly determined their decisions prior to the decision becoming conscious. Arstila provides critical remarks of both criticisms made by Raatikainen.

On the whole, we the editors would like to extend our heartfelt thanks to the authors for their ideas, efforts, and promptness, and we hope the readers will find these texts rewarding. We also give our gratitude to the four anonymous reviewers who read and commented on the chapters, and to our two interns, Amanda Kimari and Elisa Viitasaari, whose help with the manuscript was invaluable. Lastly, we congratulate Panu once more for reaching this milestone in his productive career. Four Hurrays, or more! And as a customary ending, and an inside joke for those in the know, we would like to quote Cato: *"Ceterum censeo Carthaginem esse delendam."*

Part I
Philosophy and Truth

2
On 'naturalism'

Timothy Williamson

Introduction

I have known Panu Raatikainen and his work since the last millennium. He has always struck me as a force for both sanity and clarity, two quite different virtues. The first time I heard him give a paper, when I was still a professor at Edinburgh, he deftly used the basic distinction between theories and languages to diagnose a fatal confusion in some famous arguments about the philosophy of language, cutting through the technicalities to expose the underlying philosophical error. Again and again, Panu deflates pretentious overblown claims and looks beneath slick formulations to see what they conceal. He does so in debates where formal logical considerations loom large, and mathematical prowess can easily be mistaken for philosophical insight; because he understands the mathematics so well, he is able to show what it does *not* imply about the problem at issue. Such work is vital to keeping philosophy honest and on the right track. Long may Panu continue to contribute in his distinctive way!

Early in his career, Panu was much concerned with the philosophy of Quine, and its proper interpretation. That connection makes the theme of 'naturalism' not unnatural for this volume. I will pursue it in a spirit of which I hope Panu will approve, asking what lies behind the word, so often deployed as a mantra. My discussion will be sketchy and schematic; its main purpose is to emphasize how much is likely to be swept under the carpet when 'naturalism' is invoked.

Williamson, T. (2025). On 'naturalism'. In A. Laitinen, M. Keinänen, J. Reinikainen & A. Honkasalo (eds.) *Language, Truth, and Reality*. Tampere University Press. (pp. 19–28). https://doi.org/10.61201/h5fbdk64

Naturalisms

When I hear someone begin a sentence with the words 'As a naturalist, I …' my reaction has always been to stiffen with resistance, just as it is when I hear someone begin a sentence with the words 'As a Christian, I …'. I smell dogma and self-righteousness. What do their loyalties matter to me? But there is a difference. Whereas I *know* that I am not a Christian, I do not know that I am not a naturalist—nor do I know that I *am* a naturalist. Although both words—'Christian' and 'naturalist'—are vague, 'Christian' is at least precise enough for me to know whether it applies to me, whereas 'naturalist' does not even achieve that level of precision.

This unclarity was brought home to me by reactions to my book *The Philosophy of Philosophy*. Some philosophers described the approach developed in the book as 'naturalist', others described the same approach as 'anti-naturalist'.[1] The reason for the clash was not so much divergence in what more specific views they read into the book, as divergence in whether those specific views count as 'naturalist' or as 'anti-naturalist'. Evidently, the use of such an unclear and perhaps ambiguous word risks doing more harm than good. The term 'naturalism' needs to be clarified, and indeed attempts at such clarification have not infrequently been attempted, although with little impact so far on how the term is used in practice.

One standard distinction is between *ontological naturalism* and *methodological naturalism*. As the distinction is typically understood, both kinds of 'naturalism' privilege *science*, but in different ways. Schematically, ontological naturalism is the view that the ontology of science is (metaphysically) privileged over all other ontologies. For example, an ontological naturalist may claim that only those entities recognized by science genuinely exist. Correspondingly, ontological *anti*-naturalism is the view that the ontology of science is *not* (metaphysically) privileged over all other ontologies. Equally schematically, methodological naturalism is the view that the methodology of science is (epistemically) privileged over all other methodologies. For example, a methodological naturalist may claim that only those methods used by science yield genuine knowledge. Correspondingly, methodological anti-naturalism is the view that the methodology of science is not (epistemically) privileged over all other methodologies.

Neither ontological naturalism nor methodological naturalism strictly entails the other—if you doubt me, just try constructing a rigorous deduction of one view from the other. Nevertheless, ontological naturalism looks much easier to motivate from methodological naturalism than from methodological anti-naturalism. We can try to unfold this connection.

First, assume methodological naturalism. Thus, the methodology of science is (epistemically) privileged over all other methodologies. Now, one can reasonably

[1] The edition of *The Philosophy of Philosophy* that elicited these reactions is the first (Oxford: Wiley-Blackwell, 2007). Chapter 11 of the enlarged edition (2021) collects together my engagements with self-identified naturalists of several kinds (Andrea Bianchi, Hilary Kornblith, Penelope Maddy, Alex Rosenberg, and Robert Stalnaker) and develops some themes of the present remarks in more detail.

expect the conclusions delivered by an (epistemically) privileged methodology to be (metaphysically) privileged over the conclusions delivered by an (epistemically) unprivileged methodology. Therefore, by methodological naturalism, one can reasonably expect the conclusions of science to be (metaphysically) privileged over all other conclusions. As a special case, one can reasonably expect the ontological conclusions of science to be (metaphysically) privileged over all other ontological conclusions. In other words, one can reasonably expect the ontology of science to be (metaphysically) privileged over all other ontologies. In sum, methodological naturalism makes ontological naturalism a reasonable expectation.

Analogously, assume methodological anti-naturalism. Thus, the methodology of science is not (epistemically) privileged over all other methodologies. Now, one cannot reasonably expect the conclusions of a methodology to be (metaphysically) privileged over the conclusions of another methodology when the former is not (epistemically) privileged over the latter. Therefore, by methodological anti-naturalism, one cannot reasonably expect the conclusions of science to be (metaphysically) privileged over all other conclusions. As a special case, one cannot reasonably expect the ontological conclusions of science to be (metaphysically) privileged over all other ontological conclusions. In other words, one cannot reasonably expect the ontology of science to be (metaphysically) privileged over all ontologies. In sum, methodological anti-naturalism makes ontological naturalism an unreasonable expectation.

Those two arguments are far from watertight. The term 'privileged' is obviously imprecise, even as qualified by 'epistemically' or 'metaphysically', and 'reasonably expect' is at least as vague. Philosophers of science will wince at the crude talk of 'the methodology of science' and 'the ontology of science', as though all of science had the *same* methodology and the *same* ontology. Moreover, the terms 'methodology' and 'ontology' are themselves vague. At best, the two arguments provide a defeasible, *prima facie* connection.

One suppressed complexity is the relation to whatever specific question happens to be at issue in a given context of inquiry. A methodology may be very good at answering some questions and very bad at answering others. For example, the methodology of deductive proof is very good for answering questions in mathematics, but very bad for answering questions in biology or history. It is epistemically privileged as applied to the former, but not as applied to the latter.

Still more pernicious in practice is an ambiguity in the term 'naturalism', even as qualified by 'ontological' or by 'methodological', which derives from a pervasive ambiguity in the word 'science' itself, as used in their definitions. In a broad sense, 'science' means any kind of systematic, critical, evidence-based inquiry. In a narrower sense, 'science' means specifically *natural* science, comprising physics, chemistry, biology, and other sciences which use experiments, measurements, technical instruments, and the like. The most salient example of a *non-natural* science is *mathematics*, which is primarily proof-based. Another non-natural science is *history*, which is primarily document-based. Both mathematics and history are kinds of systematic, critical, evidence-based inquiry, in their very different ways, as befits the very different kinds of question they address, but they do not normally use experiments, measurements, technical instruments, and the like. In the broad sense, 'science' *includes* both mathematics and history. In the narrow sense, 'science' *excludes* both mathematics and history. *Soft naturalism* privileges science in the broad sense. *Hard naturalism* privileges science in the narrow sense.

The distinction between hard and soft naturalism cross-cuts the distinction between ontological and methodological naturalism. All four combinations are at least logically consistent:

> hard ontological naturalism with hard methodological naturalism
> hard ontological naturalism with soft methodological naturalism
> soft ontological naturalism with hard methodological naturalism
> soft ontological naturalism with soft methodological naturalism

However, the hard/hard and soft/soft combinations look more stable than the hard/soft and soft/hard combinations. For *hard* ontological naturalism looks comparatively easy to motivate with *hard* methodological naturalism, and comparatively difficult to motivate without it, while *soft* ontological naturalism looks comparatively easy to motivate with *soft* methodological naturalism, and comparatively difficult to motivate without it. For the hard/hard combination, one can run the two *prima facie* connecting arguments sketched above, with 'science' read throughout in the narrow sense. For the soft/soft combination, one can run the two arguments with 'science' read throughout in the broad sense. These motivating connections make the hard/soft and soft/hard combinations look correspondingly ill-motivated. We can reasonably use the term 'hard naturalism' for the combination of hard ontological naturalism with hard methodological naturalism, and 'soft naturalism' for the combination of soft ontological naturalism with soft methodological naturalism.

The hard/soft ambiguity in 'naturalism' is not innocent. For philosophers who self-identify as 'naturalists' not infrequently exploit the ambiguity by arguing *for* soft naturalism but then arguing *from* hard naturalism. That is cheating. They pay the price of the cheap version but walk out of the shop with the expensive one. Presumably, they are unaware of doing so. The pervasive ambiguity of the term 'science' as used in ordinary English—though not of the corresponding term in some other languages, such as German—facilitates the confusion. For example, one may find a self-identified naturalist arguing that 'science' is privileged by appeal to the advantages of systematic, critical, evidence-based inquiry, but then dismissing some philosophical discourse as 'unscientific' because it does not involve experiments, measurements, technical instruments, and the like. That is to equivocate on the word 'science'.

For clarity, we do best to examine hard naturalism and soft naturalism separately from each other.

Hard naturalism

Perhaps the most striking challenge to hard naturalism is *mathematics*. For, as noted above, mathematics is not a natural science, and so is not a science in the narrow sense; thus, it is not methodologically privileged, according to hard naturalism. Yet mathematics is as rigorous, exact, and 'hard' a form of inquiry as we have. Moreover, the natural sciences comprehensively depend on mathematics. How can natural science be methodologically privileged if it relies on the results of a methodologically unprivileged discipline, mathematics?

Quineans may respond that mathematics derives its privilege from that very indispensability to natural science: empirical confirmation goes to the total package of natural science and mathematics, not directly to just a part of it. But that holistic response ignores the methodological autonomy of mathematics as actually practiced; mathematicians do not vet new developments in mathematics (for example, in axiomatic set theory) for their integration with natural science. Moreover, the holistic response fails to vindicate hard naturalism proper, since it makes the conclusions of a non-natural science as epistemically secure as the conclusions of the natural sciences. Indeed, if mathematics can attain that epistemic status indirectly, through its relation to the natural sciences, may not the same apply to other disciplines too, perhaps even to philosophy?

Another challenge to hard naturalism is this: it cannot be established by the methods which it privileges. For it claims that the methodology of natural science is privileged over all other methodologies—for example, that only those methods used by natural science give knowledge. But such claims about methodological privilege are of a general epistemological nature. The characteristic methods of natural science are quite unsuited to testing such claims. Of course, one can imagine statistical studies of the reliability, or at least level of consensus, achieved by different methodologies. But to design, motivate, and implement such tests of diverse methodologies would itself require abstract epistemological reasoning, rather than the use of experiments, measurements, technical instruments, and the like. If hard naturalists reply that the methodology of natural science includes such abstract epistemological reasoning, they risk watering down their 'hard' naturalism to a point where it no longer serves their dialectical purposes. In particular, they will be unable to dismiss abstract epistemological reasoning as 'unscientific'. Thus, by its own standards, hard methodological naturalism has a low epistemic standing.

I have pressed both these challenges on hard methodological naturalists, without ever receiving an effective response.

As for hard ontological naturalism, the other half of hard naturalism, its motivation comes from hard methodological naturalism, as explained above, and so is undermined by the problems just explained for hard methodological naturalism. But there is also a more specific problem for hard ontological naturalism, understood as saying that there is only what natural science says there is. For natural science itself does not say that there is *only* what natural science says there is: it does not address such general metaphysical questions. For example, particle physics does not say that there are *only* particles. It does not say that there are no non-particles such as wars or societies or suchlike; it simply does not raise the question whether there are wars or societies. Thus, a denial that there are wars or societies does not have the authority of natural science behind it. Equally, of course, an *assertion* that there are wars or societies does not have the authority of natural science behind it, but failing to answer a question does not amount to giving it a negative answer.

At this point, hard ontological naturalists may invoke Ockham's Razor, arguing that if our best methodology (by hypothesis, that of natural science) does not require us to postulate a more populous ontology than that of natural science, we are justified in the parsimonious postulate that there is nothing beyond that sparse ontology. But that conclusion is a *non sequitur*. For even if the methodology of natural science is epistemically better than that of all other methodologies, it does not follow that those other methodologies are epistemically worthless, especially on questions about which natural science has nothing to say. In particular, if history tells us that there are wars and societies, and natural science does not tell us otherwise, it may be a good bet that there are wars and societies. The testimony of moderately reliable sources may make a proposition much more probable than not, when our most reliable sources do not address the question. Otherwise, law courts would have to revise their procedures drastically.

A more positive observation is in order. Even if a given theory in natural science does not posit entities of a given kind, one may still have to posit entities of that kind in order to explain how there is evidence for the theory. For example, a theory in particle physics may not posit macroscopic objects, but explaining the nature of the empirical evidence for it may involve bringing in macroscopic observers and their macroscopic instruments of observation. Scrutinizing a theory involves scrutinizing the confirming or disconfirming evidence.

In short, the further the debate goes beyond slogans and bluff, the harder hard naturalism is to take seriously.

Soft naturalism

The preceding challenges to hard naturalism pose no threat to soft naturalism. After all, for soft naturalism, science includes mathematics, history, and even epistemology, at least when they are done in a systematic, critical, evidence-based way, as they often are. Nevertheless, even soft naturalism faces some residual challenges.

Despite the soft naturalist's inclusive view of science, the emphasis on systematic, critical, evidence-based inquiry tends to privilege *reflective* cognitive steps—the conclusions of systematic inquiry—over non-reflective steps. But reflective steps depend on non-reflective ones, on pain of an infinite regress. For reflection consists in consciously chaining together many individual steps: a simple paradigm is a mathematical calculation. Those individual steps are not themselves reflective. This does not mean that we cannot later criticize or justify those non-reflective steps, just that such a process can never be brought to completion: at any point, we are relying on some steps on which we have not yet reflected. For finite inquirers, full reflection is an impossible ideal. This is a much less severe challenge than those considered above to hard methodological naturalism, but it is not trivial. After all, it is not obvious that more reflection must always lead to better cognition. There is such a thing as overthinking. Decision-making on the basis of elaborate conscious reflection does not always end better than unreflective decision-making.[2]

We can reasonably expect that any limitations of soft methodological naturalism will tend to have repercussions for soft ontological naturalism too. If we cannot justify awarding exclusive methodological privileges to science, broadly understood, why should we assume that only those entities recognized by science, broadly understood, exist?

After all, it is far from obvious that whatever exists can be known (scientifically or unscientifically) to exist. To put the point crudely, if the epistemic privilege of science means that whatever can be known can be known scientifically, that privilege does not entail that whatever is true can be known scientifically. To bridge the gap, one needs the additional lemma that whatever is true can be known, but what is the evidence for that lemma? The inductive case that sooner or later science always succeeds in finding the answers to its questions is far from convincing. Questions about the ultimate constitution of the universe have been around since the ancient beginnings of science, and are still nowhere near to being answered. Moreover, we are now reading 'science' in the broad soft naturalist sense, so we also need to consider whether non-natural sciences sooner or later always succeed in finding the answers to their questions. Mathematicians are nowhere near to establishing new axioms of set theory that would enable them to prove or refute Cantor's Continuum Hypothesis, and philosophers of mathematics are nowhere near to establishing whether it even has a determinate truth-value. Many questions in ancient history will forever remain unanswerable because too little potential evidence has survived. Why should we even assume that all entities, states of affairs, properties, and relations are capable of being picked out in thought? If they cannot be thought of, they cannot be known.

Naturalists themselves (hard or soft) often make the point that current science is not final; we must expect science to continue making new discoveries and revising its current theories. Thus, if all truths of some kind will sooner or later be known to a given science, and some putative truths of that kind are not currently known to that science, it does not follow that they are not genuine truths. That science may come to know them in a few centuries. The gap between a science in its current state and its ideally completed version may be potentially so large that applying naturalist slogans (hard or soft) in practice may have to be a highly speculative business.

[2] For the limits of reflection, see Hilary Kornblith, *On Reflection* (Oxford: Oxford University Press, 2012).

Concluding reflections

I have not attempted to survey all the claims to which the label 'naturalism' has been attached, but I hope to have given a sense of why the word as philosophers currently use it does not denote a theory in good shape. Nevertheless, one might still feel, although the term is associated with some negative tendencies, such as scientism, it is also associated with some positive tendencies. To put it crudely, 'If you want to know the answer to a question in physics, ask a physicist, not a preacher' is good advice. The point is not restricted to natural science. 'If you want to know the answer to a question in history, ask a historian, not a politician (or a physicist)' is also good advice.

More generally, science is an amazing source of knowledge, and so of evidence that one can bring to bear in assessing other claims. In philosophy, the term 'naturalism' can serve as a useful reminder that scientific evidence may be relevant in unexpected ways to philosophical theories. For instance, evidence for Einstein's theory of special relativity is at least relevant to the philosophy of time, even if the connection is not as direct as some may assume. Such connections may be far more widespread than philosophers have fully recognized. For example, the whole internalist tradition in epistemology, which grounds the justification of belief in the subject's conscious states, is at risk of being undermined by neuroscientific evidence that conscious processes are too slow to implement internalist models of justification for most ordinary perceptual beliefs.

Some sub-traditions of philosophy have a tendency to parochialism, a habit of not considering such 'alien' evidence, even if they have no principled justification for that habit. But they may have been put off by over-eager self-identified 'naturalists' who apply results from natural science too crudely to philosophy, riding roughshod over subtle logical distinctions between the natural scientists' questions and those the philosophers are asking, perhaps because natural scientists themselves ignore those distinctions when they become amateur philosophers—such as neuroscientists who ignore compatibilism when claiming to have refuted free will. Even philosophers who make an effort to apply relevant research in linguistics, psychology, biology, or whatever may be put off by the high levels of disagreement among the scientists, and how fast the science changes, making it hard to extract well-established conclusions from the science to use as constraints in their philosophizing. But that does not justify treating the science as simply irrelevant. Those disappointing levels of disagreement amongst the scientists are evidence that theorizing in that science is more like familiar messy theorizing in philosophy than those philosophers had idealistically hoped.

The word 'naturalism' may sometimes function as a flag: seeing it encourages one to keep engaging with evidence from other sciences, undaunted by the difficulties. What invoking 'naturalism' as a general theory cannot do is act as some sort of all-purpose enforcer, *making* scientific evidence relevant to philosophical theories. If there is a gap between theory and evidence, as there usually is, 'naturalism' does not stand for any plausible general doctrine that can somehow bridge the gap, mediating between theory and evidence. The typically non-deductive evidential connection must be assessed on its own merits, for that specific theory and that specific evidence; it is not strengthened by something called 'naturalism'. The evidence may just *be* relevant to the theory, irrespective of whether some further theory says it is.

Sometimes, brandishing the word 'naturalism' may even act as a *substitute* for serious engagement with the relevant science. For example, Quine developed his naturalized epistemology mainly by armchair reflection, though under the influence of already outdated behaviourist psychology, with little interest in the rapidly developing experimental cognitive psychology of his time, despite its obvious relevance. Similarly, naturalists who take physics to have shown that really there are just 'atoms in the void' have not paid much attention to the actual development of physics over the past century. Prioritizing experimental methods over armchair reflection does not make much difference if you decide by armchair reflection what results experimental methods must have.

Of course, a bad metaphilosophical theory may still deny the relevance of empirical evidence to philosophical questions. Thus, on an old-fashioned, simple-minded metaphilosophical view, philosophical questions are conceptual questions, and empirical evidence is irrelevant to conceptual questions. Invoking 'naturalism' may signal one's rejection of such metaphilosophical views. But the 'naturalism' did not make the connection between the philosophical theory and the empirical evidence; it was there all along. One can recognize the connection without invoking any specific metaphilosophical theory, just by reflecting on what the theory says and what the evidence says. Still, if repeating the word 'naturalism' helped remove metaphilosophical blinkers that prevented philosophers from seeing evidential connections, it did some instrumental good.

No form of naturalism has the power to show that every good philosophical argument must invoke 'empirical' evidence, just as it has no power to show that every sound proof of a mathematical theorem must invoke such evidence. For all that, evidence from natural science often *is* relevant to philosophical claims. But one can acknowledge all that without endorsing any distinctive theory of naturalism. One need only reject exceptionalism about philosophy.

References

Kornblith, Hilary (2012): *On Reflection.* Oxford: Oxford University Press.

Williamson, Timothy (2007/2021): *The Philosophy of Philosophy.* 2nd edition. Oxford/ Malden: Wiley-Blackwell.

3
In search of analytic philosophy

Anssi Korhonen

1. Analytic philosophy, along with phenomenology, was the leading philosophical trend in twentieth-century European philosophy. Arguably, analytic philosophy is still alive today (this is not entirely uncontroversial, though), and the division between analytic and continental philosophy is still a valid one as an institutional matter of fact. The question what analytic philosophy is and what it has been, has cultural significance, and it may have philosophical significance as well. For instance, if one thinks that analytic philosophers have made some important discoveries and that these discoveries and insights are now in danger of being lost, then one way to resist this development would be by articulating what was characteristic of analytic philosophy. The question may be significant for one's self-understanding, too. Both these motives are present in different degrees in Georg Henrik von Wright's contributions to the topic, for instance (von Wright 1993, 2000).

The question "What is analytic philosophy?" became a topic of debate in the early 1990s. This was largely due to the appearance of Michael Dummett's book *Origins of Analytical Philosophy* (Dummett 1993). The debate is no longer as active as it used to be, but the topic has not become defunct, either, and new branches have grown into it, such as the question of the identity of analytic philosophy vis-à-vis continental philosophy.

The single most important factor behind the original debate was the phenomenon known as the *historical turn in philosophy* (Beaney 2013, Reck 2013). Analytic philosophy had enjoyed the reputation of being "philosophy without history", and analytic philosophers had enjoyed the reputation of being *ahistorical* or even *antihistorical* philosophers, who "think for themselves" and do not lean on history and tradition

Korhonen, A. (2025). In search of analytic philosophy. In A. Laitinen, M. Keinänen, J. Reinikainen & A. Honkasalo (eds.) *Language, Truth, and Reality*. Tampere University Press. (pp. 29–48).
https://doi.org/10.61201/h5fbdk64

(unlike they distant, continental cousins). Now, this topical and timeless orientation did not disappear, of course; but there arose a marked interest in the roots and origins, of analytic philosophy, and a completely new discipline was created within academic analytic philosophy: *early analytic philosophy* (Sluga (1980) and Hylton (1990) were two important early landmarks here). Of course, the historical turn itself didn't come out of nothing. An easy and quick partial diagnosis would refer to an *identity-crisis:* ever since the early 1960s, analytic philosophy had grown more and more heterogeneous and diffuse, and philosophers' self-image was becoming less clear and distinct. Not everyone cared about this, but many did, and one reaction was "to subject analytic philosophy to a historico-critical scrutiny" (von Wright 1993, 26).

My own interest in the question is related to the historical turn. The analytical tradition in philosophy is a philosophically and historically exciting phenomenon. I also think that the best way to *introduce analytic philosophy* is through its history. To explain what analytic philosophy is, we may turn to contemporary work in the discipline. But even this perspective is difficult to understand without considering the developments of, say, the past fifty years (it seems though that the border between "this is contemporary and, therefore, relevant" and "this is past and, therefore, of historical interest only" is continually moving closer and closer to us). On the other hand, a case can be made that that our philosophical understanding *is* partly historical and that, therefore, the study of philosophical past is "of more than historical interest". Personally, for what it's worth, I am inclined to think that this is, indeed, so; but quite apart from that, the study of past philosophy ought to be pursued on its own as well.

2. My aim here is to say something constructive about the twin-question, *what analytic philosophy is and what it has been.* I try to explain, at a relatively general level, what in my view is the best – the most reasonable and fruitful – approach to the twin-question. The following two points will serve as starting-points. They may appear as self-evident; but as we will see, they are not quite that:

I) Analytic philosophy is a genuine and distinctly recognizable *philosophical* and *historical* phenomenon, whose identity differs from that of, say, phenomenology and the phenomenological tradition.

II) Analytic philosophy as a historical phenomenon I shall refer to as 'analytic tradition' and shall identify it in a way that is entirely uncontroversial. First, it is a key tradition in twentieth century western or (if you prefer) European thought. Second, the tradition has existed for at least one hundred and twenty years, maybe over a hundred and forty years. The question which philosophers belong in this tradition, has been answered differently by different participants in the debate over the identity of analytic philosophy. Since we are not trying to define a previously unknown phenomenon, we must accept as our starting point a more or less agreed upon understanding. A handy ostensive definition is given by Sluga:

> Following common practice, I take analytic philosophy here as originating in the work of Frege, Russell, Moore, and Wittgenstein, as encompassing the logical empiricism of the Vienna Circle, English ordinary language phi-

> losophy of the post-war period, American mainstream philosophy of recent decades, as well as other worldwide affiliates and descendants. (Sluga 1997, 17fn.)

3. An intuitive starting-point, such as the one by Sluga, is inevitably imprecise. Unhappy about this, some scholars have adopted the radical measure and have, in fact, *rejected* the entire category of "analytic philosophy", arguing either that no genuine analytic tradition has ever existed or, else, that it is nothing but an arbitrary construction, or imposition, created by misinterpreting such allegedly analytic philosophers as Russell and Moore.[1] Usually, the complaint has been that the term "analytic philosophy" *cannot be given an analytic definition:*

> Analytic $=_{\mathrm{df}}$ *philosophy* that…

Here one is looking for a distinctive characteristic (more likely: a class of such characteristics) with which to distinguish *analytic* from *non-analytic* philosophy (or philosopher). An analytic definition is reminiscent of an Aristotelian definition *per genus et differentiam*, although no one is likely to think of real definitions here. Rather than definitions, we may simply speak about *necessary* and *sufficient conditions*: a philosophy (or philosopher) is analytic if and only if…

There is a legion of such distinctive characteristics that could be used here. They are quite familiar, and commentary would be superfluous here:

> Conceptual analysis, linguistic turn, use of formal logic, anti-psychologism, rejection of metaphysics, rejection of philosophical systems, rejection of history of philosophy, scientism, naturalism, argumentation, pursuit of inner clarity, pursuit of rigour.

This list could easily be expanded. The idea that 'analytic philosophy' or 'analytic philosopher' could be defined by means of such distinctive marks runs into an evident difficulty. For every such list, whatever its members, will inevitably exclude philosophers that we would, with good reason, like to classify as 'analytic'. Another likely consequence is that our chosen list of marks picks up a philosopher whom we do *not* wish to classify as a philosopher, and again with good reason. To put the point simply: the analytic tradition is much too heterogeneous or diverse to permit an analytic definition in the above sense.[2] Therefore, the very idea that a satisfactory analytic definition could be framed is likely to appear very much like a stillborn venture. What are we to do in this situation?

4. Three strategies are available here: first, we could *stipulate* a meaning for the term, if we believed that the introduction of 'analytic philosophy' into discourse as if

[1] Cf. Preston (2017).

[2] For an elaboration, see Raatikainen (2001, 191–197).

it were a fresh technical term served some useful purpose; second, we could *dismiss* putative analytic definitions, if we believed that looking for necessary and sufficient conditions for 'analytic philosophy' is a misguided enterprise; third, we could formulate a *revisionary* definition, if we believed that a partial revision of 'analytic philosophy' helped us to gain some insight into the analytic tradition (the dividing line between the stipulative and revisionary strategies is not very sharp).

Michael Dummett's well-known definition of analytic philosophy includes a significant stipulative element. He used 'linguistic turn' for the purpose: 'analytic philosophy' (or 'analytical philosophy', as Dummett liked to call it) is distinguished from other philosophical schools by two beliefs: first, that a philosophical account of thought can be attained through a philosophical account of language; secondly, that a comprehensive account can only be so attained (Dummett 1993, 4–5). Its first clear manifestation is to be found in Frege's *Foundations of Arithmetic*, but the decisive step was taken by Wittgenstein in the *Tractatus* (*ibid.*, 127–128). It follows from Dummett's definition, for example, that Russell and Moore were not really analytic philosophers at all, no matter how much they may have contributed to the formation of the tradition. This goes against the established use and what we think we know about the analytic tradition. Most of us would have believed that Russell and Moore were among "the founding giants of analytic philosophy" (Soames 2014), but now it turns out that they were not really analytic philosophers at all, but were at best its uncles or great-uncles, while Frege qualifies as its grandfather (Dummett 1993, 171).

Note that Dummett's definition is primarily *stipulative and normative and not classificatory at all*. In his view, Frege's philosophy was an important step in the right direction, which is the insight that a philosophical study of language (philosophy of language, theory of meaning) ought to be the foundation of all philosophizing. If considered as a piece of serious historiography, Dummett's definition must have struck many as downright bizarre. Once we take into account his real intentions, however, we see how different they were from those of an ordinary, down-to-earth historian of philosophy.[3]

For us who take the historical turn seriously, the concern is with real history and not with a stipulative and normative use of past philosophers and their ideas. We acknowledge, then, that the term 'analytic philosophy' does have an established use; that it is, indeed, a *lexical item* in standard philosophical terminology; and, finally, that the lexical item either can or cannot be turned into a useful tool in our historical inquiries by means of an analytic definition. This is the approach in Glock's well-known study (2008, Chapter 1), and I concur with it, up to a point.

Being a lexical item with an established use, the term has a tolerably clear extension and hence *clear positive* and *clear negative* cases. Such figures as G. E. Moore, Bertrand Russell, the early Wittgenstein, Susan Stebbing, Rudolf Carnap, G. E. M. Anscombe, Georg Henrik von Wright, David M. Armstrong, David Lewis and Timothy Williamson are clear examples of analytic philosophers. And Edmund Husserl,

[3] Cf. here Matar (2017).

Martin Heidegger, Edith Stein, Jean-Paul Sartre, Hannah Arendt, Michel Foucault, Jürgen Habermas and Slavoj Źiźek are equally clear cases of non-analytic philosophers. There are also *unclear* cases, philosophers who "look like analytic philosophers" but whose relationship to the tradition is somehow problematic: Bernard Bolzano, the later Wittgenstein, Karl Popper and Paul Feyerabend would be good examples.

At this point it may be argued, however, that even if our *terminus technicus* does possess a reasonably well-established use, a closer inspection will nevertheless show that its *extension is arbitrary*. This claim is the gist of the dismissive strategy: as Dagfinn Føllesdal (1997) puts it, whatever "principle of classification" we use in our analytic definition, it *cannot generate a class possessing genuine unity*. Føllesdal argues that any classification of "current philosophical trends" inevitably suffers from flaws that are not unlike the flaws of the famous Chinese imperial taxonomy of animals in Jorge Luis Borges' essay "The Analytical Language of John Wilkins"; in 'a certain Chinese Encyclopedia', animals were divided, among others, into those that belong to the Emperor, tame, sucking pigs, mermaids, stray dogs, those drawn with a very fine camel hair brush, and those that from afar look like flies.

We are familiar with our homely analytic tradition, and a list of 'analytic philosophers' will not strike us as arbitrary and fanciful like Borges' charming list. But why not? Føllesdal puts forth the valuable question: *What* are we trying to define, when we define analytic philosophy? What kind of thing or phenomenon is it? Is analytic philosophy:

> a) a doctrine or set of doctrines, b) a set of characteristic problems, c) a set of canonical texts, d) a set of philosophical virtues, e) a school, f) a movement, g) a tradition, h) a progressive philosophical program; or something else?

Føllesdal's own reply is subversive: no 'analytical trend' can be identified within contemporary philosophy. The reason is not that no such trends exist; they do exist, he holds, because suitable distinctive marks *can* be found for phenomenology, hermeneutics, etc. The trouble is specifically with the alleged analytical tradition itself: considered as a twentieth century philosophical movement, it lacks genuine unity.

Føllesdal is not entirely dismissive of 'analytic philosophy', though. No such movement exists, he argues, but there is a general and timeless *analytical approach* to philosophy. It is not a method but has to do with the most general philosophical virtues; it is the approach *by justification and argumentation*. If you are "very strongly concerned with" justification and argumentation, Føllesdal (1997, 7) suggests, then you qualify as an 'analytic' philosopher. (This may look rather thin, but the impression would be somewhat misleading, as Føllesdal uses the notion of reflective equilibrium to elaborate on the relevant notion of 'argument and justification'.) Of course, the exercise of these virtues is not confined to any philosophical current of today, or of the past century: Aristotle, St. Thomas of Aquinas, Descartes "as well as a large number of other truly great philosophers" were analytic philosophers in this sense (Føllesdal 1997, 14).

This purely methodological conception of 'analytic philosophy' has three consequences:

(iii) 'Analytic' is a term that applies *independently of school and era*; an ancient sceptic, a medieval schoolman, a German idealist, a twentieth-century phenomenologist and a logical positivist can all of them insist that arguments must be given to support philosophical theses.

(iii) It makes sense to talk about *degrees of 'analytic'*; it makes sense to say, for instance, that Husserl was more analytic than Heidegger.

(iii) Although analyticity is a virtue that philosophers have always exercised, it is nevertheless *not a trivial characteristic*; one can be a philosopher without putting much emphasis on this virtue: Pascal, Kierkegaard, Nietzsche and Heidegger *might* be examples of philosophers who put less emphasis on analyticity in this sense.

To use 'analytic' in this way, exclusively as a virtue category, has an obvious weakness: we can no longer speak about the *analytic tradition in philosophy* in the ordinary, well-established sense. In this sense, Aristotle and Saint Thomas, for instance, were *not* analytic philosophers, although their works are replete with arguments; and in this sense, Husserl, for instance, was not one of twentieth century analytic philosophers, although he may have been more analytic than most phenomenologists and undoubtedly was at least as analytic as many analytic philosophers.

The dismissive strategy, in my opinion, is too radical. The real problem is not in 'analytic philosophy' or 'analytic tradition'; it's in the idea that we should use an analytic definition in their delineation. The problem is (to quote David Hilbert from an entirely different context) that here "one is looking for something one can never find because there is nothing there; and everything gets lost and becomes vague and tangled and degenerates into a game of hide and seek".

Before we conclude that the provision of an analytic definition for 'analytic philosophy' really is just a game of hide and seek, however, we should consider the revisionary strategy. Unlike Føllesdal, it does not dismiss analytic philosophy as a unitary phenomenon but *redefines it within reasonable limits*. Unlike Dummett, it does not propose to stipulate a precise meaning for 'analytic philosophy'; it complies up to a point with the established and familiar usage, but revises our pre-analytic understanding of the term in order to turn it into a useful tool for classification. In brief, the revisionist provides a Carnapian *explication* for the term 'analytic philosophy'.

5. Panu Raatikainen, in his search of analytic philosophy, has proposed just such an explication.[4] He proceeds in two steps. First, he focuses on what he calls the *original meaning of 'analytic philosophy'* by considering how the term was introduced into philosophical vocabulary. This was a lengthy process, extending from the 1930s until the 1950s (more of this below). The original meaning is what fixes the reasonable limits for his revisionism and covers what he calls *orthodox analytic philosophy;* the heyday of analytic philosophy extended, roughly, from the late 1920s until the late 1950s, and hence Raatikainen's provocative title "What *was* analytic philosophy?".

Raatikainen's second step is the *extension of the original meaning*. The term 'analytic philosophy', although it was introduced in the 1930, made its real breakthrough only in the 1950s, by which time it had come to mean, very roughly, *the sort of philosophy where the focus is on language and the clarification of meanings*. The term caught on, but as the analytic tradition kept on developing in new directions, the sense acquired fresh layers (and lost older ones). Furthermore, the term's coverage was extended *backwards* as well, so as to cover the *roots* of 'analytic philosophy' in Cambridge (Moore and Russell) and in Jena (Frege). These extensions of the analytic canon were based primarily on perceived lines of influence, with the consequence that the doctrinal shape of 'analytic philosophy' rapidly grew less and less clear.

Raatikainen argues that we obtain terminological clarity if we stick to the original meaning of 'analytic philosophy'. In this way, we can still use the term "as a clear and distinct, serviceable, contentually classifying expression of the history of philosophy" (2013, 21). Furthermore, he has a straightforward answer to someone who is accustomed to thinking of Moore or Russell, say, as paradigmatic analytic philosophers: "[T]he problem is solved [...] when one distinguishes, on the one hand, the philosophical movement or school of thought proper, and, on the other hand, its essential predecessors and background figures" (*ibid.*). Moore and Russell were not yet genuine analytic philosophers *sensu stricto*. Orthodox analytic philosophy is what *derives* (partly) from these gentlemen. And similarly for later developments: "They could perhaps be called, if one wants to emphasize their background, 'post-analytic philosophers'" (*ibid.*, 23).[5]

[4] Raatikainen (2001, 2013).

[5] Skorupski (2013) offers a similar construction.

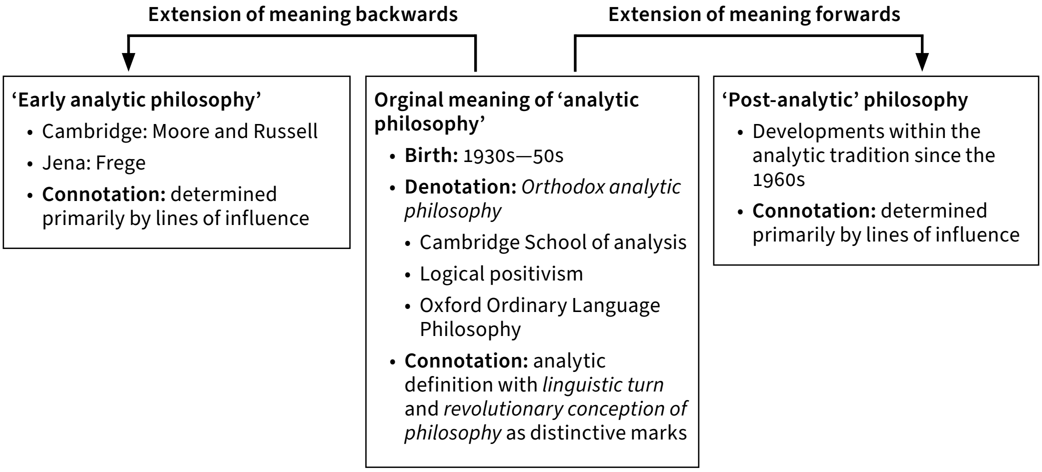

Extension of meaning backwards
Extension of meaning forwards
'Early analytic philosophy'
• Cambridge: Moore and Russell
• Jena: Frege
• Connotation: determined primarily by lines of influence
Orginal meaning of 'analytic philosophy'
• Birth: 1930s—50s
• Denotation: Orthodox analytic philosophy
• Cambridge School of analysis
• Logical positivism
• Oxford Ordinary Language Philosophy
• Connotation: analytic definition with linguistic turn and revolutionary conception of philosophy as distinctive marks
'Post-analytic' philosophy
• Developments within the analytic tradition since the 1960s
• Connotation: determined primarily by lines of influence

Figure 3.1: The genesis of the analytic tradition, according Raatikainen's (2001, 2013) revisionary definition of 'analytic philosophy'

Raatikainen's key point is that orthodox analytic philosophy does, indeed, form a genuine unity; that, in fact, *orthodox analytic philosophy can be given an analytic definition*. His definition is based on two distinctive marks, both of which derive from the *Tractatus*. First, there is the linguistic turn, or the idea that "the sole task of all legitimate philosophy is the analysis of language, the clarification of meaning, or such" (2013, 20–21). Second, there was the revolutionary ethos accompanying the linguistic turn, that "one was witnessing a definite turning point in the history of philosophy, a wholly new revolutionary way of understanding the task of philosophy and the nature of philosophical problems" (*ibid.*, 20).

In the next section, I shall argue that the conceptual clarity created by Raatikainen's revisionary definition is spurious: the three-fold distinction as in the above diagram (figure 3.1) is in itself a good way of looking at the analytic tradition, but Raatikainen makes it rather too principled. I argue that we can come to see this if we first look at the relevant facts about the original meaning of 'analytic philosophy' and then consider the true shape of Raatikainen's "orthodox analytic philosophy"; there were 'schools' or 'movements' in the analytic tradition, but as long as we consider real life phenomena, they do not really permit any definitions in the strict, analytic sense.

6. As Raatikainen points out, the term 'analytic philosophy' is of surprisingly late origin: it was introduced in the 1930s but did not really catch on until the 1950s. Here's an outline of the earlier developments, different in some important respects from Raatikainen's version of the story.[6]

The term came to use in the early 1930s when the English philosophical community began to use it to denote a particular group within that community, a group that came to be known as *the Cambridge School of Analysis*. They were, by and large,

[6] An avid reader might want to consult Frost-Arnold (2017) as well.

followers of G. E. Moore, they held the view that the analysis of common sense and scientific facts was the proper field for philosophers, and they put considerable effort into the clarification of the notion of analysis itself. A. E. Duncan-Jones, himself one of these analytic philosophers, observed in 1937:

> "The question asked in this title ["Does Philosophy Analyse Common Sense?"] relates, of course, to philosophy as understood and practiced by a particular limited group of philosophers; primarily the contemporary philosophy of the people in his country who have commonly been called analytic philosophers." (Duncan-Jones 1937, 139)

The wording here suggests that by 1937 the term 'analytic philosopher' enjoyed a well-established use in Britain. Looking at published sources, we find R. G. Collingwood criticizing 'analytic philosophy' and 'analytic philosophers' as early as 1933, in Chapter 7 of *An Essay on Philosophical Method*. The current scholarly consensus seems to be that this is the first literary occurrence of the term 'analytic philosophy', while John Wisdom had used 'analytic philosophers' in 1931 in his book on Bentham and philosophical method (Wisdom 1931; Beaney 2013, 42).[7] These terms then occur several times in a Symposium organized by the Aristotelian Society in 1934, which was entitled "Is Analysis a Useful Method in Philosophy?", with contributions from John Wisdom, Maurice Cornforth, and Max Black.[8] Similar discussions continued throughout the rest of the 1930.

The philosophers of this Cambridge School of Analysis were the original 'analytic philosophers'. In 1935, A. J. Ayer mentioned four of them by name: Susan Stebbing, John Wisdom, C. A. Mace, and A. E. Duncan-Jones.[9] In 1938, Max Black gave a fuller list under the title "Some of the analytical philosophers in England". It included Frank Ramsey (who had died 1931), Stebbing and a dozen or so 'younger philosophers'.[10]

In 1935, A. J. Ayer lectured in Paris to an international audience about the "analytic movement in contemporary British philosophy". Ayer himself sought to blend together logical positivism and British empiricism (as in *Language, Truth and Logic*), but he clearly identified himself with the analytic movement. In this way, he already took a step towards *widening* the extension of 'analytic philosophy', as he recognized an important affinity between a number of British philosophers and their Continental colleagues. Ayer, though, was critical of colleagues both at home and abroad. Log-

[7] To the best of my knowledge, however, the very first occurrence of 'analytic philosopher' is as early as 1922 and is due to none other than Bertrand Russell. It occurs in a somewhat casual book review and is brief but not without interest. I won't discuss this early specimen here, however, as it was just a foretaste of something that was still in the future and appears to have had no effects. Of course, it would be unfair not to mention here *the first ever analytic philosopher by name;* as you might expect, the title goes to G. E. Moore (Russell 1922, 406).

[8] Black et al. (1934).

[9] Ayer (1936, 57).

[10] Black (1939, 34–35).

ical positivists, he claimed, were prone to exaggerate the revolutionary character of their conception of philosophy as analysis; *this* had always been a standing feature of British empiricism. On the other hand, while Ayer found the gist of philosophical analysis in Russell's method of logical constructions, he argued that colleagues at home had not sufficiently appreciated its true nature, as they continued to formulate it using misleading metaphysical vocabulary.

In 1938, Max Black gave yet another survey of contemporary British philosophy, focusing on the Cambridge School of Analysis. The label, he argued, was convenient but still an exaggeration, as it was hard to find a single principle that all supporters of the analytic method would have accepted. More fitting, he explained, would be talk of "analytic movement" or just "analytical philosophy in England", characterized by an "unmistakable climate of opinion that was hostile to metaphysics and speculative philosophy, and sympathetic to analysis" (Black 1939, 24).

Now, insofar as *this original analytic philosophy* had a defining feature, it was an emphasis on *method*. Collingwood (1933) was critical of what he termed 'analytic philosophy', precisely because an exclusive focus on given facts and their analysis left no room for constructive philosophical thinking, and led to a kind of scepticism. He instances Moore and Stebbing as 'analytic philosophers', thus making clear that his criticism is of real-life philosophers and not some ideal type. Almost as a reply to this criticism, the English analytic philosophers of the 1930s debated the nature of analysis intensely. What are logical constructions? Is analysis concerned with worldly facts or with language? Does analysis possess a "direction", or is the *analysans* on the same level as the *analysandum?* Is analysis concerned with facts licensed by common sense? These are examples of their questions, and nothing in the debate is indicative of a convergence of opinions. Susan Stebbing, in her last contribution to the debate, simply declared that she had grown tired of the entire topic.[11]

We see that 'analytic philosophy' was originally a very British phenomenon. Then, in 1936, the term was used more freely by Ernst Nagel in his paper, 'Impressions and Appraisals of Analytic Philosophy in Europe', published in *The Journal of Philosophy*. Nagel, born European, was an American philosopher who received his education at Columbia. He spent the academic year 1934–1935 in Europe as a Guggenheim scholar, visiting five major philosophical centres: Vienna, Prague, Warsaw, Lwów (Lviv) and Cambridge. Likely – and this is my conjecture—he picked a useful term in Cambridge and used it *to make a bold generalization:* 'analytic philosophy' was a European and not just a narrowly British phenomenon. With this generalization, he wanted to assure his fellow Americans that "a romantic irrationalism had not completely engulfed Europe" (1936, 5). By Nagel's reckoning, analytic philosophy existed in Europe in several places and forms: (i) in Cambridge, which was dominated by G. E. Moore and Wittgenstein; (ii) on the Continent in Vienna, Berlin and elsewhere, where it existed as different versions of 'logical positivism'; and (iii) in Poland, where it existed as 'nominalistic naturalism, dominated by the logico-analytic method'.

[11] Stebbing (1938–1939), 71.

Nagel used four features to describe European 'Analytic philosophers'. As we would expect, they were occupied with philosophy as analysis. That is, they took scientific results for granted and didn't expect to add to it but to clarify it; discussions of *method* dominated all these places. They had little patience with philosophical systems in the traditional sense. Also, they didn't care about history of philosophy, the only point of which was to see how philosophers of a previous generation had committed a particular logical blunder. Finally, they subscribed to a kind of common sense naturalism, the gist of which was the conviction that philosophy couldn't deliver anything that conflicted with "informed practice and common experience" (1936, 7). This, though, was not a doctrine but an underlying tenet at best; for analytic philosophers never asserted a *Weltanschauung* as a part of their philosophy. Although he is more elaborate than Black, their respective pictures of analytic philosophy are quite similar: analytic philosophers often differ in their doctrines (Nagel sketches out recent developments in Cambridge and in the *Wiener Kreis*), and what they have in common is an attitude, or a number of basic convictions, both positive and negative.

Sketchy as it is, the above account of the original meaning is hopefully enough to convince the reader of the following two points. First, from its inception, 'analytic philosophy' was recognized by everyone to be a heterogeneous phenomenon. Second, its unity was less a matter of doctrine than of certain attitudes. The second point is strengthened by noting that Black's observation about the English climate of opinion in the 1930s can, in fact, be generalized so as to cover all of what Raatikainen calls *orthodox analytic philosophy*.

7. In many ways, the Vienna Circle is the paradigm of a philosophical school. After all, it was an actual group of philosophers and scientists who organized themselves into a regular discussion group and published a manifesto telling what their *Weltanschauung* was and who were their friends and adversaries. And the core of their doctrines can apparently be summarized by a few theses or doctrines (consult any textbook). But if you were to ask serious historians, they would tell you a rather more intricate story. For instance, Juha Manninen has argued in a study of the emergence of the Circle that "the features common to the Circle are to be found in a successful institutionalization and attitudes related to it, rather than in any set of collectively accepted and developed theses" (2002, 101). He argues further:

> The Vienna Circle was a process which involved a wide spectrum of sometimes conflicting ideas. The process never took a final shape. To understand the continuities and discontinuities of the Circle, we have to consider its individual members and the wider social interaction. The most dramatic manifestation of this is to be found in their views on language, which were subject to continual revision.[12]

[12] Manninen (2002, 103). Translation by AK.

Analogous points apply to philosophy in the post-war Oxford. Paul Grice, who himself worked there in the 1950s, has emphasized that there were no dogmas uniting Oxford-philosophers and that the only position accepted – with a varied measure of enthusiasm – was that philosophical thinking must be founded on "a careful examination of the detailed features of ordinary discourse", a view that implied nothing definite about the relationship between linguistic phenomena and philosophical theses.[13] Basically, Grice explains, there were two reasons why Oxford-philosophy had the appearance of a philosophical school. First, it was associated with a loose social structure, 'The Play Group', which convened for discussions on Saturday mornings (similar to but presumably looser than Moritz Schlick's famous Thursday evening Seminars). Second, Oxford-philosophy was rigidified into a "School" by its relentless critics like Russell and Gustav Bergmann. For them, talk of a "school" was a handy rhetorical device: once you define a philosophical school by reference to a few characteristic doctrines, you will have refuted all its members once you show the doctrines to be false.

8. I conclude that no feasible analytic definition can be given for 'analytic philosophy'. All such proposals, including Raatikainen's sophisticated revisionary definition, inevitably misrepresent the nature of the phenomenon under consideration. How, then, are we to proceed in our search of analytic philosophy?

My own proposal is that we should tackle the problem *historically*. We should stick to 'analytic philosophy' as a lexical item, thus obtaining a pre-theoretical and agreed-upon notion of *analytic tradition* as a starting-point. We can then put forth the following schematic characterization:

> [AP] Analytic philosophy consists of a series of connected phases—schools, movements, trends as much as individual philosophers—that together constitute the analytic tradition.[14]

[AP] is, indeed, schematic and does not say anything contentual about analytic philosophy. But it does make a point: *the unity of analytic philosophy is historical unity;* the category "analytic philosophy" is first and foremost a historical category.

We observed above that an average analytic definition fails to specify the genus that it seeks to define. Given [AP], we can say that the unity and continuity of analytic philosophy is (ultimately) supplied by the analytic *tradition*; and the very notion of tradition contains the idea that this unity and continuity need not be grounded just upon shared similarities or common features; it may be as much a matter of confrontations, changes of direction, etc. For instance, in the late 1940s, Oxford-philosophers did not regard any obscurantist metaphysics à la Heidegger as their *bête noir;* this

[13] Grice (1986, 49–51).

[14] I was pleased to find the following statement by von Wright: "The unity of the phenomenon [of analytic philosophy] I have tended to see in a chain of historically related, successive stages" (from a letter to Peter Hacker, quoted in Hacker 2016, 82.)

role was reserved for Carnap, the philosophical technologist.[15] Russell, on the other hand, argued that what he called the Oxford 'cult of common usage', among its other sins, was insincere, provided an excuse for laziness, and rendered philosophy trivial.[16] And yet, both Oxford-philosophy and Carnap and Russell all belong to the hardest core of analytic philosophy.

People like Føllesdal see heterogeneity as an existential threat to analytic philosophy. The fact is that heterogeneity belongs to the nature of the phenomenon. To begin with, it has been a standing element in analytic philosophy as long as it has been called by that name. Evidence for this claim was given above. Secondly, there is nothing exceptional about analytic philosophy in this respect. For instance, if we took a closer look at the phenomenological tradition, we would at once perceive similar heterogeneity (how does the realist phenomenology of Adolf Reinach and others relate to Husserl's endeavours?) And of course, this applies outside the sphere of philosophy, too. The analytic tradition is an intellectual formation, and as such, its structure, heterogeneity and dynamics could be readily compared, say, to the tradition of modernism in twentieth-century music, which exhibits an almost endless variety and is nevertheless a genuine and distinct phenomenon

Being historical, the concept of analytic philosophy cannot be defined. As Nietzsche observed in the second essay of *On the Genealogy of Morality:*

> All concepts in which a whole process is semiotically concentrated defy definition; only something which has no history can be defined.[17]

We can come to understand analytic philosophy by considering various aspects of the process that has been semiotically summarized in the concept; the key to 'analytic philosophy' is the analytical tradition, and understanding analytic philosophy is an essentially historical undertaking. But what are the "various aspects"? Briefly, they are (i) the various features that have been characteristic of the tradition; (ii) the inner dynamics of the analytic tradition; and (iii) the outer dynamics of the analytic tradition, that is, its relations to other relevant traditions.

Following Hans-Johann Glock's (2008) well-known analysis, we may say that, on this approach 'analytic philosophy' is at the same time a *family-resemblance* and *genetic-historical category*. This means two things. The unity of analytic philosophy is, first of all, a matter of "various resemblances" which "overlap and criss-cross like the similarities between the members of a family", to use Wittgenstein's language from *Philosophical Investigations*, § 67. The analytic tradition consists of distinct and different phases, between which there are similarities or resemblances, without there being any single feature or a group of features that should run through all these stages. Considering similarities alone, however, we would soon find ourselves outside

[15] See Ryle (1949).

[16] Russell (1953).

[17] Nietzsche (1887, *Second Essay*, § 13).

the analytic tradition. Therefore, and this is the second factor behind unity, similarities must be tied to a *particular historical tradition*. It's in the context of this tradition that we are to consider the similarities, and membership in this tradition is what makes a philosopher 'analytic' in the relevant sense. The essential point is that family-resemblances and membership in a particular tradition only work together, so that 'analytic philosophy' cannot be a purely historical and genetic concept, either. Merely considering who influenced whom and who was influenced by whom would soon take us outside analytic philosophy; lines of influence do not follow borders of traditions. Practically all European philosophers in the period between 1830 and 1930 were influenced by Kant, but that does not suffice to make them 'Kantian', not even 'neo-Kantian'.

This two-pronged approach is not fully satisfactory, however. The problem is that talk of 'family-resemblance' tends to obfuscate the *diachronic* side of the matter. The notion of *tradition* has *temporal continuity* and *change* as its key elements, but when a concept is said to be a family-resemblance concept, that is usually just a synchronic claim about taxonomy and classification; as when it is said that "things in the extension of a family-resemblance concept are brought together, not by any single feature that is common to all of them, but by a group of overlapping similarities". To repeat, the unity of the analytic tradition is primarily historical; and we should add, it's the unity of a *living tradition*. Mere features do not work here very well. They are static, supposedly repeated and transmitted within a tradition, whereas a living tradition is one that changes over time; a single so-called 'feature', moreover, may in fact cover several different, sometimes even opposite instances.

To do justice to analytic philosophy, we have to make these features *dynamic* and consider them as characteristics of a living tradition; we have to see analytic philosophy itself as "an historically extended, socially embodied argument", to quote Alasdair MacIntyre's well-known definition of tradition.[18] As MacIntyre also points out, a tradition has an outside as well as an inside: tradition is maintained and transformed by internal, interpretative debates (*internal conflicts*), and also by *external conflicts* with critics and enemies.[19] We have already met this notion of tradition. It is not very natural to call the Vienna Circle a "tradition" (it did not live long enough to developed into one); rather, it was a "school" or, better, a "movement". And yet, we saw an eminent historian arguing that even the Circle and, indeed, its so-called logical positivism ought to be considered as a "socially embodied argument", not as a set of fixed doctrines.

9. A good deal ought to be said to render the message of the previous section more transparent and convincing. Here I can do no more than draw the reader's attention to a few salient points about "features", as explained above. I shall use *linguistic philosophy* (and its cousin, *the linguistic turn*) as illustration. In the past, people were wont to use such phrases to explain the very idea of analytic philosophy. We know

[18] MacIntyre (1984, 222).

[19] MacIntyre (1988, 12).

now that this will not do, not even in a definition of classical, hard core analytic philosophy. Many philosophers in the analytic tradition, though, have shared the very general conviction that *language matters to philosophy*. And there is no doubt that an increasing attention to matters involving language was relevant to the emergence, evolution and transformations of the analytic tradition. But we have to ask: *why and in what way?*

Linguistic philosophy in this minimal and abstract sense is not a single phenomenon: in the analytic tradition, there have been many ways and many why's behind its linguistic turns. A rough typology distinguishes *three types*, all of them diachronic and dynamic (that is, historical).

Type 1 linguistic philosophy (LP-1) is the most radical one. According to it, philosophical problems, theories, theses, etc. are inextricably married to confusions and misunderstandings about language and how it works. Unsurprisingly, how this is supposed to come about depends on what philosophical phenomenon is at stake and what aspect of language is connected to it and how. To illustrate, Wittgenstein held throughout his career that philosophical problems owe their existence to "our misunderstanding the logic of our language"; but as he understood this logic differently at different times, the diagnosis in fact changed over time. In the *Tractatus*, "the logic of our language" is a deeply metaphysical matter (although the metaphysics is hidden and is officially not there at all), whereas in his later thought, beginning with the *Blue Book* of the mid-1930s, it was connected with a completely different set of ideas. Or think of Gilbert Ryle. In the early 1930s, he gave a somewhat simplistic account of philosophical mistakes as based on the notion of "systematically misleading expressions" (Ryle 1932); then, in the *Concept of Mind* (1949), he formulated an intriguing diagnosis of how the Cartesian theory of mind comes about when we misconstrue the logical geography of our mental language.

Type 2 linguistic philosophy (LP-2) is less radical. LP-2 people think that genuine philosophical problems exist and need not be based on confusions. They may think, for instance, that philosophical investigations are conceptual in nature and that concepts and conceptual distinctions are tied down to language. In addressing their problems, LP-2 people wield "a linguistic method". For example, J. L. Austin, along with many kindred spirits, argued that concepts live in our language, in "our common stock of words", which therefore embodies all the distinctions and connexions our ancestors have found worth drawing "in the lifetimes of many generations" (1961, 130). The crucial point is this: "When we examine what we should say when, what words we should use in what situations, we are looking again not *merely* words [...] but also the realities we use the words to talk about: we are using a sharpened awareness of words to sharpen our perception of, though not as the final arbiter of, the phenomena" (*ibid.*).

The distinction between LP-1 and LP-2 is not very sharp. For instance, how should we classify the Carnap of his syntactic phase, who held that genuine philosophical problems do not exist? This sounds like LP-1, but he also held that once all the relevant confusions have been eliminated, there remains the hard, scientific core of philosophy, which is logic. In this way, philosophical questions sort of disappear, as their place is taken by genuinely scientific questions about the proper formulation of the language of science (Carnap 1934). That the distinction should be vague, though, is only to be expected: we are here concerned with actual philosophers' actual thoughts and their contours, and not with ideal structures with sharp delimitations.

Type 3 linguistic philosophy (LP-3) has been present in the analytic tradition ever since its inception. Unlike Types 1 and 2, LP-3 does not see the questions, problems, subject matter or methods of philosophy as essentially linguistic: philosophy is about the real world, and "goes to the things themselves"; LP-3is just the awareness that philosophers must become conscious of the workings of language, or the ways of meaning, as this is a necessary condition of all valid philosophizing. Now, I am inclined to say that if by "the linguistic turn" we just mean an acceptance of LP-3, then it has, indeed, been a key characteristic of the analytic tradition – but it would still not be a distinguishing feature, because one can advocate LP-3 without thereby becoming an analytic philosopher.[20]

Russell and Moore are supreme examples of analytic philosophers who took the linguistic turn in the sense of LP-3. They began their careers as analytic philosophers with a resolute *denunciation* of the relevance of language to philosophy. Then, however, came a growing awareness that symbols and meaning are not as transparent as they had assumed at first. The first fruit of the new awareness was Russell's theory of definite descriptions, which Moore, too, came to accept. Russell, then, delved deeper into how symbols mean (*The Philosophy of Logical atomism* is mostly about this), not because this was what philosophy was about but because he found out that misunderstandings about "symbolism" were the veritable treasure trove behind much of traditional philosophy; here we perceive a certain overlap between LP-1 and LP-3, but they nevertheless remain distinct.

In Russell's case, there were other exciting developments, which took place as direct consequences of LP-3, including a sort of naturalistic turn. The phenomenon of meaning itself began to take on new philosophical importance for him, and since meaning, he now thought, was largely a matter of psychology and physiology, this brought about a more general change in his philosophical perspective. (You get a picture of this if you first read Russell's *The Problems of philosophy* (1912) and then his *An Outline of Philosophy* (1927), two books that stand so far apart that they were clearly written by two distinct philosophers).

Moore took a linguistic turn that probably, in the end, took him beyond LP-3. Methodologically, his version of analytic philosophy started from "transparency of appearing", as we may call it; the objects of philosophical analysis, he held, were *propositions*, or meaning structures which are independent of our minds but whose constituents and composition are something that we can become conscious of; at the end of the day, then, we just have to *see* that something is thus and so, and not some other way (Butler's maxim, which was the motto of *Principia Ethica*). When

[20] Franz Brentano would be an exciting early example (see Aho 1990).

this method turned out to be rather too simplistic (Russell's influence), Moore had to come up with a new one, and here Common Sense truisms and the inspection of actual linguistic usage become the benchmark. It is likely that Moore ended up being a linguistic philosopher in the stronger sense of LP-2; but at any rate he got there *via* LP-3.

10. So much for typology. You might raise a question at this point: What features should we include in a characterization of the analytic tradition? My preferred answer is: *any feature that a serious historian considers worthwhile*. It may be a big feature, like 'linguistic philosophy', one that runs through much of the analytic tradition. But it may be a small one too. The important point is that features *are not really meant to be typological at all but explanatory.* [21] We, as 'serious historians', want to understand the analytic tradition and explain things within it as well as about it, that is, at different levels of granularity, as they say: individual philosophers, interaction between individual philosophers, groups, schools, movements, and maybe entire segments of the tradition. This, indeed, is my main message.

Finally, I mention a special virtue of the present notion of a feature: it helps us see the analytic tradition as a broad intellectual movement. Specifically, it shows early analytic philosophy to have been so much more than just the handful of (male) names that make up the standard story, as in Soames (2014, 2018).[22] To be sure, Soames has his reasons for adopting a narrow perspective on the analytic tradition and its evolution. He is a philosopher who cares about what he takes to be *progressive* in contemporary analytic philosophy, and also about the past of such progressive elements. He would not really care about the analytic tradition as an intellectual movement; studying it only ever leads to endless contextualizations, from which no philosophical lessons can be derived. This, I think, would be wrong on several counts, but an elaboration must be preserved for another occasion. Here my concern has been with a preliminary investigation of 'analytic philosophy'.

[21] Cf. here Kremer (2013).

[22] As Janssen-Lauret (2022, Chapter 1) points out, this, indeed, remains a common blind spot.

Bibliography

Aho, Tuomo (1990): ’Brentanon suhteesta fenomenologiaan’, in M. Kosonen (ed.) *Phenomenology/Fenomenologia. Proceedings of the Symposium on Phenomenology in Jyväskylä 18.5.1988,* Julkaisu 43, Jyväskylän yliopisto, 72–83.

Austin, John (1961): ’A Plea for Excuses’, in J. L. Austin, J. O. Urmson & G. J. Warnock (eds.), *Philosophical Papers*, Clarendon Press, 123–152.

Ayer, Alfred (1936): ‘The Analytic Movement in Contemporary British Philosophy’, in *Actes du Congrès International de Philosophie Scientifique, Sorbonne, Paris 1935*, *Facs VIII,* Hermann & C^{le}, Éditeurs, 53–61.

Beaney, Michael (ed.) (2013): *The Oxford Handbook of the History of Analytic Philosophy.* Oxford University Press.

Black, Max (1939): ‘Relations Between Logical Positivism and the Cambridge School of Analysis’, *Erkenntnis* 8: 24–35.

Black, Max, John Wisdom and Maurice Cornforth (1934): ‘Symposium: Is Analysis a Useful Method in Philosophy?’, *Aristotelian Society Supplementary Volume* 13(1): 53–118. URL = https://doi.org/10.1093/aristoteliansupp/13.1.53

Carnap, Rudolf (1993 [1934]): *The Unity of Science, translated with an Introduction by M. Black*, Thoemmes Press, originally published by K. Paul, Trench, Truebner & Co, edited by Max Black.

Collingwood, Robin George (1933): *An Essay on Philosophical Method*, Clarendon Press.

Dummett, Michael (1993): *The Origins of Analytical Philosophy*, Harvard University Press.

Duncan-Jones, Austin Ernest (1937): ‘Symposium: Does Philosophy Analyse Common Sense?’, *Proceedings of the Aristotelian Society Supplementary Volume* 16, 139–161.

Føllesdal, Dagfinn (1997): ‘Analytic Philosophy: What is it and why should one engage in it?’, in H.-J. Glock (ed.), *The Rise of Analytic Philosophy*, Oxford: Blackwell, 1–16.

Frost-Arnold, Greg (2017): ‘The Rise of ‘Analytic Philosophy’: When and how did philosophers begin calling themselves ‘analytic philosophers?’’, in S. Lapointe and C. Pincock (eds.), *Innovations in the History of Analytic Philosophy*, Palgrave MacMillan, 27–67.

Glock, Hans-Johann (ed.) (1997): *The Rise of Analytic Philosophy*, Blackwell Publishing House.

Glock, Hans-Johann (2008): *What is Analytic Philosophy?*, Cambridge University Press.

Grice, Paul (1986): ‘Reply to Richards’, in R. E. Grandy and R. Warner (eds.), *Philosophical Grounds of Rationality*, Clarendon Press, 45–106.

Hacker, Peter (2016): ‘An Epistolary Friendship’, in G. Meggle and R. Vilkko (eds.) *Georg Henrik von Wright’s Book of Friends,* Acta Philosophica Fennica 92: 75–93.

Hylton, Peter (1990): *Russell, Idealism and the Emergence of Analytic Philosophy*, Oxford University Press.

Janssen-Lauret, Frederique (2022): *Susan Stebbing*, Cambridge Elements: Elements on Women in the History of Philosophy, Cambridge University Press.

Kremer, Michael (2013): ‘What is the Good of Philosophical History?’ in E. Reck (ed.) *The Historical turn in Analytic Philosophy*, New York, NY: Palgrave-Macmillan, 294–325.

MacIntyre, Alasdair (1984): *After Virtue,* 2nd edition, Notre Dame University Press.

MacIntyre, Alasdair (1988): *Whose Justice? Which Rationality?,* Notre Dame University Press.

Manninen, Juha (2002): ’Uuden filosofisen liikkeen ja sen manifestin synty’, in I. Niiniluoto and H. J. Koskinen (eds.), *Wienin piiri,* Gaudeamus, 27–128.

Mater, Anat (2017): ‘Dummett’s Dialectics’, in A. Preston (ed.), *Analytic Philosophy: An Interpretive History*, Routledge, 254–268.

Nagel, Ernest (1936a): ‘Impressions and Appraisals of Analytic Philosophy in Europe I’, *Journal of Philosophy* 33(1): 5–24

Nagel, Ernest (1936b): ‘Impressions and Appraisals of Analytic Philosophy in Europe II’, *The Journal of Philosophy* 33(2): 29–53.

Nietzsche, Friedrich (2006 [1887]): *On the Genealogy of Morality*, Cambridge Texts in the History of Political Thought, edited by K. Ansell-Pearson, translated by C. Diethe, Cambridge University Press.

Raatikainen, Panu (2001): ’Mitä oli analyyttinen filosofia?’, *Ajatus* 58: 189–217. URL = https://philpapers.org/archive/RAAMOA.pdf

Raatikainen, Panu (2013): ‘What was analytic Philosophy?’, *Journal for the History of Analytical Philosophy* 2(2): 11–27. URL = http://jhaponline.org/jhap/article/view/18/17

Reck, Erich H. (ed.) (2013): *The Historical Turn in Analytic Philosophy*, Palgrave MacMillan.

Russell, Bertrand (1922): “Analytic and Synthetic Philosophers”, review of *Philosophical Studies* by G. E. Moore and *The Misuse of Mind: A Study of Bergson’s Attack on Intellectualism* by Karin Stephen, *The Collected Papers of Bertrand Russell*, Vol. 9: *Essays on Language, Mind and Matter 1919–26*. Edited by J. G. Slater. Unwin Hyman. 1988, 406–410. First published in *The Nation and the Athenaeum* 31 (15 July 1922), 538–539.

Russell, Bertrand (1953): ‘The Cult of ‘Common Usage’’, *The British Journal for the Philosophy of Science* 3(12): 303–307. URL = https://doi.org/10.1093/bjps/iii.12.303

Ryle, Gilbert (1949): ‘Discussion: *Meaning and Necessity*’, *Philosophy* 24(88): 69–76. URL = http://www.jstor.org/stable/3747236

Ryle, Gilbert (1971 [1932]): ‘Systematically Misleading Expressions’, reprinted in G. Ryle, *Collected Essays 1929—1968* (*Collected Papers*, vol. 2), London: Hutchinson, 39–62.

Skorupski, John (2013): ‘Analytic Philosophy, the Analytic School, and British Philosophy’, in Michael Beaney (ed.), *The Oxford Handbook of The History of Analytic Philosophy*, Oxford, England: Oxford University Press, 298–317.

Sluga, Hans (1980): *Gottlob Frege*, London: Routledge & Kegan Paul.

Sluga, Hans (1997): ’Frege on Meaning’, in H.-J. Glock (ed.) *The Rise of Analytic Philosophy*, Blackwell Publishing House, 17–34. URL = https://doi.org/10.1111/j.1467-9329.1996.tb00160.x.

Soames, Scott (2014): *The Analytic Tradition in Philosophy, Volume 1: The Founding Giants.* Princeton University Press.

Soames, Scott (2018): *The Analytic Tradition in Philosophy, Volume 2: A New Vision.* Princeton University Press.

Stebbing, L. Susan (1938—1939) “Some Puzzles About Analysis”, *Proceedings of the Aristotelian Society*, New Series, vol. 39, Oxford University Press, 69–84. URL = https://www.jstor.org/stable/4544320

Wisdom, John (1931): *Interpretation and Analysis in Relation to Bentham’s Theory of Definition,* London*:* Kegan Paul, Trench, Trubner & Co.

Wittgenstein, Ludwig (1976 [1953]): *Philosophical Investigations,* translated by G. E. M. Anscombe, Basil Blackwell.

von Wright, Georg Henrik (1993): ‘Analytic Philosophy, A Historico-Critical Survey’, in Georg Henrik von Wright, *The Tree of Knowledge and Other Essays*, E. J. Brill, 25–52.

von Wright, Georg Henrik (2000): ‘Philosophy – A Guide for the Perplexed?’, in Daniel O. Dahlstrom (ed.), *Contemporary Philosophy,* The Proceedings of the Twentieth World Congress of Philosophy, Volume 8. Philosophy Documentation Center. Bowling Green State University, 275–293.

4
A categorical theory of truth[1]

Juhani Yli-Vakkuri & Zachary Goodsell

Introduction

Tarski’s method for defining truth for languages of finite order is generally understood to be his most important contribution to semantics. Tarski sets a precise standard for a definition of truth to be ‘adequate’, and he proves that definitions constructed by his method meet it. The standard is *Convention T*:

Convention T. *A formally correct definition of the symbol ‘Tr’, formulated in the metalanguage, will be called an* adequate *definition of truth if it has the following consequences:*

(α) all sentences which are obtained from the expression ‘$x \in Tr$ if and only if p’ by substituting for the symbol ‘x’ a structural-descriptive name of any sentence of the language in question and the symbol ‘p’ the expression which forms the translation [[2]] *of this sentence into the metalanguage; (β) the sentence ‘for any x, if $x \in Tr$ then $x \in S$’ (in other words ‘$Tr \subseteq S$’).* (Tarski 1955[1933]: 188.)

A ‘formally correct’ definition of the symbol ‘Tr’ is a sentence of the form

[1] Thanks to Beau Madison Mount, Peter Fritz, Volker Halbach, John Hawthorne, Lavinia Picollo, Tim Williamson, Dan Waxman, and the audience at TimFest: A Conference in Honor of Timothy Williamson at Magdalen College, University of Oxford on August 30th–July 1st, 2023 for helpful comments.

[2] We depart from Tarski in assuming that the object language is included in the metalanguage, so the translation of any object language sentence in the metalanguage will be just that sentence itself.

Yli-Vakkuri, J. & Goodsell, Z. (2025). A categorical theory of truth. In A. Laitinen, M. Keinänen, J. Reinikainen & A. Honkasalo (eds.). *Language, Truth, and Reality.* Tampere University Press. (pp. 49–58).
https://doi.org/10.61201/h5fbdk64

$$\mathrm{Tr} = \theta,$$

where θ is a predicate of the metalanguage in which 'Tr' does not occur. Here, the word 'definition' is reserved for such identity sentences. Conditions (α) and (β) concern the consequences of such an identification in the metatheory. (α) requires that every instance of the T-schema for the object language be derivable from the definition. So, if the object language includes, for example, the sentence '1 + 1 = 2' then the adequacy of a definition requires the sentence

$$\text{'}1 + 1 = 2\text{'} \quad \in \mathrm{Tr} \text{ if and only if } 1 + 1 = 2$$

or, in presently preferred notation, the sentence

$$\mathrm{Tr}\,\overline{1 + 1 = 2} \leftrightarrow 1 + 1 = 2$$

to be derivable from it in the metatheory. 'S' is Tarski's symbol for *sentence of the object language*, so condition (β) requires that in the metatheory we can prove by means of the definition that only sentences of the object language are true: 'in other words', as Tarski puts it,

$$\mathrm{Tr} \subseteq \mathrm{S} \qquad (1)$$

—a truism, since 'Tr' is simply an abbreviation for 'true sentence of the object language'. Let us call the class of metalanguage sentences comprising (1) together with all instances of the T-schema 'T'. A formally correct definition of truth, then, is deemed adequate by Convention T iff T is derivable from it in the metatheory.

Against convention T

Tarski's method for constructing definitions of truth is clearly a significant achievement, because the definitions constructed using his method are enormously fruitful, as he showed. However, the significance of Convention T is not as clear. By general agreement, T captures one important aspect of truth, but it is obvious that T does not include every generalization—not even every 'obvious' generalization, such as 'every sentence or its negation is true', that we expect to be able to prove in a good theory of truth. As Tarski himself was the first to point out, not even the theory that results from adding T to his preferred metatheory, which is (n+3)rd-order syntax formulated in a language that includes the object language, where n is the order of the object language, includes such obvious generalizations.

A policy of accepting definitions of truth based on whether they satisfy Convention T either fails to vindicate Tarski's definition or overgenerates. A policy of merely accepting *some* definition or other that can be proved adequate in the sense of Convention T fails to vindicate Tarski's definition, since in any ω-incomplete theory with an adequate definition, there is another adequate definition which is not provably equivalent. On the other hand, a policy of accepting *all* definitions that can be proved adequate in the sense of Convention T does vindicate Tarski's definition, but it also requires going far beyond Tarski's definition, by requiring us to accept a theory that, if consistent, is not recursively enumerable.[3] Such theories cannot be presented by a recursive set of axioms and rules, so are of limited use.

The reason why we should accept Tarski's definition of 'Tr', if we have any reason to accept it at all at present, is the fruitfulness of that definition, not that the definition satisfies Convention T. The derivability of is only a criterion of *minimal* adequacy for definitions of truth: we can rule out definitions that don't satisfy it (in the sense of not accepting any such definition, not in the sense of accepting its negation), but adopting a policy of accepting any definition that satisfies it, or even every definition that can be proved to satisfy it, would be either unmotivated or impossible to follow (insofar as it is impossible to accept a theory that is not recursively enumerable).

As criteria of minimal adequacy go, it is unclear why such a thing is needed, and it is also unclear why, supposing that such a thing is needed, Convention T should be it. Why should we not also require the derivability of compositional principles such as the principle

$$\forall x \in \mathrm{S}.\ (\mathrm{Tr}(\ulcorner\neg\urcorner \frown x) \leftrightarrow \neg\mathrm{Tr}x)$$

which says that every sentence or its negation is true? A definition of truth that does not satisfy this condition would not be deemed minimally adequate, so, it seems, our criterion of minimal adequacy should be at least this strong. But it is easy to come up with further desirable theorems. A survey of the literature on axiomatic theories of truth[4] will turn up a large number of attractive combinations, and it is unclear why the derivability of any of them should be designated *the* condition of minimal adequacy for a definition of truth, if such a thing is needed at all.

Categoricity

It would be preferable to avoid Convention T altogether and to formulate an acceptable theory of truth in which Tarski's definition can simply be proved (rather than proved "adequate", whatever adequacy might be). Such a theory would be, in the terminology of Tarski's 1933 paper (1956: 257), a *categorical* theory of truth.[5]

[3] *Proof sketch.* Let M be any recursively axiomatizable metatheory such that 'Tr = ' is proved to satisfy Convention T, such that we can also prove in M

For all $x \in S$, either θx or $\theta(\ulcorner\neg\urcorner \frown x)$.

Then let θ^M be the sentence:

$\lambda x \in S.\exists n$ (the length of x is n and and a contradiction cannot be derived in M in fewer than n steps).

The adequacy of 'Tr = θ' can be proved in M. So by Convention T we should accept '$\theta = \theta^M$', from which the consistency of M is derivable. The same will go for every recursively enumerable extension of M. □

[4] See Halbach 2011 for review.

[5] 'Categoricity' is also commonly used for a semantic rather than the present proof-theoretic property of theories. In contemporary model theory, a theory is called categorical if all of its models—including those with deviant interpretations of the quantifiers—are isomorphic. Tarski's theory of truth is not categorical in this sense (only negation-complete theories are). The term 'categorical' originates with Veblen 1904, where

Definition 1 (Categoricity)**.** A theory Γ is *categorical* with respect to the class of constants $\Delta = \{c_1, c_2, \ldots\}$ if and only if, for any theory Γ' obtained by replacing the constants in Δ by previously unused constants $\Delta' = \{c'_1, c'_2, \ldots\}$,

$$\Gamma, \Gamma' \vdash \overline{c_i = c_i'}$$

for each $c_i \in \Delta$.

Of interest here is the case where $\Delta = \overline{\{\mathrm{Tr}\}}$.

Tarski agreed that a categorical theory of truth would be preferable to Convention T, but he resorted to Convention T because he thought it would not be possible to formulate an acceptable such theory:

> [. . .] it seems natural to require that the axioms of the theory of truth, together with the original axioms of the metatheory, should constitute a categorical system. It can be shown that this postulate coincides in the present case with another postulate, according to which the axiom system of the theory of truth should unambiguously determine the extension of the symbol 'Tr' which occurs in it, and in the following sense: if we introduce into the metatheory, alongside this symbol, another primitive sign, e.g. the symbol 'Tr′' and set up analogous axioms for it, then the statement 'Tr = Tr′' must be provable. But this postulate cannot be satisfied. For it is not difficult to prove that in the contrary case the concept of truth could be defined exclusively by means of terms belonging to the morphology of language [i.e., syntax], which would be in palpable contradiction to Th. I.[6] (Tarski 1956[1933]: 257)

As the authors interpret this passage, Tarski is making a mistake. Tarski seems to be saying that any categorical set of axioms for 'Tr' from which T is derivable in the metatheory is inconsistent in the metatheory. But this simply cannot be the case. For consider the theory whose sole novel axiom is Tarski's own definition,

$$\mathrm{Tr} = \mathrm{Tr}_{\mathrm{Tarski}}, \quad (2)$$

where '$\mathrm{Tr}_{\mathrm{Tarski}}$' abbreviates the complex truth predicate that Tarski showed us how to construct out of the object language, the constants of syntax, and quantifiers and variables of orders higher than those found in the object language. (2) is certainly categorical in Tarski's sense, and, as Tarski showed, T is derivable from (2) in the metatheory. And adding this one axiom cannot make the metatheory inconsistent

it is used for a geometrical theory which has only one model up to isomorphism where the non-geometrical constants are given the intended interpretation. Generalizing Veblen's geometric notion to semantics, we find that Tarski's original truth theory consisting of the sentences described in conditions (α) and (β) *is* categorical in the sense that among interpretations that agree with the intended one on all constants besides 'Tr', there is only one on which the theory comes out true. However, this fact could not possibly serve to justify Tarski's definition, since it takes Tarskian semantics for the metalanguage for granted.

[6] Th. I is Tarski's undefinability theorem.

(nor could adding any definition of an otherwise unused expression, since such additions yield conservative extensions of the theories to which they are added). Of course, in justifying Tarski's definition it would not do to adopt that same definition as an axiom, but the point is that there is no in-principle problem with searching for a theory of truth that is both categorical and consistent.

A categorical theory of truth

Preliminaries

We work within finite-order fragments of Henkin's 1950[7] extensional higher-order logic plus standard axioms of syntax (Tarski used finite-order fragments of what he called the *calculus of classes*, which is essentially Henkin's system formulated with relational rather than functional types, and which he obtained by simplifying the extensional fragment of the *Principia Mathematica* system). A range of systems would work equally well for present purposes. In particular, systems to which Tarski's original hierarchy of definitions of truth satisfying Convention T can be adapted, and which are *intensional* in the sense that provable coextensionality of properties, relations, and propositions (if propositional quantification is included, as it is in Henkin's system) suffices for identity,[8] will generally also permit a modification of the theory presented here that is also categorical and a conservative extension of the system in question. For definiteness, we will ignore possible variations and adhere to the system of *extensional n^{th}-order syntax*, S^n, which has the following features.

- Simple functional types (as in Church 1940) with base types e and t and functional types—$(\sigma\tau)$ for any types σ and τ—of order $n + 2$ or less, where the order of a type is defined recursively as follows:

  – $O\ (e) = O\ (t) = 1$ and
  – $O\ (\sigma\tau) = \max \{O\ (\sigma) + 1, O\ (\tau)\}$ (i.e., the order of a functional type is one plus the order of the type of its argument, or is the order of the type of its output, whichever is greater).

- Infinitely many variables of each type of order n or less.
- λ-abstraction with the usual axioms asserting the substitutivity of β-equivalent terms (Church 1940).

[7] Henkin's system is what we get when we add the axiom of Boolean extensionality to Church's 1940 system—an addition considered and rejected by Church on the c.

[8] In contrast with *hyperintensional* systems like those of *Principia Mathematica* and Church 1940, where provable coextensionality is not sufficient for identity (Church's system lacks the axiom of Boolean extensionality). Notice that intensionality does not require that coextensionality implies identity, but only the substitutivity of provably coextensional terms (as in typical modal logics).

- Boolean connectives with classical propositional logic.
- Universal and existential quantifier symbols '$\forall_\sigma$', '$\exists_\sigma$' of type $(\sigma t)t$ for each type σ of order n or less, with classical quantifier logic at each type.
- Identity symbols '$=_\sigma$' of type $(\sigma t)\ (\sigma t)t$ for each type σ of order n or less, obeying the reflexivity of identity and Leibniz' law at each type.
- The axiom of Boolean extensionality,

$$\forall p\, q.\, ((p \leftrightarrow q) \to (p \to q))$$

- Axioms of function extensionality,

$$\forall f g^{\sigma\tau}.\, \big(\forall x^{\sigma}.\ (fx = gx) \to (f = g)\big).$$

- Axioms of choice,

$$\exists f^{(\sigma t)\sigma}.\, \forall g^{\sigma t}.\, (\exists g \to g(fg)).$$

- Standard axioms of syntax, asserting roughly that the strings are the free semigroup generated by some alphabet of characters. For definiteness, the name of a string will have type e (so the name of the class of strings, 'String', has type et).

A *finite signature* will be a finite list of typed constants separated by commas, and Σ_S will be the signature of syntax (which contains names for each character of the alphabet and a constant for concatenation of strings). Our object language, like Tarski's, will be $\mathscr{L}^n_\Sigma$ for an arbitrary order n and arbitrary finite signature Σ, which is the language described above but with the constants from Σ of order n included instead of the constants of syntax.

We will employ standard notational abbreviations for logic and syntax (e.g., '$\forall x \in \alpha.(\dots)$' abbreviates '$\forall \lambda x.(\alpha x \to \dots$'), and will take for granted standard formalizations of complex syntactic notions like the class of sentences in $\mathscr{L}^n_\Sigma$ (symbolized '$\mathscr{L}^n_\Sigma$') and provability in S^{n+3} (symbolized '$\mathrm{S}^{n+3}\vdash$'). The symbol 'Q' is used for the function which maps a string to its structural-descriptive name, and '$\frown$' for the concatenation function.

The F-schema

For the object language $\mathscr{L}^n_\Sigma$, the metalanguage Tarski uses to formulate his theories of truth is

$$\mathscr{L}^{n+3}_{\Sigma,\Sigma_s,\overline{\mathrm{Tr}}}$$

where as usual 'Tr' is the primitive truth predicate. Tarski's theory of truth is the list of sentences mentioned in conditions (α) and (β). Corresponding to condition (α) is an infinite list of sentences, one for each sentence φ of the object language:

T-schema$^{\varphi}$

$$\mathrm{Tr}\,\overline{\varphi} \leftrightarrow \varphi.$$

Condition (β) corresponds to a single additional sentence:

Sentential truth

$$\mathrm{Tr} \subseteq \mathscr{L}^n_\Sigma.$$

Call the class of all such sentences T^n_Σ (Convention T then says that a definition is adequate if every sentence of T^n_Σ can be proved from the definition).

Our metalanguage is, instead, $\mathscr{L}^{n+6}_{\Sigma,\Sigma_S,\overline{Tr}}$—the language of ($n$+6)th-order logic plus the constants of Σ and of syntax and the primitive truth predicate—and our metatheory is S^{n+6}—(n+6)th-order syntax. That is, our metalanguage and metatheory are what Tarski would use as metametalanguage and metametatheory for semantic theorising about the metalanguage. Our categorical theory of truth is given in its entirety by the following axiom schema where Φ may be replaced by any term of type *et* of $\mathscr{L}^{n+3}_{\Sigma,\Sigma_S}$:

Factivity of $S^{n+6} + \mathsf{T}^n_\Sigma(\Phi)$

$$\forall x \in \mathrm{String}.\ \left(\mathrm{S}^{n+3} + \mathsf{T}^n_\Sigma \vdash \left(\overline{\Phi} \frown \mathrm{Q}x\right)\right) \rightarrow \forall x \in\ \mathrm{String}.\ \Phi x.$$

Call the class of such sentences the *F-schema*, or F^n_Σ ('F' for 'Factivity'). The F-schema can be intuitively understood by way of example. One instance says that if the sentence 'if φ is a sentence then either it or its negation is true' can be derived in Tarski's minimal theory of truth for every string φ, then every string in fact has the property that if it is sentence then either it or its negation is true.

The F-schema is, in essence, a combination of highly plausible principles of closure and disquotation for *truth-in-*$\mathscr{L}^{n+3}_{\Sigma,\Sigma_S}$. The F-schema for a given predicate Φ can be decomposed into the following theses for a primitive notion of truth-in-$\mathscr{L}^{n+3}_{\Sigma,\Sigma_S,\overline{\mathrm{Tr}}}$ which we symbolize 'Tr^{n+3}':

Truth of $S^{n+3} + \mathsf{T}^n_\Sigma$ Every theorem of $S^{n+3} + \mathsf{T}^n_\Sigma$ is true-in-the-metalanguage (i.e., $\mathrm{Tr}^{\mathrm{n+3}}$).

$$\forall x.\ \left(\left(\mathrm{S}^{n+3} + \mathsf{T}^n_\Sigma \vdash x\right) \rightarrow \mathrm{Tr}^{\,n+3}x\right)$$

ω**-Closure of truth**$^\Phi$ If, for all sentences x, the application of the predicate Φ to the structural-descriptive name of x is true-in-the metalanguage, then the sentence '∀x ∈ String .Φx' is true-in-the-metalanguage.

$$\forall x \in \mathrm{String}.\ \mathrm{Tr}^{n+3}\left(\overline{\Phi} \frown Qx\right) \rightarrow \mathrm{Tr}^{n+3}\overline{\forall x \in \mathrm{String}.\ \Phi x}$$

T (out) schema for string quantification$^\Phi$ If the sentence '∀x ∈ String.Φx' is true-in-the-metalanguage, then ∀x ∈ String .Φx.

$$\mathrm{Tr}^{n+3}\ \overline{\forall \mathrm{x} \in \mathrm{String}.\ \Phi\mathrm{x}} \rightarrow \forall \mathrm{x} \in \mathrm{String}.\ \Phi\mathrm{x}$$

Proposition 1. *Every instance of* $\mathrm{F}^{\mathrm{n}}_\Sigma$ *can be derived from Truth of* $\mathrm{S}^{\mathrm{n}+3} + \mathsf{T}^{\mathrm{n}}_\Sigma$*, instances of* ω*-Closure of Truth*$^\Phi$*, and instances of the T-schema for String Quantification.*

In addition to being very plausible, the F-schema is categorical, and indeed proves Tarski's definition.

Theorem 2 (Categoricity). $\mathrm{F}^{\mathrm{n}}_\Sigma$ *is equivalent in* S^{n+6} *to* '$\mathrm{Tr} = \mathrm{Tr}_{\mathrm{Tarski}}$'*, where* Tr_{Tarski} *abbreviates what Tarski defined truth (in* $\mathscr{L}^n_\Sigma$*) to be.*

Proof. To derive the definition from the F-schema, it will suffice to show that each instance of

$$\mathrm{Tr}\overline{\varphi} \leftrightarrow \mathrm{Tr}_{\mathrm{Tarski}}\overline{\varphi}$$

is a theorem of $S^{n+3} + \mathsf{T}^{n}_{\Sigma}$. This holds because each instance of the T-schema is assumed for 'Tr' and is provable for $\mathrm{Tr}_{\mathrm{Tarski}}$.

To derive the F-schema from the definition, let $\mathrm{Tr}^{n+3}_{Tarski}$ be the Tarskian defined truth-predicate for the object language $\mathscr{L}^{n+3}_{\Sigma,\Sigma_S\overline{Tr}}$. Tarski shows that we can derive all instances of ω-Closure of Truth and the T (Out) Schema for String Quantification in S^{n+6} when 'Tr^{n+3}' is replaced by $\mathrm{Tr}^{n+3}_{Tarski}$. It is also easy to show that $S^{n+3} + \mathsf{T}^{n}_{\Sigma}$. is satisfied when and only when 'Tr' is interpreted as $\mathrm{Tr}_{\mathrm{Tarski}}$. □

Remark 1. *$S^{n+6} + F^{n}_{\Sigma}$, since it follows from a definition, is a conservative extension of* S^{n+6}.

Remark 2. *The F-schema is axiomatized by the single instance where Φ is*

$$\lambda y.\big(\mathrm{Tr}\, y \leftrightarrow \mathrm{Tr}_{\mathrm{Tarski}}\, y\big),$$

since this instance suffices to prove '$\mathrm{Tr} = \mathrm{Tr}_{\mathrm{Tarski}}$*' from which every other instance can be derived by Theorem 2.*

Relation to other categorical theories

The truth-definition '$\mathrm{Tr} = \mathrm{Tr}_{\mathrm{Tarski}}$', being a definition, is a conservative extension of S^{n+3}. It is also categorical, as previously mentioned. By contrast, adding the F-schema to S^{n+3} results in a non-conservative extension of S^{n+3} if S^{n+3} is consistent, because the F-schema implies the consistency of S^{n+3}. However, although the F-schema is consistency-theoretically stronger than '$\mathrm{Tr} = \mathrm{Tr}_{\mathrm{Tarski}}$', it is clearly unobjectionable from a Tarskian point of view. For anyone who adopts Tarski's unamended approach to truth also accepts S^{n+6}; they will regard S^{n+6} as the meta*meta*theory rather than as the metatheory, but they accept it just the same, and the result of adding the F-schema to S^{n+3} is a conservative extension of S^{n+6}. And the F-schema is not only unobjectionable on consistency-theoretic grounds from a Tarskian point of view; in a sense, accepting it already comes with the Tarskian approach: while those taking that approach do not make the F-schema part of their theory, in accepting $\mathrm{S}^{n+3} + \mathsf{T}^{n}_{\Sigma}$, they of course accept that whenever, 'For all strings *s*, 'Φ*s*" is derivable from these in their metametatheory, then, for all strings *s*, Φ*s* (they don't reject this, or suspend judgment on it).

Another categorical theory that has been discussed in the literature (first by Tarski himself immediately after he commits the mistake quoted above) is the closure of the Tarskian metatheory, S^{n+3}, under a syntactic ω-rule, which requires that when formulae

$$\Phi\overline{\gamma}$$

are provable for every string *γ*, then so is the formula

$$\forall \mathrm{x} \in \mathrm{String}.\ (\Phi \mathrm{x}).$$

As is widely known, theories closed under such a rule are either inconsistent or are not recursively enumerable, and here we are only considering formal (i.e., recursively enumerable) theories of truth.

References

Church, Alonzo (1940): 'A Formulation of the Simple Theory of Types', *Journal of Symbolic Logic* 5(2): 56–68. URL = https://doi.org/10.2307/2266170

Halbach, Volker (2011): *Axiomatic Theories of Truth*, Cambridge: Cambridge University Press.

Henkin, Leon (1950): 'Completeness in the Theory of Types', *Journal of Symbolic Logic* 15(2): 81–91. URL = https://doi.org/10.2307/2266967

Tarski, Alfred (1956 [1933]): *Logic, Semantics, Metamathematics*, Oxford: Clarendon Press.

Veblen, Oswald (1904): 'A System of Axioms for Geometry', *Transactions of the American Mathematical Society* 5(3): 343–384. URL = https://doi.org/10.2307/1986462

5
Putnam's transcendental arguments

Sami Pihlström

When I was invited to contribute an essay to this Festschrift honoring Professor Panu Raatikainen, a long-time friend and colleague whose contributions to both Finnish and international philosophy I admire very much, I not only immediately said yes but also knew that I should write on a topic related to Hilary Putnam's philosophy. This is because Panu and I have discussed Putnam's views (among, of course, many other things) since we first met in the early 1990s. I have vivid memories of the graduate seminar taught by Ilkka Niiniluoto at the University of Helsinki in fall 1991, which both Panu and I attended. Panu was, in fact, the opponent of my seminar presentation on Putnam's internal realism – a hotly debated topic in those days – and Ilkka, of course, was both his dissertation supervisor and mine. As far as I remember, my seminar paper focused on the ways in which Putnam criticized metaphysical realism and defended internal realism in some of his seminal work on these issues collected in *Realism with a Human Face* (Putnam 1990), a book that had just come out with fresh formulations of ideas he had developed since the late 1970s, and Panu's thoughtful critical comments raised fundamental issues challenging Putnam's epistemic conception of truth, in particular. What Panu's own presentation in the same seminar explored I unfortunately cannot recall.

I suppose Putnam has always been something like a philosophical hero for both Panu and me. However, we differ in our "favorite" Putnams. For Panu (if I am right), Putnam's philosophy reached its culmination in the 1970s when Putnam defended the causal theory of reference (with a canonical formulation in the famous 1975

Pihlström, S. (2025). Putnam's transcendental arguments. In A. Laitinen, M. Keinänen, J. Reinikainen & A. Honkasalo (eds.). *Language, Truth, and Reality*. Tampere University Press. (pp. 59–69).
https://doi.org/10.61201/h5fbdk64

article, “The Meaning of ‘Meaning’”),[1] functionalism in the philosophy of mind, and an influential version of scientific realism based on the “no miracles” argument – and then it was at least partly downhill from there on. While I also very much appreciate those lasting earlier achievements (among the many that Putnam reached but also self-critically reconsidered during his long career), I have always found the Putnam of the 1990s closer to my own philosophical temperament (borrowing a term from William James, one of the great old pragmatists, about whom Putnam wrote a number of important essays). This is the Putnam whose internal realism had already been established (throughout the 1980s) to the point of starting to fade out or merge into something he later came to call “commonsense realism” (with a significant touch of pragmatism), the Putnam who had finally (partly due to the influence of his wife Ruth Anna Putnam) recognized his crucial indebtedness to the pragmatist tradition, and had even started to see his own work as continuing it, and the Putnam who had become sharply critical of the fact-value dichotomy and even willing to contribute to the philosophy of religion, utilizing not only Kantian, pragmatist, and Wittgensteinian sources but also his inherited Jewish tradition.[2]

Putnam’s thought, clearly, is not the only philosophical interest Panu and I share,[3] but it is undoubtedly the point at which our philosophical concerns most explicitly converge. Although nowadays our paths unfortunately cross relatively infrequently, I am sure our discussions and disagreements on realism – a major Putnamian theme whose relevance of course extends far beyond Putnam’s work – in the 1990s were significant for both of us. These shared interests have also led us both to enormously appreciate the work of our teacher, Ilkka Niiniluoto; in fact, Panu and I co-edited (in collaboration with other colleagues) two Festschrifts for Ilkka when he turned 50 (in 1996) and 60 (in 2006). Not only Ilkka’s version of critical scientific realism in general but also his criticisms of Putnam in particular (cf. Niiniluoto 1999) have presumably had an equally lasting impact on Panu as they have had on me.

Putnam, realism, and conceptual relativity once again

One of Putnam’s best-known arguments for internal realism and against “metaphysical realism” is the so-called Carnapian world argument, which he developed in the 1980s (see Putnam 1987, 1990) but revisited in later work (see, e.g., Putnam 2004, 37–38, 78–84). This is also the key argument that Panu insightfully criticized in a paper in *Dialectica* (Raatikainen 2001). Panu’s criticism focuses on Putnam’s appeal to

[1] This essay is collected in Putnam 1975.

[2] For Putnam’s mature views on pragmatism and philosophy of religion, see Putnam 2008; Putnam and Putnam 2017. (I have discussed these aspects of his thought at some length elsewhere, e.g., Pihlström 2023, chapter 2.)

[3] I have, for example, repeatedly returned to Panu’s very helpful small book in Finnish on the philosophy of the human sciences (Raatikainen 2004), which I reviewed for the journal *Tieteessä tapahtuu* and which I continue to find highly relevant regarding, e.g., the issues of value-dependence vs. value-freedom in research (see also, e.g., Raatikainen 2006).

mereology, arguing that Putnam confuses languages and theories when suggesting that using the language of mereology commits us to the existence of mereological sums. I will briefly pursue this issue not by responding to Panu's paper in any great detail, nor by examining Putnam's central works – a lot has been said about them – but by referring to Putnam's recent posthumously published collection, *Philosophy as Dialogue* (Putnam 2022), which contains a number of highly interesting critical comments by Putnam on other philosophers' views and criticisms of him (albeit no response to either Panu or me).

The key to the Carnapian world argument is that when imagining a world of three individuals (x1, x2, x3), as soon as we allow the use of mereological language (enabling us to "count" the sums of individuals as individuals), we realize that there is no single privileged or definite answer to the question of how many objects there are in this mini-world. The "Carnapian logician" will say that there are three objects there, but the "Polish logician" employing mereology can claim that there are seven objects (x1, x2, x3, x1+x2, x1+x3, x2+x3, x1+x2+x3), or even eight (including the "null object"), in the "same" world. Panu pointed out that the mere use of mereological language as such does not yield this outcome; one has to theoretically postulate, in addition, that mereological sums exist (see Raatikainen 2001).

As sharp as Panu's response to Putnam is, I am (still) not entirely convinced that it succeeds in its criticism of what Putnam tried to demonstrate by means of his argument, viz., the phenomenon of conceptual relativity.[4] I don't think Putnam intended to simply claim that the mere use of the language of mereology (or any logic or language, for that matter) guarantees that there are certain objects that the language enables us to speak of. Clearly more is needed for existence than language. Rather, the availability of a certain language enables us to formulate theories that may postulate some kinds of objects rather than some others. Mereology enables us to postulate mereological sums if we wish, but it does not guarantee that such postulations are plausible or accurate.

In a reply to David L. Anderson's criticisms of his views (in 1992)[5] available in the posthumous collection, Putnam (2022, 108) maintains that "there is no fact of the matter as to whether numbers, or mereological sums, are objects or not [...] and no fact of the matter as to whether 'mereological sums exist'". If Panu intended to claim that the mere use of the language of mereology would, according to Putnam, fix the relevant ontology or bring about such a "fact of the matter", this seems to pre-empt Panu's charge.[6] In a response to Simon Blackburn (in 1994), Putnam – using phrases

[4] Raatikainen (2001) repeatedly speaks of "conceptual relativism", while Putnam consistently uses the expression "conceptual relativity", explicitly distinguishing between that phenomenon and any form of relativism. Panu's wording may be interpreted as suggesting that Putnam fails to adequately draw that distinction.

[5] In this paper, I have provided no bibliographical information on the contributions by other philosophers to which Putnam responds in the material collected in Putnam 2022; all the details are available in that volume.

[6] Admittedly, many of the formulations by Putnam over the years cited in Raatikainen (2001) *are* unclear and possibly misleading. Obviously, Panu made an important point simply by reminding us how important it is to distinguish between merely using a language and drawing up an ontology by that language-use.

such as "[i]f our ontology includes individuals but not mereological sums" and "if we adopt an ontology which includes mereological sums" (ibid., 124) – also seems to indicate that our merely having a language (e.g., mereology) is not the same thing as being committed to an ontology expressible by using that language. Certainly, the sheer use of the language of mereology does not miraculously create mereological sums into existence. The point of the example is different, namely, that there is no fact of the matter whether mereological sums exist because the very idea of there being such metaphysical facts of the matter (at some absolute or fundamental metaphysical level) is unclear at best and downright nonsensical at worst.

In other words, Putnam thus argues that it is, at least partly, a *conventional* matter whether mereological sums exist. But this is not to claim that it would be *merely* conventional, either; on the contrary, factuality and conventionality are deeply interwoven and interpenetrating.[7] The relativity of objecthood that the Carnapian world argument in Putnam's view demonstrates entails that existential expressions we use in our languages, such as "there are", "there exist", "there exists a", and "some", as well as their logical codification in the existential quantifier, "*do not have a single absolutely precise use but a whole family of uses*" (Putnam 2004, 37; original emphasis). In other words, the notion of an "object" is "inherently extendable", and no sense whatsoever can be made of the notion of a "totality of all objects" (Putnam 2022, 108). It is only by endorsing this relativity (or conventionality), and thus rejecting what Putnam earlier (e.g., 1981, 1990) called metaphysical realism, that we can truly make sense of our practice of making existence claims at all. Moreover, as far as I can see, this is something that Putnam never fundamentally reconsidered even after having given up "internal realism". I do not think he ever believed that the notion of a totality of all objects would make sense – even when he returned to something like commonsense or even metaphysical realism.

A transcendental argument against metaphysical realism?

My main aim is not, however, to defend Putnam's argument for conceptual relativity or his complex views on realism more generally but to examine the philosophical status of his argumentation in this context. What I am tentatively proposing is that we may interpret at least some of Putnam's central arguments as *transcendental arguments* in a loosely Kantian sense. Despite his obvious Kantian influences, Putnam himself never accepted those critics' views who suggested that his internal realism could be regarded as a version of Kantian transcendental idealism.[8] However,

[7] This is a major theme in, e.g., Putnam 1990 and Putnam 2004. I cannot go into any details here, though.

[8] In the editor's introduction to Putnam 1990, James Conant distinguishes between four major influences in Putnam's philosophy: Kant, the pragmatists, Wittgenstein, and Stanley Cavell. (Putnam also occasionally refers to the similarities between Kant and pragmatism, which I find important but will have to set aside here.) Taking the suggestion about Putnam's Kantian background seriously, I once made an explicit effort along these lines, wondering why he could not endorse a pragmatic transcendental rearticulation of his views, in a

I would like to suggest – continuing the dialogue not only with Panu but with Putnam himself, although he can no longer respond – that Putnam's above-described defense of conceptual relativity, in particular, might be seen as a transcendental argument.

The basic idea is this. Putnam's arguments may be interpreted as seeking to demonstrate that unless we endorse conceptual relativity (along the lines briefly explained above), we can make no sense of the idea that the "same" situation or portion of reality may be described in different ways, postulating different objects. However, we do have to make sense of this idea, because otherwise we cannot even find our practice of referring to objects intelligible at all. Without conceptual relativity, we end up with the ultimately unintelligible ideas of "self-identifying objects" and an ontologically pre-structured "ready-made" world.[9] In order to have a world of objects in the first place (something we may presumably take for granted), we must subscribe to conceptual relativity, because the very notion of objecthood can be made sense of only by accepting its "extendability".

However, conceptual relativity as such does not exhaust the matter. One of the most important moves in Putnam's reflections on his Carnapian world arguments and its significance was his response (in 2001) to Jennifer Case's articles on conceptual relativity. Putnam endorses Case's distinction between *optional* and *non-optional* languages and regrets not having made that distinction earlier. While we are free to employ or not to employ a language such as mereology, he tells us, "[w]e are not, given the material and social worlds in which we live, genuinely free not to quantify over tables and chairs" (Putnam 2022, 97). Putnam seems to be saying that the distinction between genuinely optional languages and parts of language that we cannot avoid employing is itself constitutive of our mastering a natural language (ibid.). He thus seems to argue that being able to distinguish between optional and non-optional cases of language-use is a necessary condition for the possibility of meaningful language-use. This can be regarded as a transcendental argument, albeit a pragmatist one, referring to our participation in linguistic practices and its necessary conditions. The upshot appears to be that a metaphysical realist subscribing to a metaphysics of a fixed set of mind- and language-independent objects and properties out there in a "ready-made" world cannot make sense of such participation.

Metaphysical realism can, then, (only) be transcendentally refuted – just like you have to adopt transcendental idealism in order to argue against transcendental realism in Kant's original formulation of transcendental philosophy (see Allison 2004 [1983]). To make that distinction is already to have taken a transcendental turn and thus to have embraced something like transcendental idealism. Of course, Putnam does not and cannot say this in so many words, due to his firm resistance to the transcendental vocabulary.

book symposium on Putnam 2004 (see Pihlström 2006). Putnam's (2006) response to my paper emphatically denied that any transcendental idealism could be read into his account. I further reflect on these issues in Pihlström 2009.

[9] One of Putnam's classical earlier papers on this issue, "Why There Is No Ready-Made World", is available in Putnam 1983.

The distinction between optional and non-optional languages is significant also because conceptual relativity, as Putnam acknowledges in his reply to Case, is a special case of the "wider phenomenon" of pluralism (Putnam 2022, 99). Here the examples are closer to the real world we are familiar with in both everyday and scientific experience than the Carnapian language vs. mereology example: for instance, the "contents" of a given situation or state of affairs, such as my room, can be described by using the scheme of everyday description speaking of desks and chairs and by using the scheme of fundamental physics, speaking of fields and particles. Putnam writes: "That we can use both of these schemes without being required to reduce one or both of them to some single fundamental and universal ontology is the doctrine of *pluralism*; and while conceptual relativity implies pluralism, the reverse is not the case." (Ibid., 99–100.)

Pluralism, in an ontologically relevant sense (and in my terms rather than Putnam's), is a transcendental condition for the possibility of human language-use as we know it: we must be able to use "these schemes", without a single absolute reductive ontology, in order to be able to engage in the practices we do engage in – that is, for example, to be able to simultaneously live in the world of the ordinary objects in a room and engage in research in theoretical physics (possibly in that very room). Pluralism itself thus operates within what we may call *transcendental pragmatism*, which analyzes the necessary conditions for the possibility of our engagement in our practices (and thus the necessary conditions for the possibility of something we take as given).

A transcendental argument for pragmatic pluralism (against metaphysical realism) need not, and should not, invoke optional languages and conceptual relativity in the way the Carnapian world argument does; Panu is right to point out that bringing a highly special conceptual scheme such as mereology to the picture creates new problems rather than solving any. Rather, this argument for pluralism most plausibly starts from non-optional natural languages and our unavoidable commitment to the (potential) plurality of ontologies that may be formulated within them.[10] As Putnam himself says, we are not free to avoid quantifying over tables and chairs, and it is this non-optionality (instead of optional languages like mereology) that a genuinely transcendental argument for pluralism should be grounded in, with full awareness of the pragmatic need to also employ highly specialized languages in scientific contexts.

What is at issue – even in the case of optional languages – is the *availability* of some ontology, such as mereology (the "existence" of mereological sums), rather than its actuality (given the use of the relevant language). Different ontological commitments are *enabled* by our choosing (within our practices, for pragmatic reasons) certain schemes. It is within the use of the relevant scheme that we can then further inquire into the matter and critically discuss the ontological postulations

[10] We may find interesting parallels between Putnam's and William James's pragmatic pluralisms, even though James's views hardly explicitly influenced Putnam's in this specific respect. On James and pluralism, see, e.g., Pihlström 2023, chapter 3.

we ought to, or ought not to, make. In this sense, ontology remains "internal" to our schemes and practices.

Realist philosophers like Panu may at this point remind us that it is one thing to say that we have to use a plurality of languages or schemes and quite another thing to say that this yields a plurality of (acceptable) ontologies. However, this is again where Putnam's transcendental philosophy should surface more explicitly than it does. A full-blown "Kantian pragmatist" can argue that there is no privileged or fundamental level of metaphysical theorizing apart from our use of the schemes we use or the practices we engage in; by assuming there is such a level we also assume that the idea of "the totality of all objects" (or something similar) would make sense, and here Putnam's arguments strike with their full force. Ontology is – in Putnam's earlier vocabulary – "internal" to such uses of and engagements with our practice-embedded schemes. Putnam himself might have acknowledged that some versions of his internal realism did go too far to the Kantian direction, but even this he was reluctant to admit.

Another problem a realist might raise emerges from the fact that Putnam seems to unproblematically help himself to a pre-given ontology of (say) states of affairs when speaking about the different ways in which the "same" state of affairs (e.g., the three Carnapian individuals, or the contents of a room) could be described by using different schemes (see, e.g., Putnam 2022, 109). Here, however, he explicitly denies that any such ontology of either states of affairs, events, or anything else can be given priority (ibid.). No one description can be privileged. This is the very point of the ideas of conceptual relativity and pluralism.

Furthermore, one reason for attributing something like a transcendental idealism – or, better, transcendental pragmatism – to Putnam is that he continued even in the 2000s to endorse a certain kind of "mind-dependence" of properties.[11] This is clear in his 2001 reply to Charles Travis, in which he explicitly notes that he agrees with Travis's view according to which "[a] property (the sort of thing we describe as 'a way that objects might conceivably be') requires *interpretation*", while "the reasonableness of property-interpretations [...] depends on the human interests and practices that figure in the particular contexts of speaking"; moreover, this "mind-dependence" of properties "goes all the way down" and is thus "*deep*" (Putnam 2022, 67; original emphases). I agree with this, of course, and these statements provide, in my view, a useful summary of how a pragmatist might look at the metaphysics of properties. But I also think, *pace* Putnam, that a reasonable pragmatist interprets this mind- or interpretation-dependence transcendentally, analogously to the way in which empirical reality is taken to be dependent on the human cognitive faculty in Kant's transcendental idealism (which, according to Kant, is fully compatible with empirical realism). If interpreted in any other way, mind-dependence becomes a

[11] In this context, he speaks of the mind-dependence of properties rather than objects, but I see no reason why the argument could not be extended to the mind-dependence of objects as well, given that the very notion of an object is "infinitely extendable" and that it is up to us (or our "mind") to make such extensions.

highly implausible factual, causal, or empirical claim, and no pragmatist or realist should suggest that any objects or properties are mind-dependent in that sense.

A transcendental argument for realism (about truth)?

While Putnam, as we just saw, never accepted the suggestion that his own views on realism would come even close to Kantian transcendental idealism (even if pragmatically "softened" or "naturalized"), he did acknowledge that his own famous "brains in a vat" argument, according to which (simplifying dramatically) we cannot possibly be brains in a vat because we could not refer to brains and vats and could thus never truly say or think that we are brains in a vat if we were (see Putnam 1981, chapter 1), is a close relative of a transcendental argument, in a fallible and empirically contextualized sense:

> In spite of the fallibility of my procedure, and its dependence upon assumptions which might be described as 'empirical' (e.g. the assumption that the mind has no access to external things or properties apart from that provided by the senses), my procedure has a close relation to what Kant called a 'transcendental' investigation; for it is an investigation [...] of the *preconditions* of reference and hence of thought – preconditions built in to the nature of our minds themselves, though not (as Kant hoped) wholly independent of empirical assumptions. (Ibid., 16; original emphasis.)

Thus, Putnam did, at least in a qualified sense, accept the possibility of arguing transcendentally, although (as far as I know) he wrote little explicitly on transcendental arguments. It should not therefore have been impossible for him to view his own argumentation in favor of pluralism (and against metaphysical realism) as transcendental in an analogous sense. Those arguments, too, address the preconditions of reference and thought, particularly of reference and ontological thought about objects or reality.

More generally, Putnam can be said to have engaged in transcendental argumentation not only in his defense of pragmatic pluralism but also in reminding us that we cannot just give up our realistic understanding of truth as something that cannot be reduced to epistemic concepts such as justifiability under ideal epistemic conditions (as he at one point, at the peak of his internal realism, himself maintained). Transcendental reflection thus cuts both ways, both in favor of a certain kind of realism and against metaphysical forms of realism that go too far. Again, Putnam himself, however, never explicitly explicated the transcendental status of these arguments.

In particular, it may be suggested that for a pragmatist and pluralist like Putnam, a transcendental argument is needed to block the possibly threatening Rortyan slippery slope into a radically pragmatist conception of truth that in the end gives up the concept of objective truth entirely. If this is a correct analysis, then we can draw a simple moral: if you want to be a Putnamian pragmatist, and not a Rortyan one, you had also – despite Putnam's doubts – better be a transcendental philosopher!

That we just cannot get rid of the notion of truth does not mean that truth would have to be defined as correspondence with a metaphysically pre-structured reality. Putnam's own arguments against standard forms of realism and for an ontologically relevant pragmatic pluralism have given us good reasons to avoid such metaphysically realist theories of truth. But even if we do subscribe to some form of pragmatism about truth, this does not entail that the objectivity of truth would or could be sacrificed; moreover, we should remember that the notorious "pragmatist conception of truth", as formulated by James and others, does not abandon objectivity or even "agreement with reality", either (cf. Pihlström 2021).

While maintaining[12] that all truths we human beings can understand are "made true by conditions that are, in principle, accessible to some human beings *at some time or other*" (Putnam 2022, 105; original emphasis), Putnam finds it "absurd to suppose", with antirealists, that "there *could not* be intelligent beings so much smarter than we that some of their thoughts could not even be understood by us; and surely [...] some of those thoughts could be *true*" (ibid., 104). This could again be rephrased as a transcendental argument. In order to avoid absurdity in our use of the concept of truth, we must accept the possibility of truths reaching far beyond any human capacities of cognition and understanding. This is necessary for our being able to use the concept of truth in the ways we (arguably inescapably) have to use it in our linguistic practices – for example, in speaking about the possibility of intelligent beings smarter than us thinking true thoughts incomprehensible to us. This, according to Putnam, by no means threatens the view (which he around the time of this particular argument still saw as crucial to his internal realism) that the concepts of truth and warranted assertibility are interdependent (ibid., 106–107).

However, this would only be to endorse the "spirit" of Kant's empirical realism, not any full-blown Kantian doctrine. Putnam explains:

> But that does not mean that I accept Kant's transcendental claim that space and time are "inside us," or the idea that our knowledge fails to reach to the "intrinsic properties" of the "things in themselves," claims whose intelligibility I have repeatedly challenged. Like Peter Strawson,[13] I believe that there is much insight in Kant's critical philosophy, insight that we can inherit and restate; but Kant's "transcendental idealism" is no part of that insight. (Ibid., 107.)

I have to say it is a disappointment that Putnam seems to have taken Strawson's Kant as a (at least roughly) correct picture of Kant's transcendental idealism. He could

[12] In his response to Anderson cited above.

[13] The reference, of course, is to Strawson's 1966 volume, *The Bounds of Sense*.

have been more open to a transcendental rearticulation of his own arguments, if he had found his Kant in, say, Henry Allison's one-world reading instead of Strawson's influential but (according to many later scholars) flawed account.[14]

Employing the concept of truth, at any rate, is no more dispensable in our practices than our habit of referring to tables and chairs. We cannot "cope" (borrowing one of Richard Rorty's favorite terms) without remaining committed to the objectivity of truth. We may thus even speak of truth as playing a transcendental role in our practices: a sincere commitment to pursuing the truth is a necessary condition for the possibility of our genuine participation in not only practices of inquiry but of thought itself. On the other hand, a fallibilist pragmatic pluralist must acknowledge the possibility that even our most deeply entrenched transcendental commitments may change in the course of history. It remains an open philosophical question whether we can even coherently pose the question of whether it would be possible for us to "live" without a robust concept of truth, and what this would mean.[15]

Conclusion

The purpose of this brief discussion has not been to persuade Panu or any other realist philosopher to endorse a form of pragmatism or transcendental idealism (which remain realist in their own way, though). I have only tried to suggest, by drawing attention to some of Putnam's arguments (including those that I discussed with Panu already in the early 1990s) and some of his responses to his critics recently collected in a posthumous volume, that our engagement with the realism issue may enormously benefit from taking seriously the transcendental argumentative strategy which Putnam himself arguably employed but only in a very qualified sense acknowledged as his own. It would be an entirely different ask (which I have tried to undertake in some of my own work over the years and decades, presumably in a continuous implicit dialogue with realist friends like Panu – and of course Ilkka) to demonstrate that transcendental arguments for a pragmatic pluralism are both philosophically sound and capable of securing a sufficiently robust notion of objective truth.

[14] My own attempts to develop a Kantian pragmatist approach are indebted to Allison (2004 [1983]); see, e.g., Pihlström 2009. Allison's "anthropocentric" (as distinguished from "theocentric") formulation of transcendental idealism is particularly relevant to Putnamian engagements with realism.

[15] Thus, it also remains a task for a transcendentally sensitive pragmatist philosophy to explore a question like this.

References

Allison, Henry E. (2004) [1983]: *Kant's Transcendental Idealism: An Interpretation and Defense*. Revised and enlarged edition. New Haven, CT and London: Yale University Press.

Niiniluoto, Ilkka (1999): *Critical Scientific Realism*. Oxford: Oxford University Press.

Pihlström, Sami (2006): "Putnam's Conception of Ontology". *Contemporary Pragmatism* 3:2, 1–15.

Pihlström, Sami (2009): *Pragmatist Metaphysics: An Essay on the Ethical Grounds of Ontology*. London and New York: Continuum.

Pihlström, Sami (2021): *Pragmatist Truth in the Post-Truth Age: Sincerity, Normativity, and Humanism*. Cambridge: Cambridge University Press.

Pihlström, Sami (2023): *Humanism, Antitheodicism, and the Critique of Meaning in Pragmatist Philosophy of Religion*. Lanham, MD: Lexington.

Putnam, Hilary (1975): *Mind, Language and Reality*. Cambridge: Cambridge University Press.

Putnam, Hilary (1981): *Reason, Truth and History*. Cambridge: Cambridge University Press.

Putnam, Hilary (1983): *Realism and Reason*. Cambridge: Cambridge University Press.

Putnam, Hilary (1987): *The Many Faces of Realism*. La Salle, IL: Open Court.

Putnam, Hilary (1990): *Realism with a Human Face*. Ed. James Conant. Cambridge, MA and London: Harvard University Press.

Putnam, Hilary (2004): *Ethics without Ontology*. Cambridge, MA and London: Harvard University Press.

Putnam, Hilary (2006): "Replies to Commentators". *Contemporary Pragmatism* 3:2, 67–98.

Putnam, Hilary (2008): *Jewish Philosophy as a Guide to Life*. Bloomington and Indianapolis: Indiana University Press.

Putnam, Hilary (2022): *Philosophy as Dialogue*. Eds. Mario De Caro and David Macarthur. Cambridge, MA and London: The Belknap Press of Harvard University Press.

Putnam, Hilary and Putnam, Ruth Anna (2017): *Pragmatism as a Way of Life: The Lasting Legacy of William James and John Dewey*. Ed. David Macarthur. Cambridge, MA and London: The Belknap Press of Harvard University Press.

Raatikainen, Panu (2001): "Putnam, Languages and World". *Dialectica* 55, 167–174.

Raatikainen, Panu (2004): *Ihmistieteet ja filosofia*. Helsinki: Gaudeamus.

Raatikainen, Panu (2006): "The Scope and Limits of Value-Freedom in Science". In Heikki J. Koskinen, Sami Pihlström, and Risto Vilkko (eds.), *Science – a Challenge to Philosophy?* Peter Lang, 323–331.

6
A defense of Academic skepticism

Markus Lammenranta

I will defend a form of Academic skepticism that denies the possibility of knowledge about the external world. The standard argument for it relies on internalism and infallibilism, doctrines that were widely accepted in the history of epistemology until the late 20th century. Contemporary epistemologists typically deny at least one of them, because together they lead to skepticism. Skepticism is thought to be bad because it conflicts with common sense, our ordinary epistemic practices, and linguistic data. I will argue that this is not so, that Academic skepticism gives in fact a better explanation of our intuitions and linguistic data than dogmatic epistemology. Finally, following the steps of Arcesilaus, Carneades and Hume, I will show how Academic skepticism can give a good response to the Stoics' Apraxia objection that skepticism makes rational action and good life impossible. On the contrary, it is skepticism that makes a good and flourishing life possible.

The skeptical argument

Arguments for Academic skepticism raise possibilities of error. Skeptical hypotheses describe such possibilities. A famous example is the contemporary version of Descartes' evil demon hypothesis:

> *The brain-in-a-vat hypothesis*: I am a brain in a vat wired to a computer that stimulates it so that I have the experiences and beliefs I have now, but these beliefs are false.

Lammenranta, M. (2025). A defense of Academic skepticism. In A. Laitinen, M. Keinänen, J. Reinikainen & A. Honkasalo (eds.). *Language, Truth, and Reality.* Tampere University Press. (pp. 70–79).
https://doi.org/10.61201/h5fbdk64

We can use the hypothesis to distinguish between two cases: in one of them the hypothesis is false and in the other it is true.

> *The good case*: Things are the way I think they are. I have hands, and it does not just appear that I have hands.
>
> *The bad case*: I am a handless brain in a vat, and it merely appears to me that I have hands.

Though I believe that I am in the good case, I cannot rule out the possibility that I am in the bad case. Because my experiences are the same in both cases, everything appears exactly the same. So, we get the following argument:

> P1 If I know that I have hands, my evidence rules out the possibility that I am a handless brain.
>
> P2 My evidence does not rule out the possibility that I am a handless brain.
>
> C Therefore, I do not know that I have hands.

The argument is valid, and the premises are plausible. Their plausibility is explained by three doctrines that are independently plausible:

> *Evidentialism*: S knows that p only if S's evidence supports p.
>
> *Infallibilism*: S knows that p only if S's evidence guarantees the truth of p (S's evidence rules out all alternatives to p, that is, the possibilities in which not-p). In short, knowledge requires conclusive evidence.
>
> *Internalism*: S has the same evidence in the good case and in the bad case.

Infallibilism explains why P1 is true, and internalism explains why P2 is true.

All these doctrines are intuitive, and this is widely conceded by philosophers. The problem is that together they lead to skepticism, which they find impossible to accept. I will try to show that skepticism is not so hard to accept. It may be the best way to save our overall intuitions. Indeed, some of the greatest modern philosophers have been Academic skeptics, such as John Locke, David Hume and Bertrand Russell.

Because it is infallibilism that is most often rejected by contemporary philosophers,[1] I'd like to say something about its intuitiveness and the costs of rejecting it.

[1] Evidential internalists reject just infallibilism. Evidential externalists, like Williamson, reject just internalism. Reliabilists reject all three doctrines. I will not discuss these alternatives to Academic skepticism in detail. If my case for skepticism is successful, we will lose motivation for them.

The madness of fallibilism

David Lewis calls fallibilism mad and defends the intuitiveness of infallibilism in "Elusive Knowledge" (1996, 249):

> It seems as if knowledge must be by definition infallible. If you claim that S knows that P, and yet you grant that S cannot eliminate a certain possibility in which not-P, it seems as if you have granted that S does not after all know that P. To speak of fallible knowledge, or knowledge despite uneliminated possibilities of error, just sounds contradictory.

Let me remind you of some problems of fallibilism. Firstly, knowledge attributions that concede the possibility of error are odd. Lewis refers to such attributions. For example, you will find it odd if I say that I know that it is Monday, but I may be wrong, or that I know that the animal in the cage is a zebra, but it may be a painted mule. However, there should be nothing odd with these claims if fallibilism were true.

Secondly, fallibilism creates Gettier problems: These are counterexamples to the analysis of knowledge as a true and justified belief. If a justified belief can be false, as fallibilism says, it is possible to imagine cases of justified beliefs that are true by luck. Intuitively, such beliefs are not knowledge.

Thirdly, the Lottery problem supports infallibilism. Assume that I have a lottery ticket. We have the intuition that I cannot know that my lottery ticket will lose (though this is very probable). So, any probability less than one seems to be insufficient for knowledge.

Fourthly, if fallibilism were able to solve the Lottery problem, it would still have the Threshold problem: If knowledge does not require conclusive evidence, then how strong must the evidence be on the scale from 0 to 1? Any threshold less than 1 seems arbitrary.

Of course, fallibilists have tried to offer various solutions, but the point is that fallibilism has a lot of problems that infallibilism easily avoids.

Skepticism, common sense, and ordinary language

Though there are plausible arguments for skepticism, philosophers typically think that there must be something wrong with their premises. That a view leads to skepticism is taken to be a *reductio ad absurdum* of it. Why?

One reason is that many philosophers follow G. E. Moore (1959), who thought that if philosophy is in conflict with common sense, common sense wins. It is a part of common sense that we know a lot. So, if skepticism denies this, it is wrong.

The skeptic naturally rejects this common-sense view, but the price may not be high. First, the view that philosophy cannot revise common sense is overly pessimistic about what philosophy can do. Second, the skeptic can explain why people believe that they know a lot even though it is not strictly speaking true.

Another objection to skepticism appeals to our ordinary use of language. John L. Austin (1979) thinks that the fact that we attribute knowledge to subjects who don't satisfy the skeptical standards shows that these standards are too stringent. Our ordinary standards are less demanding. We do not normally require of a person who claims to know something to be able to rule out the possibility that she is asleep or that she is just a brain in a vat. According to Austin, our ordinary use of “know” supports rather the *relevant alternatives theory of knowledge*:

> S knows that p only if S’s evidence rules out all relevant alternatives to p.

According to this account, I can know that I have hands even though my evidence does not rule out the possibility that I am a handless brain, because this possibility is not a relevant alternative (See also Stroud 1984, 39–82).

To respond to Austin and Moore, we need to make the following distinction:

1. What is true to say?
2. What is appropriate or reasonable to say?

As skeptics, we can agree with Austin that it is appropriate to say that someone knows a lot though she cannot rule out the skeptical alternatives, but insist that it is not strictly speaking true to say so.

Peter Unger (1971) defends this sort of response by arguing that “know” is an absolute term, like “flat” and “empty”. For example, if a plane is flat, it is absolutely flat. There is nothing that is flatter. Flatness rules out all bumps and curves. Similarly, if you know that p, no one else knows it better or to a higher degree. There are no degrees of knowledge. Knowledge rules out all possibilities of error.

It follows from such absoluteness that no plane is really flat, because there are always some microscopic bumps on it. In the same way, no one knows anything about the external world, because there are always some uneliminated possibilities of error.

Yet, according to Unger, it may be appropriate to say that the floor is flat because it is for practical purposes close enough to absolute flatness. Some small bumps do not matter if we want to dance on it. Similarly, it may be appropriate to say that you know that you have hands, because you are close enough to knowing this for practical purposes. It does not matter that you cannot rule out the handless-brain possibility.

The point is that when we use the term “know” in ordinary contexts, we speak loosely. Strictly speaking we don’t know anything about the external world, but loosely speaking we know many things. The skeptic points out that the strict use of “know” explains the plausibility of skeptical arguments, and the loose use explains our ordinary epistemic practices and the common-sense intuitions (Davis 2007).

There are two popular dogmatic or non-skeptical theories that try to do the same. Contextualism (Cohen 1988; DeRose 1995; Lewis 1996) and subject-sensitive invariantism (Hawthorne 2004; Stanley 2005) concede that when I consider skeptical arguments and conclude "I don't know anything about the external world", what I say is true, but when I in some ordinary context say "I know that the sun is shining" what I say is true as well. This is possible because there is a shift in epistemic standards between the skeptical context and the ordinary context: I can meet the low standards of the ordinary context without meeting the high standards of the skeptical context. These views try to do justice both to our skeptical intuitions and common-sense intuitions. They differ from skepticism in taking our ordinary knowledge attributions to be true, whereas skepticism takes them to be appropriate but false. According to skepticism, our epistemic standards are invariant and high in all contexts.

You may think that it is an advantage of contextualism and subject-sensitive invariantism that they make our ordinary knowledge attributions true. However, linguistic evidence seems to support skeptical invariantism:

Let's imagine the following dialogue between A and B:

> A: Do you know what that is?
> B: Yes, I do. It is a zebra.
> A: But can you rule out that it is a cleverly painted mule?
> B: No, I can't.
> A: So, you admit you didn't know it was a zebra?
> B: Yes, I do. I didn't know that.

B's concession that she didn't know is quite natural, and skeptical invariantism explains this: it is true. B takes back her original knowledge claim. She admits that she spoke loosely. She did not really know.

Subject-sensitive invariantism does not predict this answer. According to it, B should say something like this:

> B: No, I don't. I did know then that it was a zebra. But after you mentioned the painted-mule possibility, I no longer know.

This answer is odd. Yet it is true according subject-sensitive invariantism, because mentioning the painted-mule possibility raises the standards of knowledge. B does not meet the new high standards, though she met the original low standards. According to this view, knowledge is elusive. It disappears when error possibilities are mentioned, which is strange (DeRose 2009, 194–96).

Contextualism does not have this problem, because high standards do not affect knowledge itself. They determine the content of knowledge attributions. So, when B originally said "I know it is a zebra", what she said was true. B fulfills the low standards of that context. When she in a new context says, "I did not know it was a zebra", she also speaks truly. Her earlier belief does not satisfy the higher standards of this new context of utterance.

So far so good, but when we change A's last question, we get a very odd result:

> A: So, you admit that what you earlier said was not true.
> B: No, I don't admit anything like that.

It follows from contextualism that what B said in the earlier low-standard context was true. So, according to contextualism, what B says here should be quite appropriate, though it is not (MacFarlane 2005, 202–203).

This linguistic evidence suggests that the salience of error-possibilities affects neither the conditions of knowledge nor the content of "know". What it does is to make us reconsider and take back our knowledge claims. So, the linguistic evidence supports skeptical invariantism.

The Apraxia objection

The core of the Apraxia objection is that the skeptic is not able to act rationally if she has no beliefs. Rational action is not possible without beliefs. The Stoics made this objection against the Academic skeptics, such as Arcesilaus and Carneades (Vogt 2010; Perin 2010, 86–113). The objection is relevant because the skeptics were thought to be committed to the Stoic doctrine that a wise person believes something only if she knows it. So, if the skeptics deny knowledge, they should also deny beliefs.

The Stoics thought that a wise person assents only to cognitive impressions that are always true. Assuming that assenting to an impression is to believe its content, we get the view that a wise person's beliefs are based on cognitive impressions that guarantee their truth. And assuming that beliefs based on cognitive impressions constitute knowledge, a wise person has no mere beliefs, just knowledge. The skeptics pointed out that there are no cognitive impressions, because for any true impression there can be a false one that is exactly similar. So, a wise person has no knowledge and should not have any beliefs (Frede 1987; Reed 2002).

If we talk about justification instead of a wise person, it seems that the Stoics are committed to an infallibilist account of justification.

> *Infallible justification*: S is justified in believing that *p* if and only if S has conclusive evidence for *p*.

The skeptics argue that because conclusive evidence consists of cognitive impressions and there are no cognitive impressions, we are not justified in believing anything and should suspend belief. We get the same result by considering the skeptical hypotheses of the *First Meditation*. They show that we don't have conclusive evidence for beliefs about the external world.

Then the Stoics made the Apraxia objection, and the skeptics responded that we can act rationally without beliefs. We can guide our actions by following our impressions. Let me make a suggestion that is similar in spirit.

The principle of infallible justification concerns full belief. To fully believe that p one must be maximally convinced or certain of p. One must have no doubts about p. Understood in this way the principle is quite plausible: One should be certain of p only if one has conclusive evidence for p, evidence that rules out all possibilities of error. If there are any uneliminated possibilities of error, one should not be completely certain of p.

How do we then conduct our lives? Arcesilaus and Carneades say that we follow our impressions. Rational action is possible on the basis of rational or convincing impressions. We can understand impressions as propositional attitudes that contemporary philosophers call *seemings*. For example, in sense experience things seem to be in a certain way. However, because seemings can be initially in conflict, we can also speak about resultant seemings – how things seem after assessing the weight of the initial seemings. The resultant seeming is a matter of how things seem all things considered. According to Sosa (2015, 231–232), we can understand resultant seemings as credences, degrees of confidence or belief.

The response to the Apraxia objection is thus that action is possible on the basis of degrees of belief and that rational action is possible on the basis of rational or justified degrees of belief. Action does not require full belief.

What justifies degrees of belief? The popular answer in modern philosophy is evidence. The degree of belief should reflect the strength of the evidence:

> "Wise man… proportions his belief to the evidence." (Hume 1975, 110)
>
> "Perfect rationality consists… in attaching to every proposition a degree of belief corresponding to its degree of credibility." (Russell 1948, 397)
>
> *Bayesian evidentialism*: In order to be epistemically justified in her degree of belief that p, an agent's degree of belief that p must conform to her evidence for and against p.

So, one possible Academic response to the Apraxia objection is to suggest that rational action and thought are based on justified credences or degrees of belief. However, there is a worry that this view makes both theoretical and practical reasoning too complicated. We may not have cognitive resources for this sort of reasoning.

If you share this worry, the skeptic can give another response to the Apraxia objection. It is to suggest that there are outright beliefs (all or nothing beliefs) that do not require maximal confidence or certainty. They just require sufficient confidence, confidence that is above a certain threshold. We get the following theory of justification for such out-right beliefs:

> *The Lockean thesis*: S is justified in believing that *p* if and only if S is justified in having a degree of confidence in *p* that is sufficient for belief (above the belief threshold).

The central question of this view is how to determine the threshold in a non-arbitrary way (the Threshold problem). One option is to think that the only non-arbitrary threshold is maximal confidence or subjective probability 1. The problem is that we have very few such beliefs.

The other option is to think that the threshold varies with the context.

> "Intuitively, one believes p outright when one is willing to use p as a premise in practical reasoning." (Williamson 2000, 99)

It is plausible that whether one is willing to rely on p in action depends on practical considerations that vary with the context. If the costs of being wrong about p are very high, one may not be willing to act on p. If they are low, one may be willing to act. So, if Williamson is right, outright belief depends on the practical stakes. We can call this view doxastic pragmatism.

To sum up, knowledge is strong, belief is weak. Knowledge requires the highest degree of belief and justification. But because our beliefs and justifications are weak, we never or rarely attain knowledge. Though we aim at knowledge, we are quite happy to come close to it. For practical purposes, this is enough: Rational action does not require knowledge.[2] It just requires justified degrees of belief or weak beliefs.

Skepticism as a way of life

Another version of the Apraxia objection was that the skeptic cannot live a good life (Vogt 2010, 166). One common idea in ancient philosophy after Socrates was that philosophy should be a guide to good life. The Stoics followed Socrates in thinking that it is knowledge that guarantees such a life. The Pyrrhonists reported that suspension of belief is the way to a happy life. The Academic skeptics Arcesilaus and Carneades were silent about this, but there was a modern skeptic who defended the practical value of Academic skepticism compared to Pyrrhonism and dogmatism.

That skeptic was David Hume who, in the final section of *Inquiry*, says that the life of a Pyrrhonist would be miserable and short, because without beliefs she is not able to act and to satisfy her basic needs. But also, dogmatism has its dangers:

> The greater part of mankind are naturally apt to be affirmative and dogmatical in their opinions; and while they see objects only on one side,

[2] In contemporary philosophy, Hawthorne and Stanley (2008) defend the Stoic view that knowledge is necessary (and sufficient) for rational action. Brown (2008) and Comesaña & McGrath (2014) criticize it.

> and have no idea of any counterpoising argument, they throw themselves precipitately into the principles, to which they are inclined; nor have they any indulgence for those who entertain opposite sentiments. To hesitate or balance perplexes their understanding, checks their passion, and suspends their action. (Hume 1975, 161)

Hume thinks that what he calls academical or mitigated skepticism avoids the dangers of both Pyrrhonism and dogmatism. I think this is also true of the kind of Academic skepticism that I have defended (See also Hazlett 2014, 181-184).

Let's understand dogmatism in a way Sextus (2000, 3) does. The dogmatists are people who believe that they know the truth and have therefore no need to continue inquiry. It seems that one who believes that she knows that p is inclined to reason in the ways Hume suggests:

1. I know that p. If I know that p, I know that all evidence against p is misleading. So, I know that all evidence against p is misleading. So, I should pay no attention to the evidence against p.
2. I know that p. If I know that p, I know that anybody who disagrees with me about p is wrong. So, I know that anybody who disagrees with me about p is wrong. So, I should pay no attention to those who disagree with me about p.
3. I know that p. If I know that p, I may use p as a reason for action. So, I may use p as a reason for action.

All these ways of reasoning are based on plausible principles. So, Hume appears to be right about the dangers of dogmatism: Dogmatists ignore evidence and arguments against their view, do not tolerate those who have opposite views, and are inclined to act rashly. It is improbable that these inclinations would lead to a good life.

An Academic skeptic avoids both the dangers of Pyrrhonism and dogmatism. First, she has rational beliefs and is able to act rationally. Second, she does not believe that she knows that p. So, she has not terminated the inquiry about p and is sensitive to further evidence both for and against p, including evidence provided by other people. And, finally, she considers carefully whether her evidence for p is sufficient for action in the context she is.

So, Academic skepticism seems to offer a better life than Pyrrhonism and dogmatism. A further benefit is that it encourages us to cultivate intellectual virtues, such as conscientiousness, humility and open-mindedness, which are constitutive of an intellectually good and flourishing life. To sum up, there are both epistemic and practical reasons to be an Academic skeptic.

References

Austin, J. L. (1979). “Other Minds.” In *Philosophical Papers*, 3rd ed. Oxford: Oxford University Press.

Brown, Jessica. (2008). “Knowledge and Practical Reason.” *Philosophy Compass* 3: 1135–1152.

Cohen, Stewart. (1988). “How to Be a Fallibilist.” *Philosophical Perspectives* 2: 91–123.

Comesaña, Juan, and Matthew McGrath. (2014). “Having False Reasons.” In C. Littlejohn and J. Turri (eds.), *Epistemic Norms: New Essays on Action, Belief and Assertion.* Oxford: Oxford University Press.

Davis, Wayne A. (2007). “Knowledge Claims and Context: Loose Use.” *Philosophical Studies* 132: 395–438.

DeRose, Keith. (1995). “Solving the Skeptical Problem.” *Philosophical Review* 104: 1–52.

———. 2009. *The Case for Contextualism*. Oxford: Clarendon Press.

Frede, Michael. (1987). “Stoics and Skeptics on Clear and Distinct Impressions.” In *Essays in Ancient Philosophy*. Minneapolis: University of Minnesota Press.

Hawthorne, John. (2004). *Knowledge and Lotteries*. Oxford: Clarendon Press.

Hawthorne, John, and Jason Stanley. 2008. “Knowledge and Action.” *The Journal of Philosophy* 105: 571–90.

Hazlett, Allan. (2014). *A Critical Introduction to Skepticism*. London: Bloomsbury.

Hume, David. 1975. *Enquiries Concerning Human Understanding and Concerning the Principles of Morals*. Edited by P. H. Nidditch. 3rd ed. Oxford: Oxford University Press.

Lewis, David. (1996). “Elusive Knowledge.” *Australasian Journal of Philosophy* 74: 549–567.

MacFarlane, John. (2005). “Assessment Sensitivity of Knowledge Attributions.” *Oxford Studies in Epistemology* 1: 195–233.

Moore, G. E. (1959). *Philosophical Papers*. London: George Allen & Unwin Ltd.

Perin, Casey. (2010). *The Demands of Reason*. Oxford: Oxford University Press.

Reed, Baron. (2002). “The Stoics’ Account of Cognitive Impression.” *Oxford Studies in Ancient Philosophy* 23: 147–80.

Russell, Bertrand. (1948). *Human Knowledge: Its Scope and Limits*. New York: Simon and Schuster.

Sextus Empiricus. (2000). *Outlines of Skepticism*. Edited by Julia Annas and Jonathan Barnes. Cambridge: Cambridge University Press.

Sosa, Ernest. (2015). *Judgment and Agency*. Oxford: Oxford University Press.

Stanley, Jason. (2005). *Knowledge and Practical Interests*. Oxford: Clarendon Press.

Stroud, Barry. (1984). *The Significance of Philosophical Scepticism*. Oxford: Oxford University Press.

Unger, Peter. (1971). “A Defense of Skepticism.” *Philosophical Review* 80: 198–219.

Vogt, Katja. (2010). “Scepticism and Action.” In *The Cambridge Companion to Ancient Scepticism*, 165–180. Cambridge: Cambridge University Press.

Williamson, Timothy. (2000). *Knowledge and Its Limits*. Oxford: Oxford University Press.

7
Is the argument from inductive risk merely research ethics?

Inkeri Koskinen

Is the argument from inductive risk at heart simply an argument about research ethics? If so, does it really challenge the value-free ideal as an ideal? After all, proponents of the value-free ideal do not generally argue that ethics should not place constraints on research. Perhaps the argument from inductive risk merely points out that in some situations, our ethical duties restrict our ability to strive towards the ideal?

Once, while waiting for the coffee to drip in the break room at the Department of Philosophy at Tampere University, I started a discussion about this question with Panu, but then something else came up, and we never managed to finish the discussion. This, therefore, is my answer to his question as I understood it.[1] In brief: Yes, the argument from inductive risk is at heart a research-ethical one, or rather just an ethical one. But what this shows is that the value-free ideal is untenable as an ideal –precisely because even its proponents usually agree that researchers are ethical agents with all the usual ethical responsibilities.

I will start by introducing the value-free ideal and the argument from inductive risk and then argue that ideals ought to be able to guide action. Finally, I claim that

[1] It is fully possible that I misunderstood Panu's argument (and I could not check this, because he is not supposed to know about this book). If that is so, this paper is an exercise in argumentation against an imaginary stance – a fairly common thing in philosophy.

Koskinen, I. (2025). Is the argument from inductive risk merely research ethics? In A. Laitinen, M. Keinänen, J. Reinikainen & A. Honkasalo (eds.). *Language, Truth, and Reality.* Tampere University Press. (pp. 80–87). https://doi.org/10.61201/h5fbdk64

the argument from inductive risk does not just point out some constraints to our ability to follow the value-free ideal but shows that it is undesirable as an ideal.

While much earlier versions of the demand that science be value-neutral or value-free can be found (see Proctor 1991), it is common to name Max Weber as the first explicit proponent of the value-free ideal. He formulated it to argue that social sciences can and should strive for objectivity:

> Accordingly, cultural science in our sense involves "subjective" presuppositions insofar as it concerns itself only with those components of reality which have some relationship, however indirect, to events to which we attach cultural significance. [...] But it obviously does not follow from this that research in the cultural sciences can only have results which are "subjective" in the sense that they are *valid* for one person and not for others. Only the degree to which they interest different persons varies. In other words, the choice of the object of investigation and the extent or depth to which this investigation attempts to penetrate into the infinite causal web, are determined by the evaluative ideas which dominate the investigator and his age. In the *method* of investigation, the guiding "point of view" is of great importance for the *construction* of the conceptual scheme which will be used in the investigation. In the mode of their *use,* however, the investigator is obviously bound by the norms of our thought just as much here as elsewhere. For scientific truth is precisely what is *valid* for all who *seek* the truth. (Weber 1904/1949, 82–84.)

In other words, while "the value-ideas which dominate the investigator and his age" unavoidably influence what is studied in the social sciences, the actual research can and ought to be free of values. After Weber's time, the ideal became more widely and explicitly accepted, and for example Kuhn's (1977) distinction between epistemic and non-epistemic values has influenced its formulations. Here are two characterisations by its contemporary proponents:

> The ideal of value free science states that the justification of scientific findings should not be based on non-epistemic (e.g. moral or political) values. (Betz 2013, 207.)

> For example, it is no abandonment of epistemic ideals to reject a research project aimed at developing a doomsday device. Building a doomsday device is simply not the sort of research project most people consider valuable to pursue. But it would be an abandonment of epistemic ideals and scientifically unacceptable if one used ethical criteria in the evidential assessment that a doomsday device is technologically feasible. (Hudson 2016, 187–188.)

Adversaries of the ideal have also given characterisations of it – here are two examples:

> It does not hold that science is a completely value-free enterprise, acknowledging that social and ethical values help to direct the particular projects scientists undertake, and that scientists as humans cannot completely eliminate other value judgements. However, the value judgements internal to science, involving the evaluation and acceptance of scientific results at the heart of the research process, are to be as free as humanly possible of all social and ethical values. (Douglas 2009, 45.)

> There are various ways one might interpret the value-free ideal (VFI), but the most common way is the claim that only scientific or "epistemic" values can influence scientific reasoning or inference, while the only place for other values, including social and ethical values, should be in external aspects of science, such as choice of research projects or decisions about acceptable methods. [...] [T]he VFI is an all-or-nothing affair – either social and ethical values should play a role in the internal phases of scientific reasoning, or they should not. (Brown 2024, 2/31.)

To summarise, the value-free ideal states that while non-epistemic values can legitimately influence the "external aspects" of science, such as the choice of research projects, only epistemic values – that is, values that promote the attainment of truth – have a legitimate role in the central stages of scientific research, especially in the assessment of evidence and the justification of findings.

The argument from inductive risk (AIR) is one of the most influential arguments against the value-free ideal. There are several earlier versions, notably one by Rudner (1953), but here I will focus on Douglas's (2000; 2009) more recent and stronger formulation.

The argument starts by stating that scientists have the same moral responsibilities as everyone: they are responsible for their actions as scientists in the same way they are responsible for their actions as human beings in general. Therefore, it is their responsibility to consider the predictable, non-epistemic consequences of any errors they make in their research: a scientist, as a scientist, has no special licence to recklessly or negligently cause risk to others. Because of this, the predictable consequences of their research, including the predictable future use of their results, must influence their decisions when they face risks of error. Such risks, inductive risks, are ineliminable in all empirical research. For instance, when a scientist chooses between a method that is known to produce some false negative results but rarely false positive ones, and another that is known to produce some false positive results but rarely false negative ones, they must consider the predictable consequences of the choice: would one type of error be more perilous than the other? Or when they evaluate whether they have strong enough evidence to make an inductive leap to the acceptance or rejection of a hypothesis, they must take into account what the non-epistemic cost of an error might be. Such considerations require non-epistemic values. An Assyriologist interpreting weathered cuneiform signs can legitimately take more and different kinds of risks of error than a medical researcher developing a vaccine. Researchers face inductive risks throughout the research process. Therefore non-epistemic values must also influence the internal stages of the process.

This contradicts all the formulations of the value-free ideal that I mentioned above. It clearly contradicts the idea that non-epistemic values must not influence the internal stages of the research process. And more specifically, it contradicts the idea that the justification of scientific findings or the assessment of evidence should not be based on non-epistemic values. The Assyriologist can legitimately accept a hypothesis with weaker evidential support than the medical researcher, meaning that non-epistemic values have a legitimate role in the assessment of evidence and the justification of scientific findings.

As noted, AIR is not so much a research-ethical argument as simply an ethical argument. It is based on the idea that a researcher is responsible for the foreseeable consequences of their actions; recklessness and negligence are unacceptable, even in the role of a researcher. As Panu said, many proponents of the value-free ideal have no objection to this.

For what follows, it is important to note that AIR does not only demonstrate that a researcher must allow non-epistemic values to influence their decisions in situations where their research has foreseeable non-epistemic consequences. It also shows that a researcher has a duty to assess whether there are any such foreseeable consequences that should be taken into account. As Douglas emphasises, such assessments are often done collectively in the field in question. But sometimes it is the individual scientist conducting cutting-edge research who is in the best position to grasp the potential implications and risks of their work. (Douglas 2009, 83–84.)

Does this threaten the value-free ideal as an ideal? Several philosophers have argued that it does not: even if unattainable, it remains a good ideal (e.g. Hudson 2016) or worth pursuing (Menon & Stegenga 2023). In a recent article, Matthew J. Brown (2024) has presented what I find to be strong arguments against various versions of this idea.[2] I will focus on what I take to be a version of one of these arguments.

Suppose we gave up the requirement that the value-free ideal is an "all-or-nothing affair" (Brown 2024, 2/31); in other words, if we treated it like the requirement to use the research methods that are epistemically best for the task at hand. While this is ideally how we should act, we are prepared to make concessions if the epistemically best method is not ethically acceptable. Research ethics places constraints on research. Similarly, we could think that AIR simply identifies a type of situation where, for ethical reasons, our ability to follow the value-free ideal is restricted.

[2] Brown's Sisyphean paper, which addresses numerous recent attempts to defend the value-free ideal, was published a week before the completion of this piece. I can recommend it highly, even though it made finishing this paper somewhat challenging.

I will argue that this does not work, and that the analogy does not hold. For my argument, I need a criterion that I can use when assessing whether an ideal is good or not. When discussing normative ideals, Brown presents a good basis for such a criterion: he argues that normative ideals ought to be able to guide action, and normative ideals in science must be able to do so in science: "we are not concerned with what is epistemically preferable, but what is preferable all things considered. We don't want an epistemic ideal, but a *scientific* ideal, that is, an ideal to guide scientists who have both epistemic and social duties." (Brown 2024, 17/31.)

What would make the value-free ideal a poor ideal for science? In my view, it would be a poor, undesirable ideal if it guided action in a harmful or unjustified way, forbidding something that is always legitimate in science.

Could one think that the value-free ideal forbids a researcher from acting in a way that, in light of AIR, is their duty, but such cases are exceptions where the normally prevailing ideal must be temporarily set aside? Could AIR, then, simply highlight a type of situation where acting according to the ideal is not possible due to overriding ethical reasons? In other situations—that is, when research has no foreseeable non-epistemic consequences—the value-free ideal would nevertheless guide action as we would like it to do.

AIR is often taken to imply that if research has no foreseeable non-epistemic consequences, then researchers have no ethical obligations that would require non-epistemic values to play any role in the central stages of the research process. But if we accept the criterion for a poor ideal that I formulated above, the key question is whether it is still legitimate for a researcher to allow non-epistemic values some role in the central stages of the research process. If it is legitimate in situations where the research has no foreseeable non-epistemic consequences, then the value-free ideal forbids actions that are generally legitimate, making it a poor ideal.

Does AIR show that it is legitimate to give non-epistemic values a role in the central stages of research even in such situations? I believe it does. Remember that a researcher has a duty to assess whether their research could have some foreseeable non-epistemic consequences such that they should be taken into account when weighing inductive risks. Failing to make this assessment would be negligent. And one needs non-epistemic values for making such an assessment: What counts as a sufficiently foreseeable consequence? Is a remote possibility of some future application enough if the associated risks are particularly severe? One cannot answer such questions without non-epistemic values.

But a diligent researcher might not be satisfied with making this assessment just once. This is because such matters cannot necessarily be fully determined before the research begins; concerns about the possible non-epistemic consequences of a research project may well arise when the research is already in progress. Douglas (2009, 83) gives an example of a situation where researchers' judgments about the foreseeable consequences of their work changed in this way: before December 1938, nuclear physicists could not imagine the atomic bomb, but after the discovery of fission they could. Such a change can happen at any stage of research. To use my imaginary example, an Assyriologist might realise midway through their research that one possible interpretation of the weathered cuneiform inscription they are studying could be politically sensitive and presenting it might in principle lead to violence.

It is therefore legitimate for a researcher to assess the matter throughout the research process: are there, at this stage of my project, any foreseeable non-epistemic consequences of the work in progress that would require, for instance, tightening the criteria for accepting or rejecting a hypothesis, or using a different method in the analysis of the data? Such assessments require non-epistemic values even if the answer is negative – for example, if the Assyriologist ultimately decides that the risk is so small that it can be ignored, and that there is no need to make any changes in the ways in which they weigh inductive risks. Even in such a case non-epistemic values have played a role in the assessment of evidence.

Brown makes this point in his answer to Menon and Stegenga (2023), who suggest that researchers should often adopt value-mitigating strategies rather than make explicit value judgements:

> While there may be contexts where value-mitigation may be a good idea, in other cases it is crucial that scientists use (non-epistemic) values to weigh inductive risks, as they admit. More importantly, there is no way to tell ahead of time which kind of case we are in; so, on Menon's and Stegenga's own view, scientists will have to continue weighing non-epistemic values throughout inquiry in order to determine whether value-mitigation is permissible or superior to explicit value judgment. Whether to pursue value-mitigating strategies must be judged in each case according to non-epistemic values, effectively undermining the idea that this approach is value-free. (Brown 2024, 19–20/31.)

Whether the kind of monitoring and occasional reassessment I have described can be considered a duty depends on the context, but it is certainly always legitimate. As Douglas argues, scientists can be expected to "meet basic standards of consideration and foresight that any person would share, with the reasonable expectations of foresight judged against the scientist's peers in the scientific community" (Douglas 2009, 84). While this sets limits on what can be considered a duty, scientists are still allowed to go beyond their duty. We cannot predict at which stage of research it might be possible to recognise a potential non-epistemic consequence of the work we are doing. Therefore, it is always legitimate for a researcher faced with inductive risks to pause and consider whether their assessment of the foreseeable consequences of their work remains unchanged, and whether they need to adjust how they weigh inductive risk.

In other words, the argument from inductive risk shows that it is generally – and not just in some cases – legitimate for researchers to allow non-epistemic values to have a role in the central stages of the research process. It is legitimate (and in some situations desirable or even an obvious duty) to monitor whether, during the course of the research, any foreseeable consequences have emerged that would warrant adjusting the criteria used in weighing the risks of error. Such vigilance requires non-epistemic values. Therefore, the value-free ideal is a poor ideal for science: it forbids researchers from doing something that is legitimate.

References

Betz, Gregor (2013): 'In defence of the value free ideal,' *European Journal for Philosophy of Science* 3(2): 207–220. URL = https://doi.org/10.1007/s13194-012-0062-x.

Brown, Matthew J. (2024): 'For values in science: Assessing recent arguments for the ideal of value-free science,' *Synthese* 204(4): 112. URL = https://doi.org/10.1007/s11229-024-04762-1.

Douglas, Heather (2009): *Science, Policy, and the Value-Free Ideal*, Pittsburgh: University of Pittsburgh Press.

Hudson, Robert (2016): 'Why We Should Not Reject the Value-Free Ideal of Science,' *Perspectives on Science* 24(2): 167–191. URL = https://doi.org/10.1162/posc_a_00199.

Kuhn, Thomas (1977): 'Objectivity, Value Judgement, and Theory Choice,' in Thomas S. Kuhn, *The Essential Tension: Selected Studies in Scientific Tradition and Change*, Chicago, IL: University of Chicago Press, 320–339.

Menon, Tarun & Stegenga, Jacob (2023): 'Sisyphean science: why value freedom is worth pursuing,' *European Journal for Philosophy of Science* 13(4): 48. https://doi.org/10.1007/s13194-023-00552-7.

Proctor, Robert (1991): *Value-free science? Purity and power in modern knowledge*, Cambridge: Harvard University Press.

Rudner, Richard (1953): 'The Scientist Qua Scientist Makes Value Judgments,' *Philosophy of Science* 20(1): 1–6. URL = https://www.jstor.org/stable/185617.

Weber, Max (1904/1949): 'Objectivity in Social Science and Social Policy,' in E. A.Shils & H. A. Finch (ed. and trans.), *The Methodology of the Social Sciences*, New York: Free Press, 49–112.

Part II
Language

8
Quantifier Phrases with Referential Meanings

Michael Devitt

I am delighted to write a paper in honor of my old friend, Panu Raatikainen. We first met at a conference at the University of St. Andrews in July 2004. The conference was on "Truth and Realism", two things we are both enthusiastic about. We bonded immediately, particularly, I seem to remember, over a whisky tasting. We have been in frequent contact since, in Finland, in Hudson (my home town in upstate New York), in Dubrovnik, Buenos Aires, and many other places around the world. We agree on so much that it would be difficult for me to write a paper criticizing Panu's views. My paper proposes a view of quantifier phrases that I would expect Panu to approve of.

Introduction

Under the influence particularly of Keith Donnellan (1966, 1968), many hold the thesis that definite descriptions are "ambiguous", having not only the "attributive" quantificational meaning captured by Russell but also a "referential" meaning like that of a name or demonstrative. Under the influence particularly of Charles Chastain (1975),[1] some hold the same of indefinite descriptions. I called this thesis "RD" in "The Case for Referential Descriptions" (2004). The present paper will consider whether a similar case can be made that other quantifier phrases have referential meanings. I start by summarizing the case that descriptions have referential meanings.

[1] See also Strawson 1950, Wilson 1978.

Devitt, M. (2025). Quantifier Phrases with Referential Meanings. In A. Laitinen, M. Keinänen, J. Reinikainen & A. Honkasalo (eds.). *Language, Truth, and Reality*. Tampere University Press. (pp. 91–103).
https://doi.org/10.61201/h5fbdk64

It is generally agreed that definite descriptions have a referential *use* as well as an attributive *use*. When ‘the *F*’ is used attributively in ‘The *F* is *G*’ the sentence conveys a thought about whatever is alone in being *F*; when ‘an *F*’ is used attributively in ‘An *F* is *G*’ the sentence conveys a thought about some *F* or other. So, the sentences convey “general” thoughts. When either description is used referentially, its sentence conveys a thought about a particular *F* that the speaker has in mind, about a certain *F*. So, the sentences convey “singular” thoughts.[2]

Despite the agreement that descriptions have these two *uses*, two *speaker* meanings, there is no agreement that they have two *linguistic* meanings. Many, most famously Saul Kripke (1979),[3] accept the quantificational attributive linguistic meaning described by Russell, but appeal to ideas prominent in the work of Paul Grice (1989) to resist the Donnellan-inspired idea that definite descriptions also have a referential linguistic meaning. They argue that the referential use of a definite in an utterance does not affect “what is said” by the utterance. For what is said is the meaning (content) of the Russellian general thought. The meaning of the singular thought is indeed conveyed but only by a “conversational implicature” or the like. So, what is thereby conveyed is not the meaning of the sentence on this occasion and hence not the concern of semantics; rather it is the *speaker* meaning and is the concern of pragmatics.

The Gricean response to referential uses made the embrace of RD seem too hasty because the response raised the possibility that all these uses could be explained pragmatically. This possibility is made very real by the indubitable fact that, with the help of Grice’s “Cooperative Principle” - “Make your conversational contribution such as is required, at the stage at which it occurs, by the accepted purpose or direction of the talk exchange in which you are engaged” (1989: 26) – and related maxims, we seem to be able to give a pragmatic explanation of the referential use of *any* quantifier.

Here is Stephen Neale’s illustration of the point:[4]

> Suppose it is common knowledge that Smith is the only person taking Jones’ seminar. One evening, Jones throws a party and Smith is the only person who turns up. A despondent Jones, when asked the next morning whether his party was well attended, says,
>
> (7) Well, everyone taking my seminar turned up

[2] In calling such a thought “singular”, I am not endorsing the view that the meaning contributed by a referential description in expressing the thought is simply its referent; in my view (2001), it contributes a mode of reference that is partly causal.

[3] For some others, see Neale 1990, King 1988, Ludlow and Neale 1991, and Bach 1994.

[4] Kripke’s comparison of the case of the lover, involving the definite ‘her husband’, with the case of Smith raking the leaves, involving the name ‘Jones’ (Kripke 1979: 15–18) is much less persuasive: the two cases are crucially different (Devitt 1981b: 512–516).

> fully intending to inform me that only Smith attended. The possibility of such a scenario, would not lead us to complicate the semantics of ‘every’ with an ambiguity; i.e., it would not lead us to posit semantically distinct quantificational and referential interpretations of ‘everyone taking my seminar’.
>
> We find a similar situation with plural quantifiers. Suppose that Scott Soames, David Lewis, and I are the only three people in Lewis’s office. Soames has never played cricket and knows that I know this. In addition, Soames wants to know whether Lewis and I have ever played cricket, so I say
>
> (8) Most people in this room have played cricket
>
> fully intending to communicate to Soames that Lewis and I have both played cricket. There is surely no temptation to complicate the semantics of ‘most’ with an ambiguity,… (1990: 87–88)

Neale goes on to argue that Grice’s pragmatic theory of conversational implicature (1989) explains the mechanism by which, in all these scenarios, the speaker conveys a meaning that his words do not literally have. Thus, the theory explains how Neale, by assuming that Jones is acting in accordance with “the Cooperative Principle” and its maxims, derives the implicature (speaker meaning), **Only Smith turned up**, from what Jones literally said (semantic meaning).[5]

In “Case” (2004), I claimed that the case for RD had been greatly underestimated. I argued that the referential uses of both definite and indefinite descriptions exemplify referential meanings: the uses are semantically referential, not merely pragmatically so. A key part of my argument for RD was the rejection of the above Gricean defense of Russell (“Argument I” in Sec. 2).[6] Stephen Neale (2004) aptly named this sort of rejection “the Argument from Convention”. I presented the core of this argument for the referential ‘the *F*’ as follows:

> The basis for RD is not simply that we *can* use a definite referentially, it is that we *regularly* do so. When a person has a singular thought, a thought with a particular *F* object in mind, there is a regularity of her using ‘the *F*’ to express that thought. And there need be no special stage setting enabling

[5] There have also been some non-Gricean pragmatist views of referentially used definites (e.g., Recanati 1989; Bezuidenhout 1997, 2013; Powell 2010), reflecting the influence of the Relevance Theory of Sperber and Wilson (1995). For discussion, see Devitt 2021, ch. 9, particularly pp. 172–175, 178–181. It is often hard to see how these views differ, other than verbally, from the view that definite descriptions are ambiguous.

[6] Earlier presentations of such a rejection are Devitt 1997a: 125–128; 1997b: 388; Reimer 1998. A much less explicit version of the argument is to be found in Devitt 1981b: 316–318. I have often (1974, 1981a) preferred the term ‘designational’ to Donnellan’s ‘referential’.

> her to conversationally imply what she has not literally said, nor any sign that her audience needs to use a Gricean derivation to understand what she means. This regularity is strong evidence that there is a *convention* of using 'the *F*' to express a thought about a particular *F*, that this is a *standard* use. This convention is semantic, as semantic as the one for an attributive use. In each case, there is a convention of using 'the *F*' to express a thought with a certain sort of meaning/content.
>
> 'Every' and other quantifiers are different. There is no convention of using them to convey a thought about a particular object in mind. With special stage setting they certainly can be used for that purpose, as Neale illustrates. But then Grice shows us that with enough stage setting almost any expression can be used to convey almost any thought. (2004: 283)

The idea is that there is a convention for 'the *F*', and implicitly for the plural 'the *Fs*', but not for 'every *F*', that demands "saturation" by the particular object(s) in mind. So, the saturation is semantic. So, 'the' is ambiguous, having both a quantificational meaning that yields attributive descriptions and a referential meaning that yields referential descriptions.

What is it for the speaker to have a particular *F* object *x* in mind in using an expression? It is for the concept that she is thereby expressing to stand in a certain sort of causal relation to *x*, a relation involving the perceptual grounding of someone's thought in *x* and, perhaps, reference borrowings. Or so I have argued (1974, 1981a,b).

The Argument from Convention rests on the assumption that definite descriptions *are* regularly used referentially, *do* regularly have a referential speaker meaning. This is an empirical assumption about usage if ever there was one.[7] The assumption may not yet be supported by scientifically gathered data but it is by every dictionary I have consulted. Presumably lexicographers have arrived at their view by informal observation. This is in order because this regularity, like many others, is obvious to anyone who reflects on her linguistic experiences.[8] My observations lead me to think that most uses of definite descriptions are of "incomplete" ones, ones like 'the table' that fail to uniquely describe an object. And almost all nonanaphoric uses of incomplete ones are referential. All in all, setting aside superlatives and anaphoric uses, I'd guess that the vast majority of uses of definite descriptions are referential. Whether that guess would hold up to scientific testing, the regularity of referential uses surely would.

What is the best explanation of the regularity? The Argument from Convention offers a good explanation: definite descriptions have a referential meaning. I have

[7] Kasia Jaszczolt takes the referential reading of definites to be the default (2005: 106); Alessandro Capone also argues for this view (2019: 118–20).

[8] Given just how obvious it is, one wonders whether Russell's Theory of Descriptions would have been so dominant had it not been proposed by a philosophical giant.

argued (most recently, in 2021: ch. 9) that there is no good pragmatic explanation, let alone a better one. So, we should adopt RD.[9]

Can a similar case be made that other quantifier phrases have referential meanings? I turn to this question now, starting with a discussion of an interesting article by Mario Gómez-Torrente, “Quantifiers and Referential Use” (2015).

Gómez-Torrente

As Gómez-Torrente aptly remarks, Neale’s examples of referential uses are “quirky” (2015: 102). They would not tempt anyone to suppose that they exemplify a semantic convention. But Gómez-Torrente argues that there are many *non*-quirky referential uses of quantifier phrases: “for all typical kinds of quantified determiner phrases … referential uses are frequent and can be perfectly standard, arising in run-of-the-mill contextual scenarios” (p. 98). He sums up:

> as far as frequency and standardness are concerned, the phenomenon of referential uses of quantifier phrases other than descriptions is not significantly different from the phenomenon of referential uses of definite descriptions, after all… (p. 110)

Now, if this were really so, I would of course argue that these uses exemplify referential meanings. For, as Gómez-Torrente well appreciates, the fact that referential uses of definite descriptions are frequent and standard is the basis for the claim that those uses are not to be explained pragmatically but semantically; that’s the Argument from Convention. Gómez-Torrente, however, does not conclude that *any* of these uses of quantifier phrases are to be explained semantically. Nor does he conclude that they are to be explained pragmatically: “semantic theories of referential use … seem empirically feasible, but pragmatic theories seem empirically feasible as well” (p. 123). I think that none of Gómez-Torrente’s examples of the uses of quantifier phrases is referential. Still, I think that there *are* some non-quirky examples.

Consider Gómez-Torrente’s alleged examples of referential uses:

> Let’s go back to Smith’s murder case, but let’s imagine that the police investigation developed somewhat differently. Now we are to imagine that Jones and her colleagues arrested seven people, Adams, Barnes, Crane, Daniels, Evans, Foster and Green, and charged all of them with Smith’s murder; according to the police, they all acted together and played

[9] The Argument from Convention was a key inspiration for the approach to the semantics-pragmatics dispute that I take in *Overlooking Conventions* (2021; see also 2013a). The regular use of *any* expression with a certain speaker meaning provides good evidence that that meaning is conventional and so should be treated as semantic not pragmatic.

> comparable roles in the brutal slaying, and we can suppose that the police are right. Imagine further that Adams, Barnes, Crane, Daniels, Evans, Foster and Green are now standing trial in the dock, and that Jones is again present in the courtroom. Consider the following sentences:
>
> 5. (a) Every murderer of Smith is insane.
>   (b) Every guy in the dock is insane.
> 6. (a) Most murderers of Smith are insane.
>   (b) Most guys in the dock are insane
> 7. (a) Many murderers of Smith are insane.
>   (b) Many guys in the dock are insane.
> 8. (a) Several murderers of Smith are insane.
>   (b) Several guys in the dock are insane.
> 9. (a) Some murderers of Smith are insane.
>   (b) Some guys in the dock are insane.
> 10. (a) A few murderers of Smith are insane.
>   (b) A few guys in the dock are insane.
>
> It is of course easy to imagine utterances of (5)–(10) by which an utterer would not be attempting to communicate contents about any particular persons. But I think it's also easy (and I would say *easier*) to see how, if some of the detainees in the dock behave in suitable ways, Jones can use the quantifier phrases in all of these sentences of the form [Q_x: *x is a murderer of Smith*] *x is insane* intending to communicate, and successfully communicating, a variety of contents involving some particular detainees, meaning in each case that those particular detainees are or provide [Q_x: *x is a murderer of Smith*]; and hence it is mandatory to view the corresponding utterances as containing referential uses of the corresponding quantifier phrases.
>
> Imagine first that all the detainees are moving frantically in the dock. Jones may then make an utterance of either (5a) or (5b) intending to communicate, and successfully managing to communicate to an interlocutor sitting next to her in the courtroom, that *Adams, Barnes, Crane, Daniels, Evans, Foster* and *Green* are insane. Jones' utterance of (5a) thus contains a referential use of "every murderer of Smith" and her utterance of (5b) contains a referential use of "every guy in the dock". (pp. 102–103)

This is ingenious but quite unconvincing. The first thing to note is that generalizations about a domain that are not based on testimony should be, and typically are, *evidentially based* in thoughts about particular objects in the domain. Thus, a biologist expresses the belief that all echidnas have spikes based on observations of certain spikey echidnas; a diner expresses the belief that all of a town's Indian restaurants are cheap based (rashly) on experiences of a few cheap ones; and Jones utters (5a) or (5b) based on observing Adam, Barnes, etc. in the dock. But it obviously does not *follow* that these speakers are intentionally expressing singular thoughts about the particular entities that formed the evidential basis for their generalizations.

Second, it is a familiar logical fact that a universal generalization entails all its instances. So, it follows from (5b) that any particular guy in the dock is insane. So, Jones' interlocutor, who presumably notices that Jones is looking at those particular guys, will quickly infer that Jones is likely to have a singular thought about each guy that he is insane. But it does not follow that Jones intentionally expressed that singular thought as well as the generalization. Indeed, why would she express that thought, given that it can be inferred from the generalization she does express? Lots of contents can be inferred from any utterance beyond the content of the thought intentionally expressed.

Third, if Jones wished to express a singular thought about each of those detainees, there is a conventional way of doing so: "Those/the detainees moving frantically in the dock are insane". Why would Jones not have said that if she simply wanted to convey the singular thoughts about those people? Of course, if (5a) and (5b) exemplify *another* conventional way of expressing such singular thoughts, then we would have an answer. But, as noted, Gómez-Torrente does not claim that (5a) and (5b) exemplify such a convention, and we have been given no reason to believe that they do.

Now consider what Gómez-Torrente has to say about 'most', 'many', and 'several':

> imagine that Adams, Barnes, Crane, Daniels and Evans are moving frantically in the dock, while Foster and Green are calmly seated. If Jones then makes an utterance of either (6a), (6b), (7a), (7b), (8a) or (8b) intending to communicate that *Adams, Barnes, Crane, Daniels* and *Evans* are insane, she will successfully manage to communicate precisely that to an interlocutor sitting next to her in the courtroom. Jones' utterances of (6a), (6b), (7a), (7b), (8a) or (8b) contain referential uses of "most murderers of Smith", "most guys in the dock", "many murderers of Smith", "many guys in the dock", "several murderers of Smith" and "several guys in the dock", respectively. (p. 103)

On what grounds? We have been told, in effect, that singular thoughts about Adams, Barnes, Crane, Daniels and Evans provide the evidential basis for Jones' utterances, but where is the evidence that these utterances are not simply quantificational? If Jones really intended to communicate singular thoughts about those five in particular why would she not do so in the conventional way by saying, "Those/the guys moving frantically in the dock are insane"? Given that she didn't, why should we suppose that by "most guys in the dock", for example, Jones means those particular guys rather than just any old guys in the dock that would constitute most of them? Given that Adams, Barnes, etc. are moving frantically, an interlocuter may indeed infer that singular thoughts about those guys form the evidential base for Jones' utterances but, to repeat, we have no reason to believe that expressing the generalization expresses this evidential base.

Gómez-Torrente is no more persuasive about the referential use of 'some *F*s' and 'a few *F*s'. In sum, generalizations are frequently, although I would not say standardly, used in circumstances where it is apparent to the audience that certain singular thoughts form the evidential base for the generalization. But we have no reason to think that in such circumstances, speakers intentionally express those singular thoughts. Nonetheless, I think that a case can be made that 'many *F*s', 'several *F*s', 'some *F*s', and 'a few *F*s', but not 'every *F*' or 'most *F*s', do have conventional referential uses. A case can be made also for a quantifier phrase not discussed by Gómez-Torrente, the singular 'some *F*'.

A case that certain quantifier phrases have referential meanings

In "Case" (2004: 293-5), I distinguished the referential use of the indefinite 'an *F*' from that of the definite 'the *F*' and the demonstrative 'that *F*' as follows: in using 'the *F*' or 'that *F*' referentially, speaker *S* (intentionally) conveys to the audience *A* that *A* should identify the object *S* has in mind with an object that *A* has in mind *independently* of *S*'s utterance. Abbreviating, we can say that the conventional referential use of 'the *F*' and 'that *F*' is accompanied by a certain "identification expectation". *A*'s having the object in mind "independently" rules out *A*'s having it in mind *simply* as a result of "borrowing" the capacity to do so from *S* via the utterance. *A* must have some other link to the object. This independent link might have been established before the utterance or it might be immediately established by the object's perceptual salience in the context of the utterance; for example, 'the *F*' said while looking at, perhaps gesturing toward, a particular *F*. In the latter sort of case, the utterance prompts a link between *A* and the object that is additional to any that underlie *S*'s utterance; for example, in Donnellan's original story, a person looking at Jones says, "The guy in the dock is insane".

In contrast, *S*'s use of 'an *F*' referentially is *not* (usually) accompanied by the identification expectation: *S* does not (intentionally) convey that *A* should identify the object *S* has in mind with an object that *A* can identify independently. Here is an example from "Case":

> Several of us see a strange man in a red baseball cap lurking about the philosophy office. Later we discover that the *Encyclopedia* is missing. We suspect that man of stealing it. I go home and report our suspicions to my wife: "A man in a red baseball cap stole the *Encyclopedia*.". (2004: 286)

I use the indefinite because I suppose that my wife has no way independent of my remark to identify the suspect I have in mind. But suppose I knew that she had been among those who had observed the man in the red baseball cap lurking in the office.

Then I would very likely have said "The man in the red baseball cap...", or "That man in the red baseball cap ...".

My present thesis is that 'some F', 'some Fs', 'many Fs', 'several Fs', and 'a few Fs' are similar to 'an F' in having conventional referential uses without the identification expectation. The conventional referential uses of those quantifier phrases occur in circumstances where *S* does not (intentionally) convey that *A* should identify the particular object(s) that *S* has in mind with objects that *A* can identify independently.

I start with the argument for the referential 'some *F*' because that is an easy adaptation of the argument in "Case" for the referential 'an *F*': simply replace 'an *F*' with 'some F' in that argument. Thus, take the above *Encyclopedia* story as the example, but replace 'a man' with 'some man' in my report to my wife. So, the report becomes: "Some man in a red baseball cap stole the *Encyclopedia*." I wish to convey a singular thought about the particular person seen in the office not a general thought about just anyone in a red baseball cap, a thought that will be true only if that very man stole the *Encyclopedia*. I convey this thought, with no identification expectation, by using 'some man in a red baseball cap' referentially. That should be uncontroversial. Then, we offer the important Argument from Convention, Argument I (2004: 286–287). Such referential uses are not quirky: they are regular. When a person has a thought with a particular *F* object in mind, there is a regularity of her using, without any special Gricean stage setting, 'some *F*' to express that thought. This is strong evidence that there is a *convention* of using 'some *F*' to express a thought about a particular *F*, that this is a *standard* use. This convention is semantic.

Argument II (p. 288) is inspired by Kripke's idea of "Russell English". We stipulate a language, "Chastain English", in which there is a convention of using 'some *F*', as well as 'an *F*', to express singular thoughts without an identification expectation. The phenomena generated by speakers of this language would not differ from those generated by speakers of English; there would be the same regularities. So, these phenomena confirm that English simply is Chastain English.

Finally, we have Argument III, "Comparison with Deictic Demonstratives" (p. 289). A demonstrative, whether simple or complex, is a device that is regularly used to express singular thoughts about a particular object in mind. So too is the quantifier phrase, 'some *F*'. The devices differ in that demonstratives are (usually) accompanied by an identification expectation but 'some *F*' is not. But they are alike in depending for their reference on a certain sort of causal-perceptual relationship to that object. The referential role of 'some *F*' is as conventional as that of a demonstrative; these roles are semantic not pragmatic.

Turn next to the plural, 'some *F*s', and adapt the *Encyclopedia* story:

> Several of us see uniformed men lurking about the philosophy office. Later we discover that the *Encyclopedia* is missing. We suspect those men of stealing it. I go home and report our suspicions to my wife: "Some uniformed men stole the *Encyclopedia*.".

I do not wish to convey a general thought about uniformed men. Rather, I wish to convey, with no identification expectation, singular thoughts about *each* particular uniformed man that we saw in the office.[10] Each of these thoughts will be true only if the particular man in question stole the *Encyclopedia*. I convey these singular thoughts by using 'some uniformed men' referentially. Once again, we run the Argument from Convention. This use of 'some' seems to be a *conventional* way of conveying such singular thoughts in circumstances like this. Indeed, *how else* could I standardly convey such thoughts? Well, 'a few' would do as well: I could just as easily have said, "A few uniformed men stole the *Encyclopedia*." And, if there were enough uniformed men, it would have been appropriate to use 'many' or 'several'. Perhaps there are some other quantifier phrases that would do.[11] But, I emphasize, 'every' or 'most' would not do; nor would 'few'.[12] *Nor would the demonstrative, 'those', or the description 'the'* (because of their identification expectation). We could also again run an argument inspired by Kripke's stipulation of Russell English.

Suppose that we want to convey, in one simple sentence, with no identification expectation, singular thoughts about each of a group of objects. Then it seems that our *only* conventional ways of doing so are by using certain quantifier phrases.

So, I think that my example exemplifies a referential use of plural quantifier phrases. In the section on Gómez-Torrente above, I argued that his examples do not exemplify such uses. The circumstances of his examples differ crucially from those of mine. In his examples, should *S* (Jones) wish to express certain singular thoughts, she is in a position to convey to *A* that *A* should identify the *F* objects of those thoughts with objects that *A* has in mind independently of *S*; for, *S* surely knows that *A* is looking at the objects. In brief, *S* can convey an identification expectation. So, *S* is likely to use the plural demonstrative, 'those *F*s' or description, 'the *F*s', for *that is the conventional way* for a speaker to express such thoughts in such circumstances. In contrast, in the *Encyclopedia* example, *S* (me), wishing to express certain singular thoughts, knows that *A* (my wife) is not in a position to identify the objects of those thoughts with ones that *A* can identify independently of *S*; *A* has no acquaintance with the objects. In brief, *S* cannot rationally have an identification expectation. So, *S* cannot rationally express those thoughts using a plural demonstrative or description. Indeed, *there seems to be no conventional way to express those thoughts in these circumstances other than to use quantifiers such as 'some', 'many', 'several', and 'a few'.*

[10] Should we say rather that I wish to convey a singular thought about a *group* consisting of those particular uniformed men. This thought will also be true only if each of those very men stole the *Encyclopedia*. But it will be (literally) true only if *there also exists a group* consisting of those men. We should be very reluctant to explain the speaker meaning of such an ordinary use of a quantifier phrase in a way that commits its user to the existence of abstract entities. (This note was prompted by a question from Katarzyna Kijania-Placek.)

[11] Indeed, Antonio Capuano (2024) has recently argued persuasively that *numerical* quantifier phrases have referential meanings. So, if I had three people in mind as the thieves, I could convey my thoughts conventionally by saying, "Three uniformed men stole the *Encyclopedia*". (I was stimulated to write the present paper by blind reviewing Capuano's paper.)

[12] "Why not?", one wonders. I guess that they won't do because 'every', 'most', and 'few' indicate a *proportion* of a group. In contrast, 'some' and the others indicate a *quantity* of a group.

The circumstance of my knowing that my wife *cannot* make the independent identification is analogous to what I called, in discussing 'an *F*' in "Case" (2004: 293), circumstance "(a)". But we could come up with situations where *S* lacks the identification expectation because *S does not want A* to make the identification, perhaps even wants *A* not do so, even if *A* could. That is analogous to circumstance "(b)" in that discussion.

Just as there is a regular use of singular quantifier phrases ('an *F*', 'some *F*') to express *one* singular thought without an identification expectation, there are regular uses of certain plural quantifier phrases to express *more than one* singular thought without an identification expectation. What is the best explanation of this regularity? The Argument from Convention offers an answer: there are *semantic conventions* of using those quantifiers to express such singular thoughts.

For a variety of reasons, people who have singular thoughts about certain *F*s often want to express them in a simple sentence without conveying an intention that *A* identify the objects in mind with objects *A* has independently in mind; people want *A* to "open singular files" for those objects, not add to files independently opened. It would be surprising indeed if there were no conventional way for people to do this. Using the specified quantifier phrases is a brief conventional way to do it.

In sum, the Argument from Convention shows not only that 'the *F*', 'the *Fs*'. and 'an *F*' have a referential meaning but also that 'some *F*', 'some *F*s', 'many *F*s', 'several *F*s', and 'a few *F*s' do.

But doesn't this offend shockingly against Modified Occam's Razor, "Senses are not to be multiplied beyond necessity" (Grice 1989: 47)? The answer depends on how this popular maxim is construed. The maxim is usually understood as advising against the positing of a new conventional sense *wherever an utterance's message can be derived by a pragmatic inference*. Then the answer to our question would be, "Perhaps it does offend". But understood in this way, the maxim is quite false; or so I have argued (2013b: sec. 4; 2021: ch. 8). The maxim should be understood, on the model of the original Occam's Razor, as advising against positing a new sense *unless that sense is needed for the best explanation of the conveyance of the message*. Then the answer is, "No, it does not offend". For, as the Argument from Convention shows, the referential senses of the specified quantifier phrases are needed to best explain their referential uses.[13]

[13] A version of this paper was delivered at a conference, "Philosophy of Linguistics and Language", in Dubrovnik in September 2024. I am grateful for comments it received. And many thanks to Andrea Bianchi for his advice.

References

Bach, Kent (1994): “Conversational Impliciture,” *Mind and Language* 9, 124-162.

Bezuidenhout, Anne (1997): “Pragmatics, Semantic Underdetermination, and the Referential/ Attributive Distinction,” *Mind*, 106, 375–410.

Bezuidenhout, Anne (2013): “The Insignificance of the Referential/Attributive Distinction.” In A. Capone, F. Lo Piparo, and M. Carapezza (eds.), *Perspectives on Pragmatics and Philosophy*. Cham: Springer.

Capone, Alessandro (2019): *Pragmatics and Philosophy: Connections and Ramifications*. Cham: Springer.

Capuano, Antonio (2024): “The Case for Referential Quantifier Phrases,” *Philosophia*. https://doi.org/10.1007/s11406-024-00781-x

Chastain, Charles (1975): “Reference and Context”. In Keith Gunderson (ed.), *Language, Mind, and Knowledge: Minnesota Studies in the Philosophy of Science, Volume VII*. Minneapolis, 194–269.

Devitt, Michael (1974): “Singular Terms”. *Journal of Philosophy* 71 (7), 183205.

Devitt, Michael (1981a): *Designation*. New York: Columbia University Press.

Devitt, Michael (1981b): “Donnellan’s distinction”. In Peter A. French, Theodore E. Uehling Jr., and Howard K. Wettstein (eds.), *Midwest Studies in Philosophy, Volume VI: The Foundations of Analytic Philosophy*. Minneapolis: University of Minnesota Press, 511–524.

Devitt, Michael (1997a): “Meanings and Psychology: A Response to Mark Richard,” *Nous*, 31, 115–131.

Devitt, Michael (1997b): “Responses to the Maribor Papers,” In Dunja Jutronić (ed.), *The Maribor Papers in Naturalized Semantics*. Maribor: Pedagoska fakulteta Maribor, 353–411.

Devitt, Michael (2001): “A Shocking Idea about Meaning,” *Revue Internationale de Philosophie*, 218, 471–494.

Devitt, Michael (2004): “The Case for Referential Descriptions,” In Marga Reimer and Anne Bezuidenhout (eds.), *Descriptions and Beyond*. Oxford: Clarendon Press, 280–305.

Devitt, Michael (2013a): “What Makes a Property ‘Semantic’?” In Alessandro Capone, Franco Lo Piparo, and Marco Carapezza (eds.), *Perspectives on Pragmatics and Philosophy*. Cham: Springer, 87–112.

Devitt, Michael (2013b): “Three Methodological Flaws of Linguistic Pragmatism”. In Carlo Penco and Filippo Domaneschi (eds.), *What is Said and What is Not: The Semantics/ Pragmatics Interface*. Stanford: CSLI Publications, 285–300.

Devitt, Michael (2021): *Overlooking Conventions: The Trouble with Linguistic Pragmatism*. Cham: Springer.

Donnellan, Keith S. (1966): “Reference and Definite Descriptions,” *Philosophical Review*, 75, 281–304.

Donnellan, Keith S. (1968): Putting Humpty Dumpty Together Again,” *Philosophical Review*, 77, 20315.

Gómez-Torrente, Mario (2015): “Quantifiers and Referential Use,” In Alessandro Torza (ed.), *Quantifiers, Quantifiers, and Quantifiers: Themes in Logic, Metaphysics, and Language*. Cham: Springer, 97–124.

Grice, Paul (1989): *Studies in the Way of Words*. Cambridge, MA: Harvard University Press.

Jaszczolt, K. M. (2005): *Default Semantics: Foundations of a Compositional Theory of Acts of Communication*. Oxford: Oxford University Press.

King, Jeffrey C. (1988): “Are Indefinite Descriptions Ambiguous?” *Philosophical Studies*, 53, 417–440.

Kripke, Saul A. (1979): “Speaker’s Reference and Semantic Reference”. In Peter A. French, Theodore E. Uehling Jr., and Howard K. Wettstein (eds.), *Contemporary Perspectives in the Philosophy of Language*. Minneapolis: University of Minnesota Press, 627.

Ludlow, Peter, and Stephen Neale (1991): “Indefinite Descriptions: In Defense of Russell”. *Linguistics and Philosophy*, 14, 171–202.

Neale, Stephen (1990): *Descriptions*. Cambridge, MA: MIT Press.

Neale, Stephen (2004): “This, That, and the Other”. In Marga Reimer and Anne Bezuidenhout (eds.). *Descriptions and beyond*. Oxford: Clarendon Press, 68–182.

Powell, George (2010): *Language, Thought and Reference*. Palgrave Macmillan.

Recanati, François (1989): “Referential/Attributive: A contextualist proposal”. *Philosophical Studies*, 56, 217–49.

Reimer, Marga (1998): “Donnellan’s Distinction/Kripke’s Test”. *Analysis*, 58, 89–100.

Sperber, Dan, and Deirdre Wilson (1995): *Relevance: Communication and Cognition*. 2nd edn. 1st edn. 1986. Oxford: Blackwell Publishers.

Strawson, P. F. (1950): “On Referring”. *Mind*, 59, 320–344.

Wilson, George (1978): “On Definite and Indefinite Descriptions”. *Philosophical Review*, 86, 48–76.

9
No-content explanations

Genoveva Martí

The primacy of content

Typically, explanations of semantic and cognitive phenomena are given by appeal to content. For instance, the fact that two utterances of a sentence have different truth conditions is accounted for by assigning a different content to each. Philosophers inspired by Frege's approach explain differences in the cognitive value of sentences in terms of differences in content. Even philosophers opposed to traditional Fregeanism share with Fregeans the view that differences in cognitive value respond to differences in contents. For instance, proponents of mental files, such as François Recanati (2012), appeal to the contents of mental files entertained by the agent to explain cognitive phenomena. And John Perry has also appealed to contents, or propositions *created* by utterances of sentences. (1988) Belief and action are usually explicated in terms of relations of agents to contents.

There have been, from the very origins of philosophical semantics, important disagreements as regards how contents or propositions should be characterized. For Frege and his followers they are constituted by conceptualizations of the things utterances are about. For Russell and his followers, on the other hand, Mont Blanc with all its snowfields is part of the content an agent expresses and entertains when she thinks or says that Mont Blanc is 4,000 meters high. But in either case, content is at center stage in Fregean and Russellian accounts of language and thought.

Martí, G. (2025). No-content explanations. In A. Laitinen, M. Keinänen, J. Reinikainen & A. Honkasalo (eds.). *Language, Truth, and Reality.* Tampere University Press. (pp. 104–112). https://doi.org/10.61201/h5fbdk64

The primacy of the role of propositional content in semantic and cognitive explanations may be motivated by the conviction that intentionality, or aboutness, is the distinctive mark of the human mind: "As indicated by the meaning of the Latin word *tendere*, which is the etymology of 'intentionality,' the relevant idea behind intentionality is that of mental directedness towards (or attending to) objects, as if the mind were construed as a mental bow whose arrows could be properly aimed at different targets." (Jacob 2023)

Our thoughts and our words have targets: the things and states of affairs we think about and talk about. This much is uncontroversial. But the recognition of the aboutness or directedness of thought and speech has given rise to the presumption that there is a privileged form of explanation of thinking, believing or saying, a form of explanation that is also target-oriented, for it is given essentially in terms of relations to the content expressed by our words and *grasped* by our minds. This is in part due to the assumption that mind and language, thought and speech, go hand in hand, that thinking about something and referring to something are essentially the same phenomenon that requires just one form of explanation.[1] Having established content as the privileged tool with explanatory power in the realm of cognition and semantics, content is appealed to as the answer to fundamental questions about what is believed, what is known, what is said.

The idea that all cognitive and semantic phenomena have to be explained in terms of a *what-is-grasped* or a *what-is-expressed* is simply taken for granted and, as a consequence, the assumption that the explanation of any semantic or cognitive phenomenon is not satisfactory unless some propositional content or other plays the fundamental explanatory role is deeply ingrained.

I think that the assumption is questionable. In fact, I will argue, we can find in the philosophical literature some good explanations of semantic and cognitive phenomena that are not content-oriented. And the only reason to not accept their satisfactoriness is the insistence in clinging to the assumption of the primacy of content. I will argue, however, that if propositional content is conceived heuristically, as a convenient tool, it may have a useful theoretical role and contribute to clarify the phenomena here discussed.

A few no-content explanations

(i) Wettstein on cognitive value: dissolving the puzzle.

Wettstein's (1989) explanation (or dissolution) of Frege's puzzle of cognitive value makes no appeal to propositional content. Where Frege, both in the *Begriffsschrift* and in 'On Sense and Reference', feels compelled to produce two different contents for our minds to grasp, two propositions associated respectively with 'Hesperus

[1] An assumption I do not share, although it will not be my target in this paper.

is Hesperus' and 'Hesperus is Phosphorus', Wettstein sees no need to explain the difference in cognitive value in terms of a what-is-grasped.

Wettstein simply points out that one needs so little information to be competent with the use of proper names, that typically none of the information that a speaker grasps will give her a clue that the two names are co-referential. It is no wonder, then, that the speaker can doubt whether Hesperus is Phosphorus, even after she accepts that Hesperus is Hesperus.

Of course, this kind of explanation is a non-starter from the content-oriented point of view: what is it then, the content devotee asks, that the agent understands when she comes to accept 'Hesperus is Phosphorus' that she didn't understand when she accepted 'Hesperus is Hesperus'? It appears that nothing short of pointing to a content, something that the agent grasps now and didn't grasp before, can satisfy such a demand.

The content devotee's question is, certainly, legitimate. The presumption that only a content answer can satisfy the demand, I think, is not. I'll come back to this issue later, but for the moment I want to think a bit more about the form of Wettstein's explanation.

The puzzle of cognitive value, or informative identity, is often presented as follows: how can a competent speaker who accepts 'Hesperus is Hesperus' as trivial reject, be surprised at, or express doubt about 'Hesperus is Phosphorus'? Wettstein's account gives an answer to this question. It is not an answer that appeals to propositions grasped, to contents targeted by the mind, nor to the different entertained contents of mental files. It is not an answer inspired by the intentional "directed to goals" stance. It doesn't tell us what the agent's mind is directed towards. It is rather an explanation that looks back: it appeals to how the agent came to be in the situation she is in as regards her use of 'Hesperus' and 'Phosphorus'; it appeals to the obvious possibility that the agent came to acquire those names through separate channels that didn't obviously take her back to the same object. And that she, in consequence, associates with those names memories, images, perhaps also pieces of accurate or inaccurate information, happy, unhappy or neutral connotations, . . . that do not carry in their sleeves the condition that they apply to one and the same thing.[2]

(ii) Donnellan on empty names: the importance of the source.

Perhaps the first contemporary explanation of a non-cognitive, purely semantic, problem in terms that are free of an appeal to content targeted, or grasped, by a speaker's mind, is due to Keith Donnellan (1974). Donnellan addresses what is taken to be a serious problem for new theories of reference: the problem of true negative existentials such as 'Santa Claus does not exist'. If the statement is to be significant, according to new theories of reference, it appears that the name should refer; but then, how could we refer to something to say, truly, of it that it doesn't exist?

[2] Wettstein's account applies also to Paderewski-style cases. If the cognitive requirements to be competent with the use of names are in general so poor, it can definitely happen that an agent adds to her vocabulary, twice, the name 'Paderewski', or the name 'Hesperus', under conditions that make it quite possible for her to be surprised when she learns that Paderewski is Paderewski, or that Hesperus is Hesperus.

Donnellan observes that empty names have a history of use and are passed from speaker to speaker much like referring names do. The difference is that, in the case of an empty name, the chain of communication does not lead back to a referent: it ends in a block. This observation forms the basis of Donnellan's explanation of negative existentials. Obviously statements such as 'Santa Claus does not exist' (but also 'Santa Claus is coming tonight') are not just noises, for the name 'Santa Claus' like the referential 'Cicero' has a history of use and it is that history that accounts for its linguistic significance. Thus, significance, for Donnellan, is not to be equated with having a content, or expressing a proposition. The significance of our words depends on their having a stable and consolidated history of use.

'Santa Claus' does not refer, but that does not entail (contra Russell and contra the Fregeans that criticize Millianism) that 'Santa Claus' is a meaningless noise. As for 'Santa Claus does not exist' the alleged problem dissipates: the sentence is true, as most of us think, and it is true because the history of 'Santa Claus' ends in a block.

This explanation will not count as an explanation for the die-hard content devotee. For, as Donnellan himself points out the explanation "does not provide an analysis of such statements; it does not tell us what such statements mean or what proposition they express." (1974, 25).[3]

The question, though, is: what would the assignment of a proposition expressed help explain that Donnellan's explanation doesn't? Of course, assigning a proposition would tell us what 'Santa Claus does not exist' says, i.e., what proposition it expresses; but as the basis for a criterion of adequate explanation, this is a bit circular. There may be, nevertheless, theoretical reasons to insist in assigning a content to 'Santa Claus does not exist', and I will come to them later but, again, for the moment I just want to reflect a bit more on the form of Donnellan's explanation.

Donnellan's account is a paradigmatic historical explanation. Instead of expecting to find an explanation by appeal to what is expressed by an utterance of 'Santa Claus does not exist' (the content that constitutes the target of the agent's utterance and the agent's thought, so to speak), Donnellan invites us to find the explanation looking at the history of how names are bestowed and how they arrive to us. Once we realize that 'Cicero' and 'Santa Claus' arrive to us in pretty much the same way, the presumption that the referring one should have a standard linguistic usage whereas the non-referring one should sound like a meaningless noise falls to pieces. And from there it is a small step to realize that it is precisely the peculiar history of 'Santa Claus' that makes 'Santa Claus does not exist' true.

Donnellan's approach hints also at an explanation that applies to belief and knowledge. It is tempting to say that Mary's belief, which she expresses as 'Aristotle

[3] The content devotee will be tempted to convert that explanation into a content and therefore, in the case of Donnellan, will come up with the result that 'Santa Claus does not exist' actually expresses the content that the history of the name 'Santa Claus' ends in a block; as for Wettstein, the content devotee tells us that his explanation of the difference in informativeness between 'Hesperus is Phosphorus' and 'Hesperus is Hesperus', "... at best . . . suggests a meta- linguistic account of that difference, namely that the former but not the latter sentence implicitly yet informatively declares two different names to have the same bearer." (Glock 2005).

was a philosopher', and Ana's belief, which she expresses as 'Aristóteles era filósofo', are the same belief, because they both believe the same proposition about the same individual. But this account does not tell us why we are so inclined to say that in saying 'Santa Claus is coming tonight' and 'Père Noël arrive ce soir' Tim and Cléo express, or have, the same belief. Here's how Donnellan points to a possible explanation:

> The child who has become disillusioned expresses his new-found knowledge by saying "Santa Claus doesn't exist." A French-speaking child . . . might express his discovery by saying, "Père Noël n'existe pas." Although the names are different, I believe we should want to say that the two children have learned the same fact and, on that account, that they have expressed the same proposition.
>
> What we would like . . . is a reason for saying that both children express the same proposition . . . I want to suggest that we may find such a reason once more by using the idea of a historical connection, that, in our example, it is the blocks in the historical explanation of the use respectively of the names "Santa Claus" and "Père Noël" that are themselves historically connected. (1974, 27, 29)

Of course, we are all part of a tradition in which *the proposition expressed* is at the core of proper semantic explanations. So, Donnellan knows that we would like to have a reason to be able to say that both children express the same proposition. But instead of succumbing to the temptation of providing a proposition, Donnellan encourages us to look elsewhere, and he suggests that it is *the source* of the terms, of the mental states, of the beliefs and of the utterances, not their alleged targets, that plays a crucial role in the explanation.

(iii) Perry's first papers on indexicals: whats and ways.[4]

In his early papers on the semantics of demonstratives and indexicals John Perry (1977) makes a distinction between what an agent says, thinks or believes, and the agent's mental state. The former is roughly what is traditionally known as the content expressed by an utterance of a sentence, something that accounts for what is traditionally thought of as the truth conditions of the utterance.

The latter, on the other hand, can be characterized, he suggests, by the sentences the agent (ideally) would accept in the particular situation at stake (some time afterwards he moved to a characterization of mental states in terms of his more technical notion of *roles*). Those embody the way in which the agent believes (or expresses) the content in question.

When I sincerely utter 'I am about to be attacked by a tiger' and you utter 'she is about to be attacked by a tiger' we both say or believe the same thing, but we believe it in different ways—our mental states are different. The difference in mental states,

[4] See also my (2007) for discussion of Perry's approach.

in ways of believing, according to Perry, accounts in part for our different actions—I try to climb a tree, you go get help. And when we both utter ‘I am hungry', we say different things, but we say them in the same way, our mental state is the same (which explains why we do similar things).

Thoughts are not states, and *objects* of sayings and believings are not *ways* of saying and believing. They are not, because the variation of ways/states is orthogonal to the variation of thoughts/objects. To entertain the same thought *P* we may need different ways at different times and places, for different agents, in different contexts.

Perry’s idea goes against tradition: it entails that there are some cognitive and semantic phenomena that can be explained without appealing to some content that constitutes the mind's target.

The three explanations mentioned here are no-content accounts of some phenomenon or other. As accounts, they do not stem from some independent or theoretical reason to dislike or to reject the idea, or the metaphor, of content. They simply are not constrained by the assumption that an explanation is not complete until a relevant content has been assigned to an utterance, to a belief or to a thought, so they are free to look at other aspects that help explain how the situation or the phenomenon in question has emerged. They are historical explanations, they look at how the phenomena arise, or how the agent comes to be in the position she is.

Content as a convenient tool

By breaking away from the desideratum that only the assignment of content can provide an adequate explanation of semantic and cognitive phenomena we open the door to different forms of explanation. When we learn not to expect the assignment of a content to answer all relevant questions about thought and speech, we may also be able to let content play a partially helpful role.

Let us return to the legitimate questions that the content devotee keeps asking. Once we realize that the assignment of a specific content is not the one and only explanatory tool, the traditional question about cognitive significance–what does an agent learn when she comes to accept ‘Hesperus is Phosphorus’ that she didn’t know when she accepted ‘Hesperus is Hesperus’? – is less theoretically critical.

There is no dangerous commitment to this or that theory if we then say that agents, typically, come to know or understand a variety of things, that different people may learn different things, and that the importance of each one of them may be different depending on the agent and the occasion. For instance, what may be important for some speaker may be captured by saying that she understands that ‘Hesperus’ and ‘Phosphorus’ are names for the same thing. For some other speaker it may be crucial to realize that certain bits and pieces of information that she kept separate (as if they were in different files) apply in fact to one and the same thing (and so that the files can be consolidated). Surely, it might be argued that each one of those things that agents learn or may learn are, after all, contents. True, but in none of these cases we need to say that we have discovered *the specific* privileged content that is going to provide the unique satisfactory account.[5]

Similarly, once we accept the no-content explanation of the presence of empty names in language, the apparently contrived assignment of gappy propositions to sentences containing empty names may have a theoretical *raison d'être*. For if we think of content as a representation of a fragment of the world depicted by an utterance, there is a sense in which there is a representational gap corresponding to 'Santa Claus' in 'Santa Claus wears a red coat'; it even makes good sense to assign the same gappy proposition to 'Pegasus flies' and 'Superman flies', for these two sentences fail to depict fragments of the world for exactly the same reason.[6]

Surely, assigning gappy propositions to sentences containing empty names does not explain what Donnellan's account does explain: how the fact that there is a history of use makes 'Santa Claus' not be a meaningless noise (that explanation does not appeal to any content). But again, if we don't expect the gappy proposition to have a privileged explanatory role, that should not be a problem.

Finally, my sincere utterances of 'I am hungry' can also be characterized as utterances that are true just in case I am speaking and I am hungry, something that competent speakers of the language understand when they understand the utterance. So, the content *the speaker is hungry* can also be assigned to that utterance, even if one accepts the criticisms that direct reference theorists raised against descriptivism. John Perry (2001), and subsequently Kepa Korta and John Perry (2011), have defended a *content-pluralistic* approach to semantics and pragmatics, an approach that simply acknowledges that different propositions with different explanatory roles can be used to classify the different ways in which we can describe what makes an utterance true.[7]

The only problem with this strategy is that it is easy to forget that tools are just tools. Let us remind ourselves of the unfortunate confusions surrounding the notion and the apparatus of possible worlds. The moment one forgets that possible worlds are convenient metaphors for the basic idea that the world might have been different from the way it is, pseudo-problems may start looking like real problems. How can Cicero be in two different worlds? And if he is, how could we know it is him given that he is going to have different properties?

So, it is important to be vigilant and not fall into the content trap. It is important to keep in mind that many different contents may contribute to illuminating and explaining different aspects of a semantic or cognitive phenomenon. And it is also

[5] Observe also that a content such as the one expressed by 'the names "Hesperus" and "Phosphorus" co-name' is meta-linguistic. But as long as we keep in mind that that content is not the explanation of the speaker's reluctance to accept the sentence 'Hesperus is Phosphorus', it is difficult to see what harm there could possibly be in accepting the obvious: that one of the things that finally dawn on us when we accept 'Hesperus is Phosphorus' is that the two names name the same thing.

[6] See (Braun 1993) for a defense of gappy propositions (I very much suspect he would not accept any of the considerations I put forward here).

[7] And, of course, the proposal to appeal to a variety of explanations of what the agent accepts when she accepts 'Hesperus is Phosphorus' can also be read as a move towards liberal content pluralism.

important to keep in mind that in some cases (such as in Wettstein's account of what makes 'Hesperus is Phosphorus' informative, or in the Donnellanian explanation of what makes 'Santa Claus' significant, or in Perry's appeal to ways of saying and believing), assigning content is entirely irrelevant.

If content is relieved from its position as the unique tool with explanatory power, content may have, after all, a legitimate theoretical role in contributing, partially, to *some* explanations.[8]

[8] I thank the participants in the meetings of The Zoom Group, especially John Perry, María de Ponte, Kepa Korta, Shannon Bain, Eros Corazza, Andy Egan, Beñat Esnaola and Carlo Penco for their comments. Work on this publication has been supported by the AGAUR of the Generalitat de Catalunya (2021-SGR-00276) and by the Spanish Ministry of Science and Universities through grant CEX2021-001169-M (funded by MICIU/AEI/10.13039/501100011033) and grant PID2023-150066NB-I00 (MICIU/EU). I am extremely happy to have the opportunity to thank Panu for the many fruitful philosophical conversations we have had throughout the years, and for the many things I have learnt from him.

References

Braun, David (1993): ‘Empty Names’, *Noûs* 27(4): 449–469. URL = https://doi.org/10.2307/2215787

Donnellan, Keith S. (1974): ‘Speaking of Nothing’, *The Philosophical Review* 83(1): 3–31. URL = https://doi.org/10.2307/2183871

Glock, Hans-Johann (2005): ‘Review of Wettstein's *The Magic Prism.’ Notre Dame Philosophical Reviews*. URL = https://ndpr.nd.edu/reviews/the-magic-prism-an-essay-in-the-philosophy-of-language/

Jacob, Pierre (2023): ‘Intentionality’, *The Stanford Encyclopedia of Philosophy* (Spring 2023 Edition), Edward N. Zalta & Uri Nodelman (eds.), URL = https://plato.stanford.edu/archives/spr2023/entries/intentionality/.

Korta, Kepa & Perry, John (2011): *Critical Pragmatics*, Cambridge University Press.

Martí, Genoveva (2007): 'Weak and strong directness: reference and thought', *Philosophy and Phenomenological Research* 74(3): 730–737, URL = https://www.jstor.org/stable/40041080?seq=1

Perry, John (1977): 'Frege on Demonstratives', *The Philosophical Review* 86(4): 474–497. URL = https://doi.org/10.2307/2184564

Perry, John (1979): 'The Problem of the Essential Indexical', *Noûs* 13(1): 3–21. URL = https://doi.org/10.2307/2214792

Perry, John (1988): 'Cognitive Significance and New Theories of Reference', *Noûs* 22(1): 1–18. URL = https://doi.org/10.2307/2215544

Perry, John (2001): *Reference and Reflexivity*, CSLI Publications.

Recanati, Francois (2012): *Mental Files*, Oxford University Press.

Wettstein, Howard (1989): 'Turning the Tables on Frege, or How is it that "Hesperus is Hesperus" is Trivial?', *Philosophical Perspectives* 3: 317–339. URL = https://doi.org/10.2307/2214272

10
The meaning of absurdity[1]

Pasi Valtonen

Introduction

There are two adversarial views on the foundation of meaning. Referentialism claims that the basis of meaning is referential semantics while inferentialism holds that meaning is based on inferential rules. The latter has loose affinities with the Wittgensteinian slogan 'meaning is use'.

Timothy Williamson captures the dispute between referentialism and inferentialism by saying that the difference is the direction of explanation: 'referential[ism] gives center stage to the referential semantics for a language, which is then used to explain the inference rules for the language, [...] as those which preserve truth [...]'. Inferentialism, on the other hand, starts off with inferential rules 'which are then used to explain its referential semantics, [...] as semantics on which the rules preserve truth'. He adds that these directions cannot be combined because it would cause an obvious circularity in the explanation. (Williamson 2009, 137.)

It could be said that the common ground for both directions is truth conditions for connectives. Referentialism builds compositional semantics which yield truth conditions for connectives and inferentialism gives us inferential rules which confirm truth conditions. In this context, Panu Raatikainen claims that

[1] This paper was originally published in Logique et Analyse 239 (2017), 375–288. https://doi.org/10.2143/LEA.239.0.3237160. Reprinted with publisher's permission.

Valtonen, P. (2025). The meaning of absurdity. In A. Laitinen, M. Keinänen, J. Reinikainen & A. Honkasalo (eds.). *Language, Truth, and Reality*. Tampere University Press. (pp. 113–126).
https://doi.org/10.61201/h5fbdk64

referentialism has the upper hand. He utilises Carnap's considerations to reach this conclusion. Raatikainen gives Carnap's considerations a rather unique flair as he aims to convince the superiority of referentialism with Carnap's results. In short, Raatikainen sees Carnap's considerations especially problematic for inferentialism. He says that model-theoretically you can come up with valuations that fail to meet the truth conditions to which *both* accounts subscribe. The crux of the discussion involves ruling out these valuations. Raatikainen claims that referentialism can rule out these evaluations while inferentialism cannot.

Carnap's problem as a problem for inferentialism

The truth condition for negation

$$\text{(NEG)} \quad \neg \text{ A is true} \quad \Longleftrightarrow \quad \text{A is false}$$

is an essential principle in both classical and intuitionistic logic (Raatikainen 2008, 282–284 and Murzi and Hjortland 2009, 481). Carnap has shown a nonnormal model which violates this principle: For any sentence A, both A and ¬A are true. Raatikainen argues that Carnap's problem poses 'a real challenge' for inferentialists like Dummett and Prawitz. On the one hand, they hold that the rules of inference determine the meanings of logical connectives but, on the other hand, they adhere to NEG. 'Yet the standard formalisations of logic (rules of inference) do not rule out non-normal interpretations which violate these principles' (Raatikainen 2008, 284). To illustrate, let there be a classical propositional logic (CPL) in which a set of valuations V for sentences and connectives is produced in a standard recursive manner. Consider an expansion of CPL in the following way. Let there be a set of admissible valuations $V \cup \{v^*\}$ (where for any A, $v^*(A) = T$). As Julien Murzi and Ole Thomassen Hjortland explain, both semantics yield the same (semantic) consequence relation: $\Gamma \vDash_V A$ iff $\Gamma \vDash_{V\cup\{v^*\}} A$. (Since $\vDash_{V\cup\{v^*\}}$ provides no counterexample for $\Gamma \vDash_V A$. Furthermore, assuming $\Gamma \nvDash_V A$, then there is a valuation $v \in V$ according to which every member of Γ is true and yet A is false. Because $v \in V\cup\{v^*\}$, the very same valuation is also in the extended set of valuations. Hence, $\Gamma \nvDash_{V\cup\{v^*\}} A$.)

As a consequence, the formalisation of classical propositional logic, $\vdash_{CPL}$ is sound and complete with respect to standard semantics $\vDash V$ iff it is sound and complete with respect to $\vDash_{V\cup\{v^*\}}$. The problem is that the inferentialist cannot make a distinction between the semantics on the basis of soundness and completeness results. Yet there is a big difference. In $\vDash V_{\cup\{v^*\}}$, NEG 'fails massively'. (More elaborate exposition in Murzi and Hjortland 2009, 480–481.) Raatikainen concludes that inferentialism owes us an explanation as to how the problem is circumvented (Raatikainen 2008, 287).

It should be pointed out that the troublesome valuation surely affects other connectives too. For example, it is a rather intuitive thought that 'A ∧¬A' is false but with the non-standard valuation this comes out as true. It is also an intuitive thought that 'A → ¬A' is false when A is true but that is not the case with $\vDash_{V\cup\{v^*\}}$. However, NEG is the only truth condition that directly relies on the fact that A and ¬A cannot be true at same time. That is why NEG deserves special attention.[2]

Murzi and Hjortland's intuitionistic solution

Murzi and Hjortland stress that Raatikainen's dismissal of Dummett and Prawitz is too quick and they establish an inferentialist response based on Dummett and Prawitz's work. I claim that Murzi and Hjortland do not provide a solution to the problem, at least not without serious caveats. I then introduce my solution which is based on Neil Tennant's work. The solution takes Tennant's adherence to paraconsistency seriously. On the basis of this, I offer a solution which explicates negation with the principle of consistency. The key ingredient in the solution is that absurdity is viewed as a primitive expression of the principle of consistency. Finally, Murzi and Hjortland are sceptical about a bilateralist solution. I go on to show that the developed view can contribute to the bilateral solution. I argue that if a bilateralist adopts a similar paraconsistent view, then Carnap's problem is not a threat to her, contrary to Murzi and Hjortland's claim. Initially, one might object that the problem arises because of the non-normal valuation and it should be ruled as inadmissible in the first place since it violates NEG. But as Murzi and Hjortland point out, this misses Raatikainen's point:

> [I]f meanings are to be determined by the inferential rules, and if meanings are truth-conditions, logical inferentialists cannot legitimately appeal to NEG [...], on the pain of invoking a previous knowledge of the meanings they are trying to capture (Murzi and Hjortland 2009, 481).

In short, an inferentialist cannot appeal to semantics to justify the inferential rules. It should be the other way round. The situation that both A and ¬A are true has to be ruled out with the inferential rules which ultimately determine the truth-conditions such as NEG. Only after this, the inferentialist can commit to NEG.

First, intuitionists equate truth with proof. Thus, the investigation is narrowed to excluding the possibility that there is a case where both A and ¬A are provable (Murzi and Hjortland 2009, 483). Essentially, Murzi and Hjortland's solution to the proof-theoretic version of Carnap's problem relies on two points:

On canonical proof: A proof whose last step is an introduction rule.

On Prawitzian view on absurdity: The introduction rule for ⊥ is null.

They introduce Brouwer-Heyting-Kolmogorov (BHK) clauses to specify proofconditions for complex statements. BHK clauses define ¬A as A →⊥. (Murzi and Hjortland 2009, 483.) Given BHK clauses, Carnap's situation is that A is proven and

[2] Raatikainen also discusses disjunction at length but I think Murzi and Hjortland's response to that is satisfactory so here I am concentrating solely on negation. (Raatikainen 2008, 283–284 and Murzi and Hjortland 2009, 483–484.)

that $\neg A$ is also proven. However, the latter is blocked with Prawitz's view of absurdity. According to him, the meaning of $\bot$ is determined by a null introduction rule and the elimination rule is absurdity rule:

$$(\bot\text{ -E})\frac{\bot}{A}$$

in which A can be substituted with any atomic sentence of the language (Prawitz 1973, 243). Murzi and Hjortland argue that the null introduction rule makes sure that any proof of $A \to \bot$ cannot satisfy the notion of canonical proof. Because the introduction rule is null, there simply is not a way to use the rule as a last step in a proof. (Murzi and Hjortland 2009, 483–484.) There seems to be something rather odd with this response. It seems as if Murzi and Hjortland are saying that there is no way to introduce negation because the rule for the introduction of absurdity is empty. But concerning the introduction of $\neg A$, the introduction rule for absurdity is the wrong rule. The (intuitionistic) rules for negation are

$$(\neg\text{-I})\quad \frac{\begin{matrix}A\\ \vdots\\ \bot\end{matrix}}{\neg A} \qquad (\neg\text{-E})\quad \frac{A \quad \neg A}{\bot}$$

An inferentialist can introduce negation (and thereby prove $\neg A$ in canonical way) with $\neg$-I rule. For surely there must be a way to prove negated claims within the inferentialist system.[3] The upshot is that there is a perfectly good way to introduce $\neg A$. The real question is whether you can introduce $\neg A$ (with $\neg$-I) given that A is already proven.

Murzi and Hjortland are right in their contention the conception of absurdity is important in solving Carnap's problem but they should have concentrated on explicating the connection between negation and absurdity.

Hand's criticism of Dummett

Harmony between the introduction and elimination rules guarantees that nothing is added to (or left out from) the elimination rule (in respect to the relevant introduction rule). The elimination rule can only unpack the information that the introduction rule packed in. (e.g. Rumfitt 2000, 782–792). Neil Tennant criticises Prawitz's conception of absurdity because the question about harmony cannot be properly investigated. He says that it is unnatural that an introduced concept has only an elimination rule,

[3] I owe this point to the referee(s). The referee(s) also point(s) out that Murzi and Hjortland's solution applies only to proofs, not to deductions in general. The referee(s) go(es) on to point out that, surely, NEG applies to all sentences, not just proofs. In my view, this points to the fact that Murzi and Hjortland's solution is not entirely satisfactory. However, I am willing to disregard this asymmetry and concentrate on the more serious problem that Murzi and Hjortland place too much weight on the wrong rule.

'with no introduction rule to which it is genuinely answerable' (Tennant 1999, 216). Dummett also sees this problematic. He says that it is a usual practice not to impose an introduction rule for ⊥. He suspects the motivation for this is an implicit appeal to principle of consistency. In order to harmonise the introduction and elimination rules for ⊥, Dummett proposes the following introduction rule:

$$(\bot\text{-I}) \quad \frac{A \quad B \quad C \quad \cdots}{\bot}$$

where A, B, C, ...are atomic sentences of the language. The idea is that the premise set includes all the atomic sentences of the language. Dummett comments: 'The constant sentence ⊥ is no more problematic than the universal quantifier: it is simply the conjunction of all atomic sentences'. This harmonises the elimination rule from which one can infer any atomic sentence of the language.[4] At the same time, the intuitive thought is that no language is consistent and you are bound hit inconsistency at some point. However, Dummett himself observes that the intuitive thought is beside the point. As far as logic is concerned, a language *L* might be consistent. Dummett thinks that the principle of consistency is not a logical law. (Dummett 1993, 295–296.)

In his comments on Dummett's proposal, Michael Hand makes much of Dummett's point that a language need not to be inconsistent. He first admits that Dummett's introduction and elimination rules do explicate the meaning of ⊥ in a harmonious way:

> The answer to the question of the meaning of ⊥ is now obvious: it has precisely the same logical power that a conjunction of all atoms other than ⊥ would have, if we had infinitary conjunction in the language. (Hand 1999, 189).

But then he goes on to criticise the proposal:

> One's first reaction to this observation should be to balk. Surely there is something wrong if we cannot fix the truth-condition, not to mention the meaning, of ⊥ any better than this. The constant ⊥ is supposed to be false, and if meaning is use, then our rules governing ⊥ had better make it so. What Dummett points out is that the intuitionistic rules cannot even prevent ⊥ from meaning something that might be true [...] To put it differently, [the question with ⊥ is] why are we unable to formulate rules ensuring that

[4] The adding of an introduction rule for ⊥ of course undermines Murzi and Hjortland's solution but that is not the issue because it already turned out that the rules for negation should have been the real issue in their solution. Also given that Dummett adds the introduction rule, it is odd that Murzi and Hjortland insist that they show that Carnap's results are not a problem for Dummett since their solution rests on the absence of the introduction rule.

> assignments meet the consistency condition, i.e. that a sentence and its negation are not both true?

In the present context, Hand's observation anticipates Carnap's problem. Since the essence of the problem is that according to the assignment both A and ¬A are true, Hand's point only emphasises Raatikainen's claim: The current inferentialist conception of ⊥ does not rule out Carnap's problem. In the following, I will introduce an alternative way to view absurdity. I claim that the view has a central role in the solution to Carnap's problem.

Absurdity based on semantics

The solution to Carnap's problem which I am offering is based on Tennant's thinking. But before going into Tennant's thinking in detail, I will make a clarifying point concerning Hand. He advocates a semantic view of absurdity. He starts his exposition with an observation that a false sentence does not mean that it is equivalent with a conjunction of all atoms. He continues: 'To say that a sentence is false is to say something much worse.' (Hand 1999, 192). The challenge is to frame this badness and the inferential rules alone cannot explicate the badness of falsity. Hand makes the following proposal:

> False sentences are bad because they fail. This failure is a semantical phenomenon, and purely intralinguistic rules cannot be formulated to characterise it. Intralinguistic rules can be formulated for contradictions, of course: if a person asserts one, reject it immediately. Nonetheless, this rejection is based on the realisation of a semantical fact about the claim, viz., that it is bound to fail. (Hand 1999, 194.)

Hand proposes that falsity is based on pragmatic and normative obligation. To avoid to assert something which turns out to be false is the primary obligation of an assertor. Importantly, this obligation can only be framed semantically, not inferentially. Hand presents a debate concerning his dachshund in the backyard:

> The important point is that this obligation cannot be explained except in overtly semantical terms, as far as I can see. When I said that my dachshund was in the back yard, you looked for him. You sought the referent of the name, to see whether it satisfied the predicate. Your discovery was that it did not, and the fact that it did not is just what makes it the case that I failed in my linguistic obligation to avoid falsehoods. (Hand 1999, 197.)

This *semantic* realisation gives us the principle of consistency, that A and ¬A are incompatible. The fact that this realisation is semantic seems to be another point for referentialism. At this point, the exact accusation against Murzi and Hjortland could be phrased that they aimed to provide a solution to Carnap's problem which then turned out to rest on insufficient explication of absurdity. According to Hand, the explication needs to be supplemented with a *semantic* explanation why a false sentence is a bad thing. However, I do not think this is insurmountable for inferentialism. I agree with Hand that something beyond inferential rules is needed but I do not think semantics is the only place to look for this.

In sum, Murzi and Hjortland's proposal is disappointing for two reasons. First, they do not pay enough attention to the rules for negation. Secondly, their view on absurdity is insufficient. In the following, I will bring clarity to both points. First an alternative conception of absurdity is introduced. This conception respects the principle of consistency and the inferentialist order of explanation. Then I will use this conception to clarify the notion of negation in the sense that the rules for negation provide a solution to Carnap's problem.

Tennant's paraconsistency, concept mastery and intuitionistic solution

Tennant is also concerned with the badness of absurdity: 'The source of the 'badness' that ⊥ seeks to register is contrariety' (Tennant 1999, 216). He thinks that contradiction 'is a matter of deep metaphysical necessity' (Tennant 2004, 362). According to Tennant, this separates his view from Dummett's:

> Whereas Dummett seeks a logical basis for metaphysics, I think we need, at this point, to put it the other way round. One needs a metaphysical basis for logic, insofar as we seek an origin for our grasp of the meaning of negation. I believe this is to be found in our sense of contrariety [...] (Tennant 1999, 217.)

Tennant sees the order of explanation as a crucial matter. It is my contention that Tennant does agree with Hand in this. Tennant sees ⊥-I and ⊥-E rules as a *logical* explication of absurdity but neither for Hand nor for Tennant that will do. Hand thinks that at the heart of falsity is a semantic explanation why false sentences are bad. Tennant thinks that the proper way to go is to explicate badness of absurdity metaphysically. At the same, both views make a distinction to Dummett. However, I suggest that there is a third way which looks for the basis of absurdity elsewhere but, in broad terms, stays faithful to Dummett. If we look at Tennant's 'metaphysics' more carefully, we can see that he does not go very far from Dummett. To elaborate, it seems to me that there are two versions of Tennant's view. The first view he presents in 'Negation, Absurdity and Contrariety'. He notes that the consistent language which Hand toys with is not actually learnable. According to Tennant, contraries among atomic sentences are crucial in learnability. Our grasp of different concepts depends on their patterns of instantiation, i.e. the grasp of concepts is based on different extensions. (Tennant 1999, 216-218.) In my view, this is not a very good *rebuttal* of Hand's overall point. Hand's point is that the inferentialist conception of absurdity needs to be supplemented with a semantic explanation and Tennant just seems to confirm this. I think his second proposal is better. The second view emphasises concept mastery. In 'An Anti-Realist Critique of Dialetheism', Tennant holds that some antonym-pairs derive from the structure of our phenomenology (Tennant 2004, 362.) For example, any competent speaker knows that an object cannot be solidly red and solidly green at the same time. For this reason, any competent language user can make the transition from Hot, Cold to ⊥. Tennant says that this realisation stems from the mastery of 'hot' and 'cold'. A child can learn what 'cold' means without knowing what 'not-hot' means. He says that contraries that he is talking about differ from sensory experience in that they are *a priori*. That is why the contraries do not have much to do with acquisition, i.e. sensory experience and 'everything to do with mastery'. (Tennnant 2004, 361–362.) Finally, you can ask where does this mastery stem from. In the present context, there are two possibilities: semantics and the inferential rules. According to Tennant's first story, the basis of absurdity is semantic. Realisation of absurdity is based on extension of contrary concepts like 'hot' and 'cold'. In the second story, this part is open. So the second story can accommodate inferentialism. The mastery of 'hot' and 'cold' could be explained with the inferential patterns concerning these concepts. In this case, no reference to semantics is needed. It seems to me that this kind of explanation is still compatible with Dummett's inferentialist conception of concept mastery (albeit it might not be compatible with Dummett's view of absurdity).

It is essential to understand the contrast between Hand's proposal and Tennant's proposal regarding the order of the explanation. Hand thinks that a false sentence is the primitive notion and this is then intimately connected with negation as NEG involves falsity. According to Hand, a contradiction is just a special case of a false sentence. It is always false. In distinction, Tennant thinks that contrariety is the primitive notion and then 'the conception of contrariety is expressed by means of an inferential transition from the contraries in question to absurdity' (Tennant 2004, 363). After this, Tennant moves on to explicate negation with the usual (intuitionistic) rules (re-introduced as a reminder):

$$(\neg\text{-I})\qquad \dfrac{\begin{array}{c}A\\ \vdots\\ \bot\end{array}}{\neg A}\qquad\qquad (\neg\text{-E})\qquad \dfrac{A\quad \neg A}{\bot}$$

It still remains to be seen how Tennant's view differs from Dummett's account and how Tennant's view provides a solution to Carnap's problem. Tennant is a relevantist and an essential part of his relevantism is paraconsistency, characterised as a rejection of absurdity rule, i.e. ⊥-E rule.[5] Tennant sees ⊥ as a (structural) punctuation mark. It represents a logical dead-end. (Tennant 1999, 200–205.) Since the sign does not have any propositional content, it is not subject to introduction or elimination rules.

To start the positive contribution of Tennant's paraconsistency, let us make the following observation. On the basis of Tennant's paraconsistent understanding of absurdity, we should not need any logical explanation of the badness. Whenever absurdity is derived, we should shout 'enough already' (Tennant 2004, 358). That is the point of ⊥ and there is no need to show any additional *logical* badness of ⊥ with the absurdity rule. The badness is in the derivation of absurdity itself. Tennant's version of paraconsistency appeals to the principle of consistency: it cannot be consistent to assert A and ¬A at the same time. When absurdity sign appears in ¬-I and ¬-E rules, it precludes any explicit definition of negation. It does not yield any propositional content to the definition of negation (such as $A \rightarrow \bot$). Instead, the rules for negation give us instructions how to use negation in an inference.

Given all this, the solution to Carnap's problem was hidden in plain sight all along. ¬-E states that to claim that A and ¬A are both true leads to a *logical* dead-end. The contrast to Dummett's ⊥-I and ⊥-E rules is that Dummett's rules allow to equate ⊥ with the conjunction of all sentences of language but it does say anything about the semantics of the language. As far as the rules for ⊥ are concerned, all of the sentences might be true. Whereas, with Tennant's conception: 'There is no question – the possibility simply cannot arise – of ⊥ [...] ever being true. And that is why negation works in such a way that it could never be the case that both P and ¬P were true.' (Tennant 2004, 364.) This gives the inferentialist the armoury to respond to the referentialist (or anyone) who proposes $V \cup \{v^*\}$ as an admissable valuation. The inferentialist can point out that *from the inferentialist point of view* the valuations are highly problematic since the valuations allow that A and ¬A are both true and this is an absolute logical dead-end. Most importantly, the inferential rules for negation can be viewed as meaning constituting rules so that they yield NEG as truth conditions for negation in a standard way.

Paraconsistency and bilateral solution to Carnap's problem

It is clear that Rumfitt's bilateralism aims to justify classical logic but he does this in an unusual way. As Raatikainen points out, usually referentialism comes with a realist notion of truth, i.e. evidence-transcendent notion of truth and semantic anti-realist like Dummett and Prawitz adhere to warranted assertability (Raatikainen

[5] Here paraconsistency is understood as just that and nothing more. It has to made clear that Tennant's relevantism or paraconsistency is not 'inconsistency-friendly' in that it claims that not all inconsistencies lapse into absurdity. On the contrary, Tennant claims that all contrarieties do lapse into absurdity but he claims it without an appeal to absurdity rule because Tennant does not hold that the badness of absurdity is that it entails everything. More on this below.

2009, 282–283). It is precisely the adherence to the correspondence notion of truth which justifies classical logic. Because truth is not up to us, we can consider truth to be bivalent. This bivalence then justifies the crucial classical rules like Law of Excluded Middle (LEM) and Double Negation Elimination rule (DNE) even in undecidable discourse and even if DNE is non-harmonious by the inferentialist standards. Rumfitt's bilateralism provides a novel defence for classical logic. It concedes that truth might be equated with warranted assertability and hence it 'concedes the anti-realist standards for the justification of rules of inference', as Imogen Dickie points out (Dickie 2010, 163). This is the novelty in bilateralism: to admit the anti-realistic starting point in inferentialism and to justify classical logic anyway. I think Rumfitt maintains this strategy in his "Yes' and 'No" (Rumfitt 2000, 781-824). In a later response to Dummett's criticism, Rumfitt somewhat retracts this position. In his 'Unilateralism Disarmed: A Reply to Dummett and Gibbard' (2002, 305–321), Rumfitt writes:

> The oddity only arises, however, if truth is equated with the correctness of assertion and falsity with the correctness of denial; and I accept neither of these equations as generally correct theses about truth and falsity. For both Dummett and me, the notion of correctness is epistemic: to say that it is (objectively) correct to assert (or to deny) a sentence A is to say that knowledge is (tenselessly) available which, were a speaker to apprehend it, would warrant him in asserting (or in denying) A. As Dummett's reply makes clear, he wishes to equate, always and everywhere, the truth of a sentence with its being correct to assert it. I allow that there may be theories for which this conception of truth is correct; in the original paper ["Yes' and 'No"] I cited elementary arithmetic as a possible example. (Rumfitt 2002, 313.)

It seems to me that Rumfitt is proposing some form of truth pluralism here: Sometimes truth is an epistemically constrained notion and sometimes it is not. However, in my view this is a retrograding step as it takes the novelty out of bilateralism. The interest in bilateralism rests on the fact that it admits the anti-realistic starting point, truth can be equated with warranted assertability, and it still aims to justify classical logic. As soon as Rumfitt admits that truth is evidence-transcendent, bilateralism becomes redundant as the usual defence for classical logic is also available. That is why I will assume that Rumfitt adheres to warranted assertability or epistemically constrained notion of truth. At the very least, I am restricting the discourse under discussion to arithmetic, i.e. to discourses to which warranted assertability applies even by Rumfitt's standards.

That being said, Murzi and Hjortland are sceptical whether a classical bilateralist like Rumfitt is equipped to cope with Carnap's problem. Bilateralism recognises, in addition to assertion, an act of rejection. These acts are introduced to the formalisation as + A (assertion of A) and – A (rejection of A). Rumfitt forms the introduction and elimination rules for negation accordingly:

$$(+\text{-}\neg\text{-I})\ \frac{-A}{+(\neg A)} \qquad (+\text{-}\neg\text{-E})\,\frac{+(\neg A)}{-A}$$

$$(-\text{-}\neg\text{-I})\ \frac{+A}{-(\neg A)} \qquad (-\text{-}\neg\text{-E})\,\frac{-(\neg A)}{+A}$$

The main idea behind these rules is that they yield DNE in a way that satisfies the demand for harmony. Given +-¬-E and ¬¬A, we have +(¬¬A) ⊢ –(¬A) and given – -¬-E, we have –(¬A) ⊢ +A. Thereby, we have DNE: the rules take you from the assertion of ¬¬A to assertion of A.

Murzi and Hjortland suggest that the bilateralist's attempt to appeal to these rules fail to block Carnap's problem. Let +A and –A be signed formulae for any A ∈ WFF (well-formed formula) and let $\mathrm{WFF}_{\mathrm{sign}}$ be the set of signed formulae. In this case, They appeal to +-¬-E and to the set of correctness-valuations C for signed formulae with the following correctness clauses:[6]

$$(\mathrm{C1})\mathrm{v_c}(+\mathrm{A}) = \mathrm{T} \ \Leftrightarrow \mathrm{v}(\mathrm{A}) = \mathrm{T}$$
$$(\mathrm{C2})\mathrm{v_c}(-\mathrm{A}) = \mathrm{T} \ \Leftrightarrow \mathrm{v}(\mathrm{A}) = \mathrm{F}$$

They also define validity for signed formulae:

(VAL) Γ ⊨ α is valid in the case, for every correctness-valuation $v_c \in$ C, if $v_c(\beta)$ = T for every $\beta \in \Gamma$, then $v_c(\alpha)$ = T.

It appears that C2, VAL and +-¬-E block together Carnap's problem because it says that A and ¬A are both correct. Hence assertion of ¬A, i.e. + (¬A) is correct but since A is correct too, – A cannot be correct. On the basis of C2 and VAL, +-¬-E fails.

On the first appearance, Carnap's problem seems to be solved but actually it just shifts level. For the non-normal valuation can appear at the level of signed formulae (containing + and –): let there be valuation $v^*_c(\alpha)$ = T for every α ∈ $\mathrm{WFF}_{\mathrm{sign}}$. This is the troubleshooting valuation which creates Carnap's problem in the first place now applied to bilateral signed formulae. Both A and ¬A are true and more disturbingly +-¬-E is valid according to VAL: 'the assertability of the premises guarantees the assertability of the conclusion' since for every signed formulae, v^*_c(a) = T. So the bilateral rules for negation do not block Carnap's problem any more. The troublemaking valuation still violates C2 but the appeal to C2 is problematic. Murzi and Hjortland argue that syntactically C2 and NEG are exactly alike and nothing in the bilateral system prevents to view rejection as a special kind of (non-iterative) negation. Hence, C2 is comparable to NEG. So why is it all right to appeal to C2 but not to NEG? (Murzi and Hjortland 2009, 485–486.) I agree that if C2 was the only resource, bilateralism would indeed be in trouble.

[6] In my formulation, 'T' stands for epistemically constrained truth which applies only to decidable statements. In broad terms then it coins with correct assertability. It is clear that as an intuitionist, Tennant subscribes to epistemically constrained notion of truth and as I explained above, I assume that Rumfitt subscribes to this too. (See Tennant 1997, 173–177 and Rumfitt 2000, 817–820.)

Nevertheless, Murzi and Hjortland pay too little attention to the bilateral rules for negation. Rumfitt holds that bilateral logic contains a co-ordination principle:

$$(\mathrm{COP}) \quad \frac{+\mathrm{A} \quad -\mathrm{A}}{\perp}$$

Murzi and Hjortland approve this as a bilateral Law of Non-Contradiction

$$\left(\mathrm{LNC}^{*}\right)\ \alpha,\ \alpha^{*}\ \vdash\perp,$$

where α^* is the result of reversing the sign of α. (Rumfitt 2000, 816 and Murzi and Hjortland 2009, 485.) The bilateral solution rests on a seemingly trivial observation that because of – -¬-I and – -¬-E, + A and – (¬A) are interdeducible and more importantly because of +-¬-I and +-¬-E, – A and + (¬A) are interdeducible, we have – A ⊣⊢ +(¬A). Hence:

$$+A,\ -A\ \ \vdash\perp\ \Longleftrightarrow\ +A,\ +(\neg A)\vdash\perp$$

Thus, the bilateral inferential rules do rule out Carnap's problem for unsigned and signed formulae since the right-hand side of the equivalence says that A and ¬A cannot be both asserted at the same time and the left-hand side rules out valuation $v^*_c(\alpha) = \mathrm{T}$.[7]

In my view, the previous solution works only if we take Tennant's paraconsistent view on ⊥. It is not clear whether Rumfitt actually takes that view but, in the face of Carnap's problem, it might be beneficial. It seems to me that Rumfitt does take some initial steps towards paraconsistency. He says: '[I]t would be perverse to try to assign a propositional content to the expression 'contradiction'. Rather, as Tennant puts it, the expression plays the role of a punctuation mark in deduction' (Rumfitt 2000, 793–794). I take this as a sign that Rumfitt thinks that ⊥ has no propositional content to be clarified with introduction and elimination rules. As a result, Rumfitt admits that intuitionism has the advantage at least in one respect. The intuitionistic rules for negation express in a very direct manner the principle of consistency whereas the bilateral rules do not. For the intuitionistic elimination rule for negation (¬-E above) is a unilateral equivalent of the co-ordination principle (as the above equivalence shows). Therefore, the classical bilateral rules must be coordinated so that they preserve the principle of consistency and they must preserve it in such a way that 'the co-ordination principle (and hence the principle of consistency) holds for complex formulae [+(¬A) and –(¬A)] as well as for atomic ones [+A and –A]' and 'such co-ordination will be necessary if +A and +(¬A) are themselves to be contradictory' (Rumfitt 2000, 815–816). With the troublesome valuation, either side of the equivalence becomes a logical dead-end and that is why the bilateralist can repeat the intuitionistic response.

[7] It should be noted that this solution does not depend on the way Hjortland and Murzi formulate the problem. Especially, the current solution does not depend on VAL. In fact, it is likely that VAL needs to be adjusted to accommodate the non-classical consequence in paraconsistent logic. See below.

Conclusion

I agree with Murzi and Hjortland's overall contention. Inferentialism can overcome Carnap's problem. However, there are four qualifications to be made: (i) The analysis of Murzi and Hjortland shows some weak points regarding the connection between absurdity and negation. (ii) Dummett's proposal for absurdity does not clarify the notion falsehood properly. (iii) The paraconsistent solution does better by appealing explicitly to the principle of consistency. Absurdity is a primitive expression of the principle of consistency. Hence, the rules for negation reflect that A and ¬A cannot be asserted at the same time. (iv) A classical inferentialist can solve the problem by adhering to paraconsistency.

References

Dummett, Michael (1993): *The logical Basis of Metaphysics*, Cambridge, Mass.: Harvard University Press.

Hand, Michael (1999): 'Antirealism and Falsity', in D. Gabbay and H. Wansing (eds.), *What is Negation?*, Dordrecht: Kluwer Academic Publishers, 185–198. URL = https://doi.org/10.1007/978-94-015-9309-0_9

Kürbis, Nils (2005): 'What Is Wrong with Classical Negation?' *Grazer Philosophische Studien* 92(1): 51–85. URL = https://doi.org/10.1163/9789004310841_004

Murzi, Julien & Hjortland, Ole Thomassen (2009): 'Inferentialism and the Categoricity Problem: Reply to Raatikainen', *Analysis* 69(3): 480–487. URL = https://www.jstor.org/stable/40607663

Prawitz, Dag (1973): "Towards a Foundation of a General Proof Theory", in P. Suppes, L. Henkin, A. Joja & G. C. Moisil (eds.), *Logic, Methodology and the Philosophy of Science IV: Proceedings of the Fourth International Congress*, Amsterdam: North Holland, 225–250. URL = https://doi.org/10.1016/S0049-237X(09)70361-1

Raatikainen, Panu (2008): 'On rules of Inference and the Meanings of Logical Constant', *Analysis* 68(4): 282–287. URL = https://doi.org/10.1093/analys/68.4.282

Rumfitt, Ian (2000): ''Yes' and 'No'', *Mind* 109(436): 781–824. URL = https://doi.org/10.1093/mind/109.436.781

Rumfitt, Ian (2002): 'Unilateralism Disarmed: A Reply to Dummett and Gibbard', *Mind* 111(442): 305–321. URL = https://www.jstor.org/stable/3093712?seq=1

Tennant, Neil (1999): 'Negation, Absurdity and Contrariety', in D. Gabbay and H. Wansing (eds.), *What is Negation,* Dordrecht: Kluwer Academic Publishers, 199–222. URL = https://doi.org/10.1007/978-94-015-9309-0_10

Tennant, Neil (1997): *Taming Of the True*, Oxford: Oxford University Press.

Tennant, Neil (2004): 'An Anti-realist Critique of Dialetheism', in G. Priest and B. Amour-Garb (eds.), *The Law of Non-Contradiction: New Philosophical Essays*, Oxford: Oxford University Press, 355-384.

Williamson, Timothy (2009): 'Reference, Inference, and the Semantics of Pejoratives', in J. Almog and P. Leonardi (eds.), *The Philosophy of David Kaplan,* Oxford: Oxford University Press, 137–158.

11
Questions of reference

Jaakko Reinikainen

This paper defends a piece of conventional wisdom (that descriptivism fails) with conventional arguments (namely, from incompleteness and redundancy) against a recent case made by Jens Kipper and Zeynep Soysal.

Introduction

A fond adage has it that in philosophy, the questions tend to be more important than the answers. Personally, this lesson came to me most concretely from my PhD supervisor, Professor Raatikainen, who above all has taught me to question my questions.

Let me illustrate with an anecdote. I first read the paper 'Theories of reference: What was the question?' (Raatikainen 2020) in October 2020. In my notes, preserved pristine in their digital form, I wrote:

> And the answer to 'What reference is?' is that it is whatever subjects ought to take as serving as the standard of truth for their assertions; and I think Brandom is right that the mechanism of standard selection is a relation internal to language, not between language and the world, although since language is lumpy, the line is somewhat blurred anyway.

Two years later – exactly to a day, as it happens – I wrote a follow-up comment:

Reinikainen, J. (2025). Questions of reference. In A. Laitinen, M. Keinänen, J. Reinikainen & A. Honkasalo (eds.). *Language, Truth, and Reality*. Tampere University Press. (pp. 127–141).
https://doi.org/10.61201/h5fbdk64

Think again.

It suffices to say that during those two years, my answers had shifted, and with them, the questions. While the direction of influence is clear enough, I, for one, could only see it after the fact. The process was above all guided by the arguments I came to work through by myself after the prolonged, tacit, passive exposure to Professor Raatikainen's work on reference.

This paper discusses the topic which I consider to be of fundamental importance to language – theory of reference – from Kripke's causal-historical viewpoint. In particular, I extend the defence of this viewpoint made in Raatikainen (2020) against recent criticism, owing to Kipper and Soysal (2022), that favours the descriptivist alternative. I proceed by characterising Kipper and Soysal's arguments and showing how they can be responded to mostly based on arguments already made in Raatikainen (2020).

In the first section I focus on one key deficit of descriptivism, namely that its explanations of reference are essentially incomplete. After that, I will focus on another key deficit, according to which the explanations are redundant.

Against descriptivism I: Incompleteness

The first problem with Kipper and Soysal (2022) is how they define 'descriptivism':

> **Descriptivism** For any speaker, S, expression, e, and class, C, if S refers to C with e, then there is a property, F, such that (i) S intends to refer to all and only Fs with e, and (ii) C is the class of all and only Fs; and S refers to C with e because F satisfies (i) and (ii). (Kipper and Soysal 2022, 655)

It is clear that 'descriptivism' is here defined as a theory concerning only reference, whereas, as Raatikainen (2020, 70–71) reminds us, the descriptivism that Kripke targets in *Naming and Necessity* (1980, henceforth abbreviated as NN) is first and foremost defined as a theory of meaning by both its supporters and Kripke himself. The reason this is a problem is that Kipper and Soysal explicitly frame their defence of descriptivism as Kripkean in nature, meaning that their claim is that Kripke himself (along with some other notable externalists) is committed to descriptivism as they define it, based on what he says in (NN). This latter part is partially true: Kripke indeed explicitly accepts many of the claims made by Kipper and Soysal, as we shall shortly see. But it should be stressed from the start that, unlike Kipper and Soysal at times imply, these other claims do not fall under the descriptivism that Kripke criticises in (NN).

While this problem is not strictly speaking a fault in argumentation, it carries a risk of confusing the debate. One can see this in how, at the end of their paper, Kipper and Soysal (2022, 664–665) claim that their definition of 'descriptivism' is immune to Kripke's semantic arguments – namely, arguments from ignorance and error – against descriptivism. This is confusing because their definition of descriptivism (i.e., a theory of reference) is not the definition of descriptivism used by Kripke (i.e., a theory of meaning) when he formulated the arguments from ignorance and error. It is hardly surprising, then, to find the arguments weak when they are misapplied in this fashion. That being said, even when descriptivism is understood as a theory of reference only, the arguments from ignorance and error have some traction, as I will explain later.

Moving on, let's review Kipper and Soysal's arguments for descriptivism, understood henceforth as a theory of reference only. They begin by observing that Kripke in (NN) allowed that sometimes the reference of a proper name can be fixed with the help of a definite description: the famous cases he mentions are the namings of Neptune and Jack the Ripper. Based on these examples, Kipper and Soysal suggest that Kripke is committed to the following principle:

> **SI** For any speaker, S, and expression, e, if there is a property, F, such that S intends to refer to all and only Fs with e, then S refers with e to the class, C, of all and only Fs, and S refers to C with e because (i) S intends to refer to all and only Fs with e and (ii) C is the class of all and only Fs. (Kipper and Soysal 2022, 656)

First, a word on exegesis. One characteristic, well-known trait of Kripke is his extreme caution in committing himself to any philosophical theory, or to develop one himself. As explained by Raatikainen (2020, 76), at many places in NN Kripke explicitly declines to give a theory of either meaning or reference to replace the one he's rejecting – at most he gestures towards 'a better picture'. With that in mind, I would hesitate to pin a general principle any such as SI on Kripke, who always preferred to keep the argumentation on the level of concrete examples and cases.

Furthermore, even supposing that Kripke would find SI acceptable, it is another question whether it is true. One reason to think it is not is explored in a recent paper by Jani Sinokki (2021). His discussion points to good reasons to caution against drawing a generalised principle like SI from the concrete examples discussed by Kripke, and is worth quoting at length:

> It seems that most familiar instances of "refence fixing" [*sic*] by description turn on closer inspection to presuppose ordinary causal-informational connections. Consider the famous case of fixing the reference of "Neptune." Alexis Bouvard first noticed certain irregularities in the orbit of Uranus and suggested that their cause of is another planet (as opposed to the Le Verrier, who only later calculated the location of the suggested planet). If Bouvard used the name "Neptune" for that planet, then his coining this name for the planet took place by simple causal-informational connection. There was the data (visible light) that was information about an event (unexpected

> irregularities in Uranus' orbit), which originated from something. With the realization that the cause of the perturbations is singular thing, a planet, Bouvard used this causal-informational connection to tag this object (not yet directly observed) with the name "Neptune." Importantly, even before coining the name, there was a question to be asked by using the data Bouvard was trying to interpret: "What is that?" or "What is the cause of that?" The truth-values of the answers to these questions are determined already by the actual history of the causal information coming from an object about which the question is asked, so also the reference must precede the use of any description used to *introduce* the name. (Sinokki 2021, 342; footnote omitted)

Another famous example by Kripke that Kipper and Soysal (2022, 656) appeal to as defence of SI, is the coining of the word 'Gödel' as a descriptive name for whomever invented the incompleteness theorems, which Kripke makes in (NN, 91). Briefly, in such a case, the speaker uses a definite description to fix the reference of a name, which Kipper and Soysal take to mean that sometimes the intention to refer alone is sufficient to refer.[1] However, here it is left open how the reference of the words in the associated description, in this example the reference of 'the incompleteness theorems', is established. If the reference is causal-historical, then this case also is not a 'pure' example of fixing the reference by description, but a hybrid case. This would be in conflict with SI, assuming that a descriptive intention alone is sufficient to fix reference for *every* expression. Suppose, then, that the reference of 'the incompleteness theorems of arithmetic' is not causal-historical but can itself be given a descriptivist explanation in line with SI. This effectively means that the introducer of 'Gödel' is able to describe the incompleteness theorems without reference to anyone else's work. But doesn't that obviously mean that the speaker herself has independently invented the incompleteness theorems? If so, it seems that not many people can coin the name 'Gödel' in the *purely* descriptivist fashion, i.e. in the fashion where the terms in the introducing description also are given a descriptivist explanation.[2]

The previous discussion offers reasons to caution us from drawing a generalised principle like SI from the concrete examples discussed by Kripke.[3] Suppose, though,

[1] It should be emphasised that Kripke himself would not call the Gödel example as the introduction of a *proper name* solely with the means of a description, as he explains elsewhere (Kripke 1977, 260, fn.9)

[2] It could be objected that this problem is merely an artefact of the particular example, namely the incompleteness theorems and their inherent difficulty. The objection would be premature, as the following discussion will reveal. The main point is that in order to fix reference purely descriptively in the spirit of SI, the reference of the terms in the associated description must also be fixed only descriptively, which introduces the essential problems regardless of what these descriptions are.

[3] And by that token cautions us also from attributing the commitment to SI to him. Indeed, my elaboration of the Gödel example is more or less directly analogical to what Kripke says about fixing the reference of 'Einstein' by the description 'the author of the theory of relativity' (NN, 82). So, it's not unreasonable to say that Kripke was well-aware of the problems of a generalised principle like SI.

that SI, or perhaps some appropriately adjusted replacement, is true. Granted that, there is another important weakness inherent in the principle which Kipper and Soysal do not seem to notice. The weakness is that SI does not stand for an autonomous mechanism of reference fixing: it is parasitic on some other, merely presupposed mechanism of reference. So, even if SI is true, the explanation of reference it provides is essentially incomplete and dependent on some other theory of reference. Let me explain this thought in detail.

The thought is due to Michael Devitt's observation that description theories of reference are essentially incomplete (Devitt 1996, 159; see also Devitt and Sterelny 1999). We can see the incompleteness problem by asking how the speaker S is related to the property F in SI. Presumably, F must be associated with some kind of mental representation in S's mind; this is also suggested by Frank Jackson's (1998) discussion, to whom Kipper and Soysal refer. But now it follows that, in order for S to refer to C with e via F, she must already have a representational, referential relation in her mind to F. What is the medium of this association, or the nature of the mental reference? This is not what SI, or the description theory as such, is able to explain; it merely presupposes that some such relation exists.

Now, Jackson for one believes that this outcome is not a major problem for the description theory of reference:

> There is of course an important problem of reference for the words of mentalese (if such there be), and more generally for how we refer in thought, but, as signalled earlier, this is not the problem of reference that the Lockean description theory we are defending is concerned with. (Jackson 1998, 204)

In contrast, I agree with Devitt that incompleteness poses a major problem for the description theory of reference. The reason is that the incompleteness problem makes the explanations provided by the description theory parasitic on whatever explains reference at the mental level – it merely passes the buck, as Devitt puts it.

It is vital not to misunderstand the incompleteness problem as the unreasonable demand that, in order to explain anything, a theory should explain everything. If that was the case, only theories in fundamental physics would be genuinely explanatory – quite possibly not even those. Rather, the main point of the incompleteness problem is that the description theory, as Devitt says, explains the reference of some term *by appealing to the reference of other terms,* i.e. representings in Mentalese or in some other mental medium (basic representings in Mentalese are often called 'terms' or 'words' for convenience). The theory presumes what it seeks to explain. This could be a perfectly valid explanation, perhaps an interesting one to some extent, but the answer it gives is essentially incomplete as concerns the *nature* of reference. That is to say, descriptivism does not do what we most of all want a theory of reference to do, namely explain reference without appealing to reference; explain reference in an ultimate sense, as it were.

How does the causal-historical account fare better in this regard? Simply put, the causal-historical account states that the buck stops at what Devitt (1996, 167) calls 'grounding uses', which paradigmatically involves a perceptual contact between the speaker and a token referent. Reference can then be borrowed to those speakers who haven't had perceptual contact with the referent, forming a chain-like network of causal-historical relations. This is, of course, only to provide the rough gist of the account, but the important point is that this account can explain reference in an ultimate sense, i.e. without appealing to reference established at the level of some mental medium.

Kipper and Soysal argue that denying SI leads to highly implausible conclusions. Consider again the example where someone coins the word 'Gödel' as a descriptive name for the prover of the incompleteness theorems. Such a speaker, called David, is disposed to conform all the credences of his beliefs involving 'Gödel' (only) to whatever information he gleans about the prover of incompleteness theorems. They then claim that:

> Someone who denies SI holds that it is possible that the reference of David's use of 'Gödel' isn't explained by his intentions, such that he might refer to something other than {x | x is the prover of the incompleteness theorems}. (Kipper and Soysal 2022, 661)

As we saw above, however, the key anti-descriptivist objection against SI is not that it doesn't work anywhere, but rather that it cannot work everywhere (i.e., it is incomplete). So, someone who rejects SI is not committed to denying that David couldn't successfully coin a descriptive name as he does, i.e. intend to use the name as applying only to whoever proved the incompleteness theorems. Indeed, since by hypothesis David is rational, and because his intention about the reference of 'Gödel' makes it an analytic truth in his idiolect that 'Gödel', if it refers, refers to whoever proved the incompleteness theorems, it is trivial that there exists 'no possible piece of information David could get that would break this connection between his 'Gödel'-utterances, credences, and the property of being the prover of the incompleteness theorems' (Kipper and Soysal 2022, 661). This is an irrelevance, however, for the interest in the example should be on the question of how the reference of the expression 'the prover of the incompleteness theorems' is to be determined. As we saw above, there are two options. If the reference is fixed purely descriptively, that can only mean that David himself counts as a prover of the incompleteness theorems as much as anyone else because he can correctly formulate the theorems without reference to anyone else's work. (Again, this is surely not how most people refer to the incompleteness theorems!) However, if the reference is fixed causal-historically, then this is a hybrid case of explaining reference, and thus not a counter-example to externalism.

Kipper and Soysal further claim that anyone who denies SI must claim that 'we have no control over the meanings of our words' (2022, 661). (I ignore the point that here the authors speak explicitly of meaning, although at the beginning of the paper they define Descriptivism as a theory of reference.) This seems quite hyperbolic and is surely something which a reasonable form of externalism should dodge. Indeed, there is a simple way in which to dodge the accusation. The key point is to agree that we do indeed have control over the reference (and meaning) of our words, but that the control is never *total* in the sense that one could, by oneself, determine the meanings and references of *every* word one ever used, as opposed to any single word considered in isolation. This is compatible with saying that David can use his control to coin the name 'Gödel' as he does by relying on the control others have wielded in determining the meanings and references of other words, for example, by borrowing the reference of 'the prover of the incompleteness theorems'.

In any case, externalism is not so much a claim about *control* as it is about *knowledge* of meaning and reference. Since reference is fixed by use[4], and since we are firmly in control of how we use language, externalism allows that we have indirect control over the reference of our words.[5] What we often lack, especially in the case of proper names and natural kind terms, is complete knowledge about the referents and the nature of the referential relation itself.

Against descriptivism II: Redundancy

The previous section argued that there are problems in the claim, encapsulated by SI, that descriptive intentions *alone* may suffice to fix the reference of *every* term (as opposed to any term considered in isolation). I first pointed out, following Sinokki (2021), that we should be cautious about drawing a generalised principle from the concrete examples discussed by Kripke because, arguably, these cases (e.g. the naming of Neptune and Jack the Ripper) are not purely descriptive, but rather hybrid cases, of reference fixing. That is to say, the success of reference here includes both a descriptive-intentional element and a causal-informational element. Second, following Devitt, I pointed out that even if a rare case of purely descriptive explanation of reference can be found in real life, the explanation provided by SI is essentially incomplete, as it merely passes the buck to whatever explains the speaker's reference to the property F in the mental medium.

In this section I tackle the complementary idea by Kipper and Soysal, according to which descriptive intentions are not only a sufficient but also a necessary component in explaining reference. Their initial formulation of the idea goes as follows:

> **NI** For any speaker, S, expression, e, and class, C, if there is a relation, E, between S, e, and C's members, because of which S refers to C with e, then

[4] This is implicit in Kripke's notion of 'initial baptism'. The idea is developed further by Devitt (1981; 1996).

[5] As a perceptive reviewer points out, much more would need to be said here about the nature of semantic control in an externalist setting. Here I can only refer the reader to the discussion in Steffen Koch (2021) as an example of how to understand the 'indirect control' that externalism allows for the reference of our words.

> S intends to refer to all and only the things that are E-related to e and S. (Kipper and Soysal 2022, 657)

In other words, the idea is that no purely causal-historical relation can be sufficient to fix reference, because wherever such a relation explains reference, a descriptive intention to refer via the causal-historical relation is needed. But, because such an intention is, per SI, always sufficient to fix reference, it follows that descriptive intentions alone are really both necessary and sufficient to fix the reference of any term – therefore, Descriptivism is vindicated.

The key deficit in SI, we recall, was that the explanation it offers is essentially incomplete. The key deficit in NI, on the other hand, is that the explanation it offers is essentially redundant. Supposing that there is a relation, E, *because of which* S refers to C with e, why is it additionally necessary for S to intend to refer to C with e? The main reasoning provided by Kipper and Soysal is that, apparently, Kripke and Putnam thought that intentions play an important role in explaining reference (Kipper and Soysal 2022, 657). While this is true, I think that Kipper and Soysal make a mistake in *how* intentions play their important role – for Kripke, in particular.

There are two key aspects in Kipper and Soysal's understanding of referential intentions, concerning the acts of borrowing and fixing reference. First is that the intention to refer must, for them, include some *descriptive content*: it must be an intention to refer to C via a certain relation, E. Second is that the intention must be *present* in every act of referring with a word. I argue that both claims are false, and that attributing them to Kripke, at least, is wrong.

Let's start with the second claim, that the intention to refer must always be present in the act of using a word referentially for the reference to succeed. Now, Kripke said:

> When the name is 'passed from link to link', the receiver of the name must, I think, intend *when he learns it* to use it with the same reference as the man from whom he heard it. If I hear the name 'Napoleon' and decide it would be a nice name for my pet aardvark, I do not satisfy this condition. (Kripke 1980, 96; my italics)

Kipper and Soysal (2022, 659) note that according to Devitt (2006, fn.6) and Raatikainen 2020), a natural reading of Kripke's thought here is that the intention to refer in accordance with original use need only be present when the name is originally borrowed, but that it is not needed later while using the name to refer in accordance with original use. The idea is consistently implied elsewhere in (NN) as well:

> I may then say, 'Look, by "Gödel" I shall mean the man Joe thinks proved the incompleteness of arithmetic'. Joe may then pass the thing over to Harry. One has to be very careful that this doesn't come round in a circle. Is one

> really sure that this won't happen? [...] You may not even remember whom you heard of Gödel. (NN, 90)
>
> A speaker who is on the far end of this chain, who has heard about, say Richard Feynman, in the market place or elsewhere, may be referring to Richard Feynman even though he can't remember from whom he first heard of Feynman or from whom he ever heard of Feynman. He knows that Feynman is a famous physicist. A certain passage of communication reaching ultimately to the man himself does reach the speaker. He then is referring to Feynman even though he can't identify him uniquely. (NN, 91)
>
> On our view, it is not how the speaker thinks he got the reference, but the actual chain of communication, which is relevant. (NN, 93)

If the speaker need not presently remember where she borrowed a name in order to refer with it in accordance with original use, then surely no present intention to use it in accordance with original use is needed, either.

Kipper and Soysal disagree, since the truth of NI requires that the descriptive intention be present in every act of using the name referentially:

> If Devitt's suggestion is correct, then NI is false, and reference can often be explained by external relations. However, Devitt's suggestion isn't credible. Assume, for instance, that the patient in Burge's thought experiment had the intention to defer to the experts' usage of 'arthritis' when he first heard the term, but later loses this intention and decides, instead, to use it to refer to tharthritis, which includes inflammations of muscles. This seems clearly possible. Just as a speaker can stipulate a term to have a certain reference when they first hear it, they can perform such a stipulation later. But Devitt's suggestion seems to imply that this is impossible. Accordingly, if a speaker once had the intention to defer to others' usage, they won't be able to use this term with a different reference later, even if they want to. Such a view would entail that we only have control over the meanings of our words when we first encounter these words, which seems no less absurd than the view that we have no control at all over those meanings. (Kipper and Soysal 2022, 659)

The main objection to Devitt's (and by the same token, Raatikainen's) reading of Kripke, according to which the intention to defer with a term need not be present in every act of referring with it in accordance with original use, is that this would make later stipulations of novel reference impossible for the term. This, however, seems clearly possible, as showcased by Kripke's example of using the name 'Napoleon' as a name for a pet aardvark. The objection fails, though, because Devitt's (and Raatikainen's) reading of Kripke is compatible with the possibility that the speaker, after borrowing the name 'Napoleon', later decides to call her pet aardvark that. In such a case, it can be argued, the speaker has essentially generated a new name type that is homonymous with an existing name, which is a common enough occurrence. But, so long as the speaker does not form a referential intention that contradicts the one she had when she borrowed the name, it's plausible that reference according to original use can be explained without the intention being present during every instance of use.

It is curious that, after presenting their objection to Devitt and Raatikainen, Kipper and Soysal mention in a footnote that their objection fails against this more plausible construal of Devitt's and Raatikainen's interpretation (2022, 659, fn.17). As a consequence, they shift to defending a weaker version of NI:

> **NI*** For any speaker, S, expression, e, and class, C, if there is a relation, E, between S, e, and C's members, because of which S refers to C with e, then there is no property, G ≠ being E-related to e and S, such that S exclusively intends to refer to all and only Gs with e. (Kipper and Soysal 2022, 659)

With this adjustment, Kipper and Soysal's strategy is to concede that externalism is right that reference can be explained by external relations by default, although explicit speaker intentions can override these, as in the aardvark case. I shall first explain their argument before criticising it.

Kipper and Soysal argue for NI* based on two thought experiments, which seek to show that it is implausible to hold that a speaker's expression refers to something if the speaker has no dispositions to conform their credences based on the information they get about the referent (which is how they understand referential intentions). The first thought experiment involves one Ella, of whom two things are stipulated: she is rational and that she refuses to update any of her belief-credences involving 'Gödel' based on information she gets about the person at the causal-historical origin of her term 'Gödel'. Kipper and Soysal then claim:

> Assuming that Ella's second-order dispositions align with her dispositions to update her beliefs, Ella will also be disposed to say things such as "The fact that the person at the causal origin of my use of 'Gödel' won the Einstein award is irrelevant to whether Gödel won this award." Ella thus isn't disposed to conform her 'Gödel'-thoughts to information about the person that is causally related to her in this way. Assuming that Ella is rational, there seems to be no reason to think that in this case, Ella's use of the term is explained or determined by the relevant causal relations. (Kipper and Soysal 2022, 662)

How is this supposed to be problematic for the causal-historical account, or externalism at large? The key premise in this argument appears to be that if Ella is rational, then her dispositions to update credences based on information about the assumed referent of her term 'Gödel' are infallible, or at least very strong, evidence about what the reference of 'Gödel' in her idiolect really is. But why should we think Ella's dispositions are infallible or strong evidence in this regard? It seems that a tacit assumption is at play here: if Ella is rational, she is supposed to *know* what her term 'Gödel' refers to, i.e. she is supposed to know if information about an object is relevant to whether she should update her credences concerning 'Gödel'. So, if she says that the person at the causal origin of her use of 'Gödel' isn't relevant to her credences, this is evidence that she does not refer to the person at the causal origin of her use of 'Gödel'. At least, if Kipper and Soysal do not make this assumption, it's unclear why they would think Ella's dispositions to update her credences are strong evidence of reference.

Since every relevant fact about Ella is stipulated, the weight of the argument rests on its assumptions. The problem is that Kipper and Soysal don't provide justifications for these assumptions, indeed do not even articulate them explicitly. Moreover, the externalist has arguments available for why the assumptions are wrong. In particular, the assumption that Ella, being rational, is supposed to know what her terms refer to is clearly an internalist notion, and an easy target to the classical arguments from ignorance and error. Very often, it strongly seems that people who we otherwise would count as rational do not know very much about the referents of their terms, for example proper names, or then all that they believe about them turns out to be false. Likewise, it seems plausible to say that speakers can refer with words even when they have little to no idea how the referential relation itself works – barring such more or less trivial statements as 'The name *N* refers, if it refers, to whoever is called that'. A separate account is then needed about the relation of 'calling' to make this version of descriptivism interesting, as Kripke already noted (NN, 70).

There is, of course, an obvious way in which the assumptions in Ella's case are justified. They are justified if Ella has stipulated that 'Gödel' in her use does not refer to whoever is at the causal origin of her use of the term. But then this is not a counterexample to externalism, which allows that we can stipulate the reference of any term in isolation (although *not* the reference of *all* terms). Again, this is surely not how most people refer to Gödel, however.

Is there any way in which an externalist might explain Ella's refusal to update her credences based on information about the causal-historical origin of her word 'Gödel'? Very likely, Ella is a stout descriptivist theorist and thinks that her best evidence simply speaks against the causal-historical account. Naturally, the externalist is not obliged to explain why some people continue to believe in descriptivism. Rather, it should be Ella's burden to explain to us why her evidence justifies the rejection of the causal-historical account. This, she fails to do.

The second thought experiment includes one Fritz and his pet aardvark called 'Napoleon':

> An opponent of NI might try to argue that Fritz could later lose the intention to comply with his original naming ceremony and still continue to refer to his aardvark by 'Napoleon', as long as he doesn't acquire any conflicting

> intentions. To illustrate why this view isn't tenable, let us assume that one day, Fritz learns that the aardvark he is currently taking for a walk isn't the one he once named 'Napoleon'. If Fritz then gives up his belief that he is currently walking Napoleon, this indicates that he still intends to comply with his original naming ceremony. If, on the other hand, he doesn't change his belief, this indicates that he treats information regarding the initial naming as irrelevant to the reference of his use of 'Napoleon'. In this case, the reference of 'Napoleon' in Fritz' idiolect has plausibly changed. (Kipper and Soysal 2022, 662)

Now, there's a simple way in which the externalist can accommodate the case of Fritz and his aardvark. Finding that there's been a change of animal, we have two basic options. In the first option, Fritz decides to stick with his original intention regarding the use of 'Napoleon'. In the second, he rather decides to name this other aardvark also as 'Napoleon'. But does this second option mean that the reference of 'Napoleon' has changed, or that a new, homonymous name type has been added to Fritz's vocabulary? One would think that, absent an intention to withdraw the name 'Napoleon' from the first aardvark, it still retains the name. In any case, all these options are perfectly viable for the causal-historical account, as I shall now explain.

The general problem with Kipper and Soysal's strategy for defending NI and NI* is that it rests on a false dilemma. As explained above, in the dialogue as they understand it, the concession of NI* to the externalist is that reference is by default explained by external relations, but the speaker's intentions can at will override these, with the argument following that this is in fact the case everywhere – speaker intentions are always necessary to explain reference. The false dilemma here is that reference is exclusively either explained by external relations or the speaker's intentions. But it was never part of Kripke's (or Burge's or Putnam's as far as I know) plan to deny that acts of reference (including reference borrowing) are intentional acts, or to claim that intentionality does not play an important role for reference. The key insight of externalism is that the referential intentions need not, at least everywhere, have descriptive content in order to explain reference, or then that this content can be very uninformative, even false. Moreover, even in cases where the speaker intention is said to 'override' the external relation, as in the aardvark case, this is only possible because one external relation (the chain of reference that links the name 'Napoleon' to Napoleon) is replaced by another external relation (a perceptual one to a certain pet aardvark). So, there is no escaping the external relation, nor the speaker's intention: both are needed to explain reference.

In contrast, the reason why Kipper and Soysal think that the examples of Ella and Fritz vindicate Descriptivism is that the referential intentions (i.e. dispositions to update credences) are assumed to always be able to override any external relation when it comes to explaining reference:

> Generally speaking, the problem is this: If a speaker has a credence in a sentence and gets a piece of information, she will either adjust her credence in response to this information or she won't—there seems to be no middle ground here. This implies that if a speaker doesn't conform her credences to a particular property, then she has dispositions that conflict with this property, since there are cases in which she treats some piece of information about the property as irrelevant to her beliefs. It would seem misleading to say that a subject is indifferent regarding the relevance of this property in such cases. *In any case, we contend that reference cannot be explained by external relations if speakers have dispositions that are contrary to the relevance of these relations for reference.* (Kipper and Soysal 2022, 662, my italics)

Since the way that the speaker is disposed to update her credences in response to new information *determines* what she refers to, it is impossible for the speaker to find evidence about the referents of her words that would contradict her credences. In the case where contradictions between the speaker's initial credences and new information emerge, she has two options available. If she refuses to update her credences in view of the new information, this shows that she did not refer to whatever the new information is about, hence she has no false beliefs about it. On the other hand, if she does update her credences in response to the new information, this shows that she now *chooses* to refer to whatever the information is about. So, the only way in which the speaker could have a false belief about the referent is if she chose to hold some mistaken credentials about it, which she presumably will not do, being by hypothesis rational. But isn't it more credible that actual speakers often have false beliefs about the referents of their words regardless of their choosing to have them? And if so, does that really entail that such speakers must always be irrational, as opposed to merely ignorant? The true depth of the arguments from ignorance and error, of course, is to question the role of knowledge (and other epistemic notions) in determining reference.

To end this section, a word on the arguments from ignorance and error. Kipper (2012) defends the idea, originally made by Jackson (1998), that the primary intensions which determine a speaker's reference could include descriptions about the causal chains of borrowing by which the name has arrived to the speaker:

> I think that a good case has been made that even where names are concerned or other uses of terms to which the arguments from Ignorance and Error can be applied, speakers do (implicitly) know something which can determine the term's reference: They know that a name 'N', if it refers, refers to the individual called 'N' by those from whom she acquired the name. (Kipper 2012, 92)

First of all, this response misapplies the arguments from ignorance and error by understanding them as objections to descriptivism as a theory of reference, whereas their proper target is descriptivism as a theory of meaning. But even so, it does not take much to see that the ignorance argument, at least, has weight against this kind of descriptivism as well, only in a different sense. The key point is that if the reference of a proper name in the speaker's use is fixed by the causal-historical chain of borrowing, it is redundant to require that the person also has to *know* that this is how the reference of her terms is fixed (Raatikainen 2020, 91).

One last point. How should the speaker's referential intentions be described, according to externalism, supposing that it is wrong to understand them (everywhere) in terms of descriptive content? Notably, Kripke (1980, 163) left this point open. In any case, the critical arguments against descriptivism are independent of what the right account will eventually turn out to be.

Conclusions

This paper objected to a recent defence of descriptivism, understood as a theory of reference, made by Kipper and Soysal (2022). The main arguments on either side aren't especially new. The main clash point concerns what role epistemic notions such as belief and knowledge are to play in explaining reference. My conclusion echoes what has become the conventional wisdom: externalism prevails.

References

Devitt, Michael (2006): 'Responses to the Rijeka papers,' *Croatian Journal of Philosophy*, 16, 97–112.

Devitt, Michael (1996): *Coming to Our Senses: A Naturalistic Program for Semantic Localism*. Cambridge MA: Cambridge University Press.

Devitt, Michael (1981): *Designation*. New York: Columbia University Press.

Devitt, Michael and Sterelny, Kim (1999): *Language and Reality: An Introduction to the Philosophy of Language* (2nd Ed.). Cambridge MA: MIT Press.

Jackson, Frank (1998): 'Reference and description revisited,' *Philosophical Perspectives*, 12, 201–218. URL = https://doi.org/10.1111/0029-4624.32.s12.9.

Kipper, Jens and Soysal, Zeynep (2022): 'A Kripkean argument for descriptivism,' *Noûs*. 56: 654–669. URL = https://doi.org/10.1111/nous.12378

Kipper, Jens (2012): *A Two-Dimensionalist Guide to Conceptual Analysis*. Berlin: Walter de Gruyter.

Koch, Steffen (2021): 'The externalist challenge to conceptual engineering,' *Synthese* (2021) 198:327–348.

Kripke, Saul (1980): *Naming and Necessity*. [NN]. Cambridge MA: Harvard University Press.

Kripke, Saul (1977): 'Speaker's Reference and Semantic Reference,' *Midwest Studies in Philosophy II*, 255-276.

Raatikainen, Panu (2020): 'Theories of reference: What was the question?' In A. Bianchi (Ed.), *Language and Reality From a Naturalistic Perspective: Themes from Michael Devitt* (69–103). Springer.

Sinokki, Jani (2021): 'What on Earth Is Smenkhkare? WH-Questions, Truth-Makers, and Causal-Informational Account of Reference,' *Theoria*, 88, 326–347.

12
Carnapian explication and normativity

Aleksi Honkasalo

Introduction

This paper was partly inspired by a recent discussion I had with Panu Raatikainen and Jaakko Reinikainen on the relationship between conceptual engineering and Carnapian explication. Panu wondered if conceptual engineering is just rebranded Carnapian explication. My initial sentiment was that Carnapian explication should be thought of as a type of conceptual engineering rather than a rebranding, and this paper seeks to defend this claim. However, while working out the details, it became clear that the relationship between conceptual engineering and Carnapian explication is more complicated than it initially appeared.

Panu's suspicion is well justified since the term conceptual engineering was coined by Richard Creath to refer to Carnapian explication in which existing concepts are improved to better serve the needs of scientists (Creath 1990). However, nowadays the term is used to refer to a broader range of practices aimed at improving concepts. Modern conceptual engineering is not only interested in how concepts could best promote the acquisition of scientific knowledge but also in how they can help to achieve various socially valuable goals, such as increasing inclusivity, justice and democracy. Since both approaches seek to improve concepts, Carnapian explication has been viewed as a special case of a broader practice of conceptual engineering (Cappelen 2018, 3–4).

Honkasalo, A. (2025). Carnapian explication and normativity. In A. Laitinen, M. Keinänen, J. Reinikainen &
A. Honkasalo (eds.). *Language, Truth, and Reality*. Tampere University Press. (pp. 142–151).
https://doi.org/10.61201/h5fbdk64

Taking Carnapian explication as a species of conceptual engineering has its issues. Conceptual engineering is almost universally accepted as a normative practice. While some, including Creath, have interpreted explication as a normative practice set on improving our existing concepts (Creath 1990; Justus 2012), Carnap also expressed anti-normativist sentiment towards language choice in a famous passage: "in logic [and language], there are no morals." (Carnap 2000 [1934], 52)[1] Carnap argued that instead of prescribing certain linguistic forms and proscribing others, we should be tolerant towards the adoption of various linguistic forms.

How should we understand both the normativity of explication and what is meant by no morals in logic and language? How can the process of improving concepts be normative and at the same time the choice between languages be a non-normative matter? In this paper, I seek to reconcile the apparent tension between normativity of explication and the anti-normativism of Carnap's principle of tolerance. First, I show that Carnapian explication can be understood as *instrumentally normative*. Adopting concepts can be thought of as means to achieve goals. Instrumental normativity is compatible with the principle of tolerance, since (A) there can be several conceptual means to promote scientific goals and (B) if language choice is goal-relative, one ought to prefer certain concepts over others only insofar as one is willing to pursue the goals these concepts promote.

While understanding explication as instrumentally normative might be enough to dispel any worry that there is a tension between explication and tolerance, for the conceptual engineer interested in furthering social justice, mere instrumental normativity does not seem to be sufficient. These engineers may want to go beyond the conditional claim that "if you want to promote inclusivity, you ought to use more inclusive concepts" and instead asks which concepts promote goals which are worth pursuing (see e.g., Haslanger 2000, 33). Likewise, it does not seem to be the case that conceptual engineers, (or anyone really) ought to be tolerant of linguistic forms that further morally questionable goals. Using misogynistic concepts may well further the goal of increasing gender-inequality, but this fact does not itself seem to warrant tolerance.

Does this mean that what distinguishes Carnapian explication from conceptual engineering is that the former is merely instrumentally normative while the latter is in some sense more robustly normative, or that explication is not a subtype of conceptual engineering but merely a precursor to it? This conclusion would be too hastily drawn. First, it should be noted that the goals of improving the clarity, and the acquisition of new scientific knowledge are arguably goals worth pursuing the same way promoting social justice is a goal worth pursuing. I shall argue that since many concepts can promote the acquisition of scientific knowledge, there are no absolutely right concepts which everyone must adopt and therefore there is still room for tolerance of different conventions.

[1] For Carnap, the choice of logic and the choice of language are inseparable as it is evident from the sentence following directly after the quoted passage: "Everyone is at liberty to build up his own logic, i.e. his own form of language, as he wishes." (Carnap 2000 [1934], 52).

While I believe that some textual evidence, particularly in "Empiricism, Semantics, and Ontology" (1988 [1950]) and *Logical Foundations of Probability* (1971 [1950]) suggests that Carnap did treat epistemological goals of science as valuable on their own, the main purpose of this paper is not exegetical. Nor is the goal of the paper to defend Carnap against criticism laid towards his conception of explication or its broader philosophical basis (Quine 1951; 1954; Strawson 1978 [1963]). Rather the goal is to show that both instrumental normativity and stronger normative notions are consistent with Carnap's principle of tolerance. As such this paper contributes to the understanding of the practice of explication and also provides insight into how to understand normativity in conceptual engineering in its modern sense.

Principle of tolerance

Throughout his career Carnap held the view that the everyday language was vague and ambiguous, which produced misunderstandings and confusions. Carnap argued that much of philosophical literature discuss pseudo-problems, which are the results of confusions created by the use of these vague concepts (1928). To overcome the pseudo-problems, the concepts with which we conduct our philosophical and scientific inquiries must be logically exact. In *Logische Aufbau der Welt*, he referred to this process as rational reconstruction, but later he came to call this process *explication*. In "The Two Concepts of Probability" (1945) and in the first chapter of his *Logical Foundations of Probability* (1971 [1950]) he gave the most detailed discussion of this process. In explication, a prescientific concept (explicandum) is analysed making notes about the possible vagueness and ambiguities as he argued was the case with probability. Then these concepts are given formal definitions (explicatum). A pretheoretic concept may encompass multiple distinct notions – as Carnap argued was the case with the notion of probability – then explicandum can be given multiple explicata. If the concept is vague, the explicator can define the unclear instances as either belonging to the explicatum or not.

Martin Gustafsson, among others, has argued that explication is tied to the Carnap's logical pluralism (Gustafsson 2014, 510–11). Since the explicandum is by its very nature inexact (otherwise there would be no need for explication in the first place) there is no exact way to determine whether the proposed explicatum is right or wrong (Carnap 1971 [1950], 4). Similar denial of the applicability of rightness and wrongness to the choice of language can be found in the in The Principle of Tolerance the most famous statement of which can be found in the *Logische Syntax*:

> *In logic, there are no morals.* Everyone is at liberty to build up his own logic, i.e. his own form of language, as he wishes. All that is required of him is that, if he wishes to discuss it, he must state his methods clearly, and give syntactical rules instead of philosophical arguments. (Carnap 2000 [1934], 52)

Carnap was unimpressed by intuitionist arguments against classical logic, but he saw the value of studying formal systems that could capture intuitionist ideas. Indeed, Language I in *Logische Syntax* was such systems. While it may be useful to study formal systems in which certain inferences, such as the law of excluded middle, are restricted, Carnap objected to extending these restrictions to alternative systems.

Similarly, correct language is not forced on us by the considerations of what are the true meanings of the expressions. Absolute faithfulness to the prescientific concepts would only reproduce the deficiencies of those concepts and thus, even in their most logically exact form the explicated concepts would not serve the needs of the scientists. Since one is free to use one's language and logic there is no absolutely right or wrong choice between possible explicata of the same explicandum (Carnap 1971 [1950], 4–6).

Normativity and the principle of tolerance

It should be pointed out that while the principle of tolerance denies that there are absolutely right or absolutely wrong linguistic forms, this does not lead to an *anything goes* sentiment in the choice of explicatum. In *Logical Foundations of Probability* Carnap lists four requirements for a successful explication:

1. Similarity: The explicatum must be similar to the explicandum.
2. Exactness: The logical connection between the explicatum and the scientific system must be clear.
3. Fruitfulness: Explicatum must allow formulation of many universal statements (empirical laws or logical theorems).
4. Simplicity (Carnap 1971 [1950], 5–8).

Are these requirements in conflict with the principle of tolerance? Should we not tolerate linguistic forms that fail to satisfy these requirements?[2] However, Carnap's goal of stating these requirements was not to set up rules for language choice but rather to make explicit the rules that are implicitly followed by philosophers, scientist, and mathematicians who seek to make concepts more explicit (Carnap 1971 [1950], 7). In his reply to Strawson, Carnap clarifies this instrumentalism towards concepts: "Language, whether natural or artificial, is an instrument that may be replaced or

[2] Out of the four, it is easiest to show that the requirement of similarity is not in conflict with the principle of tolerance. The requirement says that, while there can allows for significant deviation from the prescientific concept, there must be some similarity between the and the explicatum. Otherwise, the proposed explicatum is not an explicatum of the explicandum, but something completely different. A completely arbitrary definition regardless of how exacts, would not be explicatum of explicandum. For example, definition of "fish" as celestial bodies orbiting the Sun, would not be an explication of the prescientific concept of fish. (Carnap 1971 [1950], 5) However, this does not imply that one should not adopt the definition of "fish" as celestial bodies in one's language. It merely means that in such a case the concept would not be explicatum of the prescientific concept fish.

modified according to our needs, like any other instrument." (Carnap 1978 [1963], 938.) Furthermore, he concedes that less exact concepts can themselves be useful for many purposes, suggesting that he does not intend to categorically prohibit the use of prescientific concepts (ibid 938–939).

Instrumentalism towards explications leads to adopting comparative rather than absolute evaluation of possible explicata. The question whether an explicatum promotes the goal of the scientist, the philosopher or the mathematician is not a matter of yes or no, but of better and worse. This idea of evaluation of language being matter of degree also appears in a later discussion of the principle of tolerance in "Empiricism, Semantics, and Ontology". While the sentiment is here very close to the earlier discussion of the principle in *Logische Syntax,* Carnap now clearly leaves room for evaluating competing systems of explicated concepts he called linguistic frameworks. Linguistic choices are to be assessed on the basis how well they serve the goals for which the language, especially the language of science is to be constructed. (Carnap 1988, 221).[3]

As the evaluation of explicata is relative to the goals of scientists, explication can be described as instrumentally normative, that is, in terms of the relationship between means and ends. One should choose the concepts that promote the goals of their inquiry. Explication of the concept of fish as celestial bodies does not promote the goals of the inquiries into marine life, but of those explications that do, some are better than others. One should prefer more fruitful concepts, (permit the verification of universal statements such as "all fishes have gills") and the method of their verification and the logical connections to other concepts are exactly given. Finally, all else being equal one should choose the concept that is simpler. But, these oughts are binding only so far as one has the stated goal. Whether or not the concept *nut* should include peanuts depends on whether one has culinary or botanical communicative goals. Thus, instrumental normativity is not in conflict with the principle of tolerance.

Conceptual engineering in its modern sense could also be described in terms of instrumental normativity. For example, Sally Haslanger suggests that we should assess our concepts in terms of how effective they are for accomplishing our (legitimate) purposes (Haslanger 2000, 33). If the concept of woman is to promote the goals of critical theory it needs to be such that it helps to "identify and explain persistent inequalities between females and males", be "sensitive to both the similarities and differences among males and females", track how gender [..] are implicated in a broad range of social phenomena", and "take seriously the agency of women" (Haslanger 2000, 36).

It is, however, important to note that Haslanger does not speak of any old purposes we may have, but specifically the legitimate ones. This suggests that conceptual engineering requires a notion of normativity that is stronger than merely

[3] I shall leave the question open, whether this difference amounts to change of heart in Carnap from 1930s to 1950s.

instrumental. There may be concepts which promote illegitimate goals, such as perpetuating oppression or disseminating misinformation, but these are not what conceptual engineers are interested in. If conceptual engineering is taken to be a normative endeavour, there is no place for concepts that promote illegitimate goals. Instead focus of a normative endeavour must be on the conceptual means for the ends worth pursuing.

Prima facie, the decision to use emancipatory over oppressive concepts seems to have more normative weight than the decision to use scientific over the prescientific concept of fish. Perhaps normativity is what distinguishes Carnapian explication from conceptual engineering. In other words, it could be that one should be compelled to revise the prescientific concept only so far as one is willing to pursue scientific goals. Nevertheless, setting aside Carnap's views for a moment, I believe that the difference is a matter of degree rather than kind. The advancement of scientific knowledge may not be as important as promoting equality, but this does not make it an unworthy goal to pursue and certainly not an illegitimate goal. The conceptual means for scientific ends could even be in conflict with the conceptual means for emancipatory ends, but this does not show that one of the goals is not pro tanto worth pursuing. There is often conflict in the pursuit of valuable goals. Sometimes this means that we have to consider alternative means in order to achieve both goals and sometimes we have to forsake one goal to achieve another, but none of these make the forsaken goals unworthy to pursue.

To treat explication as robustly normative, Carnap needed only to accept that the epistemic goals of science are valuable in their own right. While this may not seem to be much of a concession for Carnap, who throughout his career sought to advance scientific knowledge, this stronger notion of normativity could perhaps compromise his tolerant attitude towards admitting linguistic forms stance. After all, if a concept does not promote our legitimate epistemic goals, do we not arrive at "a dogmatic prohibition" against the adoption of a linguistic form, the very thing Carnap warned against (Carnap 1988, 221)?

However, even if concept A is better for the advancement of scientific knowledge than concept B, this does not mean that the use of the latter should be absolutely prohibited, for such prohibitions themselves can turn against our scientific goals. In a perfect epistemic situation, we might prohibit the use of inferior concepts where there are better ones available. However, since we are not in a perfect situation, a second-order prohibition: "do not prohibit the adoption of any concepts" better promotes scientific goals. Since even a poor tool may still permit the achievement of a goal, the risk of forbidding the use of concepts that may eventually prove to be useful is not worth the possible advantages of making the conceptual choice easier by narrowing the field. Finally, a poor concept for one scientific inquiry may eventually prove useful for another.

There still remains one significant complication. It was noted that conceptual engineers may not be tolerant towards adoption of concepts that seek to promote illegitimate ends. Is this intolerance in conflict with Carnap's principle of tolerance?

If language is a tool that is "useful for a hundred different purposes" (Carnap 1978 [1963] 938) might some of those be considered immoral or otherwise illegitimate? Furthermore, is not the claim that a linguistic form promotes immoral ends precisely the kind of philosophical argument that Carnap argues should not be used to argue for or against adoption of a linguistic form (Carnap 2000 [1934], 52)? While the choosing concepts to promote inequality, totalitarian regime and other insidious goals, is certainly not in the spirit of Carnap's philosophy, it may very well be that Carnap failed to consider the implications of illegitimate ends to his linguistic instrumentalism and that such choices are to be tolerated by the letter of the principle. Nevertheless, it is important to stress that tolerance does not mean that a linguistic form is immune to critique. In his discussion on the principle of tolerance in "Empiricism Semantic and Ontology" Carnap argues that the ultimate acceptance and rejection of a linguistic form is to be decided by the testing it in practical use. If this test can involve taking the critical attitude towards the goals language choice seeks to promote, perhaps Carnapian explication is not so different from normative conceptual engineering.

Conclusion

I have argued that instrumental normativity plays a role in understanding both Carnapian explication and modern conceptual engineering. Both seek to find out what are the conceptual means to achieve various ends. The relevant ends for Carnapian explication relate to scientific knowledge, whereas conceptual engineering deals with a broad range of goals, including furthering justice, unmasking oppression, and defending democracy. Conceptual engineering requires stronger normativity than merely instrumental normativity and given that the pursuit of scientific truth was a goal Carnap had personally adopted, it is not a farfetched idea that explication is at least compatible with treating the goals of science as legitimate ones.

While this suggests that Carnapian explication is indeed a type of conceptual engineering, a strongly normative view of explication must also be compatible with Carnap's principle of tolerance. This turns out to be a slightly more complicated matter, but I have suggested a way to combine these ideas. I argued that there are higher-order reasons that speak against the adoption of prohibitions against poor concepts.

While there is very little textual evidence supporting that Carnap saw explication as normative beyond instrumental, it is at least consistent to maintain that the notions, such as truth and reality only make sense within a chosen linguistic framework, whilst maintaining that the goals for which that framework was constructed are to be goals worth pursuing. Provided of course that these views are independently consistent. Regardless of whether Carnap would accept it, I have suggested a way of combining these ideas in a way which may prove illuminating for the contemporary discussions with Carnapian themes, including conceptual engineering, and pluralism about logic (e.g. Steinberger 2017; Kissel and Shapiro 2017).[4]

[4] I wish to thank the anonymous referee for helpful comments and corrections.

References

Cappelen, Herman (2018): *Fixing Language: An Essay on Conceptual Engineering*. Oxford New York (N.Y.): Oxford University Press.

Carnap, Rudolf (1928): *Scheinprobleme in Der Philosophie: Das Fremdpsychische Und Der Realismusstreit*. Berlin-Schlachtensee: Weltkreis-verlag.

Carnap, Rudolf (1945): 'The Two Concepts of Probability: The Problem of Probability', *Philosophy and Phenomenological Research* 5(4): 513–32. Published originally in 1950. URL = https://doi.org/10.2307/2102817

Carnap, Rudolf (1971): *Logical Foundations of Probability*. London: University of Chicago Press. Published originally in 1950.

Carnap, Rudolf (1978): 'P.F. Strawson on Linguistic Naturalism', in Paul A. Schlipp (ed.), *The Philosophy of Rudolf Carnap,* Library of Living Philosophers, XI, La Salle, Illinois: Open Court Publishing Co. Published originally in 1963.

Carnap, Rudolf (1988): 'Empiricism, Semantics, and Ontology', reprinted in *Meaning and Necessity: A Study in Semantics and Modal Logic*, 2nd ed. Chicago London: University of Chicago Press.

Carnap, Rudolf (2000): *The Logical Syntax of Language*. Philosophy of Mind and Language 4, London: Routledge. Published originally in German in 1934.

Creath, Richard (1990): 'Introduction', in Richard Creath (ed.), *Carnap, Dear Van: The Quine–Carnap Correspondence and Related Work*: *edited and with an introduction by Richard Creath*, Berkeley-Los Angeles-London: University of California Press. URL = https://doi.org/10.2307/jj.8501222

Gustafsson, Martin (2014): 'Quine's Conception of Explication – and Why It Isn't Carnap's', in *A Companion to W.V.O. Quine*, Wiley-Blackwell, 508–25. URL = https://doi.org/10.1002/9781118607992.ch24

Haslanger, Sally (2000): 'Gender and Race: (What) Are They? (What) Do We Want Them To Be?', *Noûs* 34(1): 31–55. URL = https://doi.org/10.1111/0029-4624.00201

Justus, James. (2012): 'Carnap on Concept Determination: Methodology for Philosophy of Science,' *European Journal for Philosophy of Science* 2(2): 161–79. URL = https://doi.org/10.1007/s13194-011-0027-5

Kissel, Teresa Kouri, and Stewart Shapiro (2020 [2017]): 'Logical Pluralism and Normativity', *Inquiry: An Interdisciplinary Journal of Philosophy*, 63(3-4), 389-410. URL = https://doi.org/10.1080/0020174x.2017.1357495

Quine, Willard Van Orman (1951): 'Two Dogmas of Empiricism,' *Philosophical Review* 60(1): 20–43. URL = https://doi.org/10.2307/2266637

Quine, Willard Van Orman (1954): 'Carnap and Logical Truth,' *Synthese* 12(4): 350–74. URL = https://doi.org/10.1007/bf00485423

Steinberger, Florian (2017): 'Frege and Carnap on the Normativity of Logic,' *Synthese* 194(1): 143–62. URL = https://doi.org/10.1007/s11229-015-0880-4

Strawson, Peter Frederick (1978): ‘Carnap’s Views on Constructed Systems versus Natural Languages in Analytic Philosophy,’ in Paul Schlipp (ed.), *The Philosophy of Rudolf Carnap*, The Library of Living Philosophers, XI, United States of America: Open Court. Published originally in 1963.

13
Theories of reference
What really is the question?

Jaakko Kuorikoski

In an essay in the honour of Michael Devitt, Panu Raatikainen (2020) takes up the challenge of answering to recent criticisms against the causal-historical or “new” theory of reference by advocates of the description theory. One of the key motivations for reviewing these various critiques, and the original claims by Kripke and others, is to see whether the parties in the debate even share a common understanding of what is the central question that a philosophical theory of reference is supposed to answer. Hence the title of his essay: “Theories of Reference: What is the Question?” In this essay in honour of Raatikainen, I intend to revisit that very question, as I believe more can be said about this very important, but often neglected meta-question. In doing so, I place this article within this particular causal historical chain of reference, and hopefully this will ensure that whatever it is that Devitt and Raatikainen were writing about, this essay will at least succeed in discussing the same topic: what should an adequate philosophical theory of reference be able to accomplish? As both Devitt and Raatikainen are avowed naturalists and take philosophical semantics to be a part of an encompassing empirical account of language, I will approach this question from the perspective of philosophy of science.

After briefly revisiting the history of the philosophical debate on reference from Mill to the emergence of the new theory of reference, Raatikainen answers his titular question in the following way:

Kuorikoski, J. (2025). Theories of reference: What really is the question? In A. Laitinen, M. Keinänen, J. Reinikainen & A. Honkasalo (eds.). *Language, Truth, and Reality*. Tampere University Press. (pp. 152–161).
https://doi.org/10.61201/h5fbdk64

> *Main question: In virtue of what does a referring expression refer to whatever it in fact refers to*? (Raatikainen 2020, 73)

As an answer to a question presented in the very title of an article, this might strike the reader as somewhat underwhelming, as one would think that there is not much disagreement or confusion about whether this really is the main question of philosophical theories of reference. Raatikainen quotes the key players, such as John Searle, Devitt, and William Lycan, who seem to be, more or less, in agreement that yes, this is the common task to be answered by philosophical inquiry into reference. The onus was originally on the reference of proper names, but as the hypothesis that the semantics of natural kind terms behave similarly to that of names gained traction, this seemingly semantic question began to acquire much bigger epistemological and metaphysical stakes.

Raatikainen first goes through the subtle shifts in the history of theory of reference, from discussion of meaning of proper names to reference as such, noting that none of the original theorists intended the theory to be a fully general theory of reference nor meaning, i.e., answer the main question for all possible expression types. He then argues that the modern versions of descriptivism, such as causal descriptivism and metalinguistic descriptivism, are not really up to the task of satisfactorily answering the main question. I will not review or assess Raatikainen's convincing rebuttals against the descriptivist proposals, and the reader is invited to look at the thorough and knowledgeable argumentation from the source. My intention is to take a step back and ask what kind of a question the main question is supposed to be in the first place, and whether the standard philosophical methodology of imaginary counterexamples is really fit for the task of answering it.

Before we start analysing the question in more detail, it is worth pointing out the broader philosophical stance shared by the causal-historicists and the descriptivists: that there is such a thing as the reference relation and that this relation has some important *explanatory* role in the big picture of understanding linguistic communication and perhaps even of our epistemic lot in the world. In contrast, different deflationist accounts of reference deny that there is a substantive relation between a word and its referent to which the concept of reference itself refers to, and that the meaning of 'reference' ought to instead be understood in some purely intra-linguistic way.

Now let us get back to the main question: In virtue of what does a referring expression refer to whatever it in fact refers to? There are at least two points in need of clarification here. What is the nature of the 'in virtue of' relation and what is the nature of the putative 'fact' of referring? What kind of an explanation is the theory of reference supposed to provide and what kind of a phenomenon is it that we are trying to explain?

Disregarding Kripke, at least both Searle and Devitt have stated that their accounts are to be a part of a fundamentally empirical understanding of the phenomenon of language. Especially Devitt has been very explicit about his stern commitment to (meta)philosophical naturalism. He is a self-described card-carrying naturalist and has published a number of important papers attacking the possibility of a priori knowledge (Devitt 2011). More specifically, he has defended at length a thoroughly naturalist methodology for semantics, which includes philosophical (fundamental) semantics as a key element (Devitt 1996). This stance is, I take it, also shared by Raatikainen. An important constraint in clarifying the above questions is therefore that the explanation and the phenomenon ought to be, if not identical with, then at least continuous with the kinds of explanations and phenomena investigated by empirical linguistics, psychology and the like.

"In virtue of"

Let us start with the first item of clarification: what kind of in-virtue-of-relation is at play here? All the key authors, including Raatikainen, insist that the theory of reference ought to be *explanatory* and that the relation thus carries explanatory weight. Again, Devitt is exceptionally clear in formulating the main question in explanatory terms: "The central question about reference is: In virtue of what does a term have its reference? Answering this requires a theory that explains the term's relation to its referent" (Devitt 1998). Furthermore, one of Raatikainen's key arguments against the adequacy of causal and metalinguistic descriptivisms is that they do not offer adequate explanations of reference.

This plea for explanations is not really surprising, as the explanatory commitment is the key feature distinguishing substantive from deflationary accounts of reference. For example, one of the main claims put forward by Brandom (1994), a deflationist about reference, is that representational vocabulary, including the concept of reference, is not itself explanatory, but instead an expressive and explicative metavocabulary. According to Brandom, stating that "'Moo Deng' refers to a baby pygmi hippo" is not to refer to any independently existing relation between the referent and the name explaining its meaning and use, but an act of simultaneously summarizing and instituting a set of inferential commitments and entitlements involving Moo Deng. However, this general claim is a core aspect of the whole Brandomian picture of language and Brandom does not provide any *specific* arguments against the possibility of an explanatory account of reference in particular. Next, I will consider what kind of an explanation the causal-historical theory aims to provide. Together the commitment to an explanatory account of reference and metaphilosophical naturalism mean that the main question ought to be analysable as a scientific explanation, broadly understood. At least to my knowledge no one has seriously asked this question using standard conceptual tools from the philosophy of explanation.

I will start with the assumption that the intended *explanandum* is the fact that a particular expression denotes an object in the world ('Moo Deng' refers to Moo

Deng), and the *explanans* the chain of causally linked utterances of the expression starting from the baptism event.

Even though the very name of the theory refers to causality, let us first quickly discard the possibility that the explanation offered by the causal historical theory could itself be causal in nature. The first objection to this idea is that causal explanations in general are not answers to questions of the type "in virtue of what?". The relata of causal explanations are typically events, whereas here the *explanans* is the whole of the causal historical chain and the *explanandum* the property of an expression (type). Another possibility is that the surface form of the explanation-seeking question is misleading, and that the idea is that the baptism event causally explains the reference at the time of the utterance via the causal chain. An immediate problem with this suggestion is that causal explanation is transitive only in the special case that all the implicit contrast classes in the sequence of explanations line up nicely. Even if we charitably thought that encountering an instance of an expression could in some exceptional circumstances act as a sensible causal explanation of another use of that expression, the idea of a chain of such explanations is implausible.

A more promising suggestion is that the causal historical chain is constitutive of the property of the expression denoting a specific object in the world. This interpretation is also strongly suggested by many philosophers who explicitly state that the causal-historical chain is *the mechanism* in virtue of which the expression has the property of denoting a specific object. For example, Kaplan distinguishes between the way in which an individual is represented from "the mechanism that determines what individual is represented [reference]." (Kaplan 2012, 167, quoted in Raatikainen 2020) and even the Stanford Encyclopedia entry on reference states that the third central question of the theory of reference is "What is the mechanism of reference? In other words, in virtue of what does a word (of the referring sort) attach to a particular object/individual?" (Michaelson 2024). Conceptualizing the causal historical chain as a mechanism also chimes well with the naturalist ambition, as discovering mechanisms is something that the empirical sciences are supposed to be all about.

The problem here is that the putative causal historical chains between baptism events and subsequent uses of an expression do not look or behave anything like other explanatory mechanisms in the sciences. First, the causal historical chain is curiously distributed and extended both in space and especially in time. Let us take a paradigm empirical constitutive explanation of a property or a disposition by its realizing mechanism (in a very broad sense), such as the explanation of the brittleness of an object by its chemical and structural make-up. Here the explanandum is physically and temporally co-extensional with the explanans. The chemical structure in the here and now explains the disposition in virtue of there being a synchronic ontic dependency between the structure and the disposition (e.g. Ylikoski 2013). To be fair, examples of mechanistic explanation in the social science can be more diffuse both in space and in time, as they might involve relational properties, long-term equilibria and the like (Kuorikoski 2009). A particular market mechanism can balance supply and demand of assets or goods with market participants often literally from all over the world and with transactions taking place in dramatically different timescales. Nevertheless, the way in which a property of a (type of) expression here and now would depend constitutively on things that took place hundreds or even thousands of years ago, often in far-away places, is another thing altogether. Typical mechanisms are also relatively stable configurations in which the organization of the parts has an important explanatory role with regard to the property of the whole (Machamer et al. 2000), whereas historical chains of utterances are presumably highly contingent and the pattern of "reference borrowing" does not seem to have any systematic explanatory role.

Perhaps the sensible stance is to take the mechanism-talk as purely metaphorical and admit that the explanations offered by theories of reference are more distinctly philosophical. Clearly a more natural way of understanding the in-virtue-of relation is in terms of *grounding* and theorists of grounding mostly agree that either grounding simply is a form of explanation (e.g., Fine 2012), or alternatively serves as the metaphysical determination relation grounding philosophical explanations (e.g., Schaffer 2016). Although grounding theorists routinely lump many cases of empirical explanations which I would rather call constitutive (e.g. that the bowl's brittleness is grounded on the ionic bonds of its constituent atoms, see Kuorikoski 2012) as cases of grounding, I will restrict my discussion to more conceptual or metaphysical dependencies, as the possibility of constitutively explaining reference in the empirical sense was already dismissed above. The property of referring would thus depend on the causal-historical chain in the same sense as moral and aesthetic facts (if there are any) may depend on non-normative facts about acts and objects of art respectively, truth putatively depends on truth-makers, and essential properties on essences. Such explanations are distinguished, among other things, by implying stronger modality than mere nomological necessity.

The problem for naturalists like Devitt and Raatikainen is that such philosophical grounding explanations are also distinguished by the fact that they have little or nothing to contribute to empirical theories. Whether or not normative properties are grounded in non-normative properties is, at least arguably, inconsequential to any empirical theory of human behaviour, as normative properties do not have any causal power over such matters. Whether or not the redness of a particular colour is grounded in its maroonness is, arguably, pretty much irrelevant for chemistry, optics or neuropsychology of colour perception. Although grounding claims may well have some other, perfectly legitimate, cognitive roles, this interpretation would thus put a serious dent in the naturalist hope that the theory of reference would ultimately serve an explanatory role in a comprehensive empirical theory of language. Furthermore, if the theory of reference were to be a part of such a theory, the *explanandum* itself ought to correspond to an empirically ascertainable phenomenon. Next, I will turn my attention to what kind of fact this might be.

"In fact refers to"

For the advocate of a substantive theory of reference, not only is the causal historical chain supposed to be explanatory of what expressions in fact refer to, this fact about reference itself is also taken to have explanatory value. An obvious instance of this is the idea that reference is part of the meaning of at least some expressions and meaning, whatever it may be, ought to be explanatory of human behaviour (cf. Raatikainen 2020, see also Devitt 1996, ch. 2). If the theory of reference is to be a part of the empirical theory of language, reference ought to be not just a fact, but an empirical fact capable of being investigated by empirical means. Moreover, if these facts were to have explanatory power for human behaviour, they ought to correspond to some robustly *causal* phenomenon (Devitt 2011, 429).

In some sense it seems almost absurd to even question whether matters of fact about reference exist. The point of the whole business of language is presumably to communicate claims about the world, so surely some expressions are really about the world. Only someone who has seriously messed up her worldview with philosophy could deny that there is no such thing as (successful) reference. But it is one thing to admit the reality of linguistic representation and another to claim that there is such a thing as the reference relation. For example, a deflationist like Brandom certainly does not deny that we, as discursive beings, routinely refer to objects with our words – only that the sentence "'Moo Deng' refers to Moo Deng" does not itself refer to a special explanatory relation between the name and its bearer.

It is also clear that referring is, at least in some sense, an empirical phenomenon because it, or at least something closely related to it, actually *has* been empirically studied. Much discussed survey studies in experimental philosophy have claimed to show that there is significant cross-cultural variation in semantic intuitions about reference. Machery et al. (2004) claim that their empirical survey shows that East-Asians have more descriptivist semantic intuitions whereas Westerners think more along the lines of the causal-historical theory. Machery et al. further hypothesize that this difference is linked to a broader cross-cultural cognitive difference between East-Asians and Westerners in that East-Asians' categorization judgments depend more on similarity judgements whereas Westerners focus more on causality. In principle, there should be nothing mysterious about this, as for example grammatical intuitions are known to vary across different linguistic groups.

Max Deutsch (2009) has criticized these studies for falsely portraying the theory of reference (and philosophy of language in general) as relying on "the method of cases" tested against semantic intuitions in the first place. According to Deutsch, the theory of reference "makes predictions" directly about terms and their referents, about *semantic* facts, not about the intuitions of laymen or philosophers. Intuitions about reference and reference are, according to Deutsch, different things. The semantic fact that 'Madagascar' now refers to an island in the Indian Ocean can easily be ascertained by simply opening the Atlas. There is no need to survey any intuitions. Deutsch therefore is clearly committed to the idea of the reference relation as a robust phenomenon existing independently of our intuitions about it. In contrast, Devitt agrees what the theory of reference does resort to semantic intuitions as primary evidence, but argues that the empirical evidence presented by Machery et al. is simply not strong enough. His reasons are that intuitions about hypothetical cases are not as strong evidence as those about humdrum examples, that the intuitions relevant for the case against descriptivism are really metaphysical, not referential in nature, and that philosophers' intuitions really are better evidence than those of the common folk. (Devitt 2011) Although I am not convinced by these counterarguments claiming that the surveyed intuitions are not of the right kind, I will not dwell more on the matter here, as my interest is on the use of semantic intuitions in general.

An important principle of empirical research is that phenomena ought to be multiply measurable by mutually independent means of determination. It is only by triangulating with different independent methods that we can ensure that any putative result is real and not an artifact of any particular method. (Kuorikoski and Marchionni 2016; Wimsatt 2007) The crucial question now is, what other empirical means we really have of deciding whether 'Madagascar' really refers to an island in the Indian Ocean or still to a part of the mainland of Africa, and would do so in a way which would be *independent* of our semantic intuitions about the matter? It is important to clarify here that by semantic intuitions I do not only refer to intuitions in the specific (and perhaps proper) sense of private, immediate, pre-theoretical judgements, but more broadly to also include considered and public interpretations about what people take other people to mean with their words. These interpretations also encompass such things as the Atlas with the depiction of Madagascar in it. Is there any other way of empirically investigating what expressions refer to other than surveying what we take, implicitly as well as explicitly, the said expressions to refer to? At least I have never seen a reference relation and do not know of any empirical methods of detecting or measuring one without first going through our considered judgements about what we think our words refer to. Historians of various ilk certainly produce genealogies of words and concepts, some of which can be philosophically highly enlightening, but such historical narratives are simply further interpretations of historical changes in interpretations of what words mean. One can admit the existence of semantic facts without being committed to the idea that there is a robust empirical phenomenon of the reference relation out there.

Let us finally return to the truism that we use language to communicate claims about the world and that the existence of reference is therefore undeniable. The idea of a substantive reference relation is not solely motivated by the desire to understand the nature of linguistic meaning, but to also understand how linguistic representations are linked to the world outside language. This is also an epistemological worry. An important motivation for believing in substantive reference relations is that these relations could act as semantic hooks anchoring our concepts and beliefs securely into the world. Without such anchors, our system of beliefs is surely doomed to the dreaded frictionless Davidsonian spinning in the void (cf. McDowell 1994, 11).

In philosophy of science, the presupposition that a semantic theory should carry such epistemological weight led to the use of evermore sophisticated theories of reference in the attempt to disarm the pessimistic meta-induction argument against scientific realism. As argued by Stephen Stich and Michael Bishop (1998), this train of thought leads inevitably to some rather bizarre conclusions. If ontological commitments of scientific theories depend on what its terms refer to, and this reference relation is a substantive phenomenon, then we do not really know the ontological commitments of any of our theories until we have found the true theory of reference. But this is plainly mad. Before we elevate the linguistics departments to the highest position in the hierarchy of sciences, we should perhaps rethink the very idea of referential hooks as necessary conditions for our epistemic access to the world. There is plenty of friction with the world even without such contraptions.

Theory of reference as a model of data

If we are serious about the naturalist conviction, what then, remains of the epistemic role of a philosophical theory of reference within an empirical account of language? I definitely do not want to claim that such theorizing is scientifically empty. I suggest that the theory is, in fact, a highly stylized *model of data* in a verbal form. A data model is a representation of data, which highlights some selected systematic features of the data in a cognitively salient manner. In the case of theories of reference, the primary data are the semantic intuitions, understood very liberally as above. The theory thus summarizes a systematic feature in our semantic intuitions: we tend to *judge* or *interpret* people as referring to things in accordance with the tradition of using the expression and with the assumption that at some points in the history of the use of the expression there has been direct interaction with whatever the expression is taken to mean. In fact, this is the very stance that Devitt takes to be the implicit interpretation of the role of semantics by many philosophers, an interpretation he finds deeply mistaken (Devitt 2011). It is important to note, however, that I do not intend to make any sweeping claims about semantics in general, only a suggestion concerning philosophical theories of reference in particular (as, for example, cognitive and computational semantics arguably deal with explanatory relations between language use and cognitive and computational phenomena).

As already Patrick Suppers pointed out in his “Models of Data” (1962), data models are of paramount importance to inquiry. Scientific theories can directly engage with neither phenomena *an sich* nor raw data. Theories explain and are tested by specific (claims about) phenomena, which have to be purposefully and painstakingly distilled from the cacophony of raw data. (Bogen and Woodward 1988) Theories of reference can thus be seen as crystallizations of stylized facts about (a certain aspect) of the phenomenon of language. But what is crucially important to note here is that models of data are not explanatory. They only highlight salient patterns indicative of phenomena, but do not contain epistemic resources to explain those patterns. The explanatory resources lie elsewhere, probably in psychology, socio-linguistics and related fields.

Viewing philosophical theories of reference as data models also partly salvages their standing in the face of the possibility of significant cross-cultural differences in semantic intuitions. Mallon et al. (2009) use the empirical result as grounds to dismiss all philosophical applications of theory of reference: if there is no one correct substantive theory of reference, there can be no arguments from reference. While I agree with the sentiment, it is also important to note that these cross-cultural differences are made much more salient by the vocabulary of theory of reference. Instead of finding ad-hoc arguments to downplay the significance of this data, the true naturalist would welcome the discovery of such interesting systematic differences as potentially important phenomena waiting for an explanation, just not by a philosophical theory of reference.

References

Deutsch, Max (2009): 'Experimental philosophy and the theory of reference', *Mind & Language*, 24(4), 445–466. URL = https://doi.org/10.1111/j.1468-0017.2009.01370.x.

Devitt, Michael (1996): *Coming to Our Senses*, Cambridge: Cambridge University Press.

Devitt, Michael (1998): 'Reference', in E. Craig (ed.), *Routledge Encyclopedia of Philosophy*. London: Routledge.

Devitt, Michael (2011): 'Experimental semantics', *Philosophy and Phenomenological Research*, 82(2), 418–435. URL = https://doi.org/10.1111/j.1933-1592.2010.00413.x.

Fine, Kit (2012): 'Guide to Ground', in Fabrice Correia & Benjamin Schnieder (eds.), *Metaphysical Grounding: Understanding the Structure of Reality*, Cambridge: Cambridge University Press: 37–80.

Kaplan, David (2012): 'An idea of Donnellan', in J. Almog & P. Leonardi (eds.), *Having in Mind: The Philosophy of Keith Donnellan*, New York: Oxford University Press, 122–175.

Kuorikoski, J. (2009). 'Two concepts of mechanism: Componential causal system and abstract form of interaction', *International Studies in the Philosophy of Science*, 23(2), 143–160.

Kuorikoski, J. (2012). 'Mechanisms, modularity and constitutive explanation', *Erkenntnis*, 77(3), 361–380.

Kuorikoski, Jaakko & Marchionni, Caterina (2016): 'Evidential diversity and the triangulation of phenomena', *Philosophy of Science*, 83(2), 227–247. URL = https://doi.org/10.1086/684960.

Machery Edouard, Mallon Ron, Nichols Shaun & Stich Stephen (2004): 'Cross-Cultural Style', *Cognition*, 92(3): B1-B12. URL = https://doi.org/10.1016/j.cognition.2003.10.003.

McDowell, John (1994): *Mind and world*. Cambridge, MA: Harvard University Press.

Michaelson, Eliot (2024): 'Reference', *The Stanford Encyclopedia of Philosophy* (Fall 2024 Edition), Edward N. Zalta & Uri Nodelman (eds.), URL = https://plato.stanford.edu/archives/fall2024/entries/reference/.

Raatikainen, Panu (2020): 'Theories of reference: what was the question?', in Bianchi, A. (ed.) *Language and reality from a naturalistic perspective: Themes from Michael Devitt*, Springer: 69-103.

Schaffer, Jonathan (2016): 'Grounding in the Image of Causation', *Philosophical Studies*, 173(1): 49–100. URL = https://doi.org/10.1007/s11098-014-0438-1.

Suppes, Patrick (1966): 'Models of data', *Studies in logic and the foundations of mathematics* 44: 252–261. URL = https://doi.org/10.1016/S0049-237X(09)70592-0.

Wimsatt, William (2007): *Re-engineering philosophy for limited beings: Piecewise approximations to reality*, Harvard University Press. URL = https://doi.org/10.2307/j.ctv1pncnrh.

Ylikoski, Petri (2013): 'Causal and constitutive explanation compared', *Erkenntnis* 78(2): 277–297. URL = https://www.jstor.org/stable/24010965.

14
How ideal was Ockham's universal mental language?

Mikko Yrjönsuuri

Linguistic universality was a rising trend from the 1960s onwards among contemporary philosophers of language and of mind. Noam Chomsky and Jerry Fodor were among the main champions in this trend as systematic philosophers. Historians of philosophy of the analytic bend quickly noticed that William Ockham's theory of mental language resembles in many ways what was then known as a fashionable trend. That is, there was a clear medieval predecessor in philosophy of language and of mind for what seemed to be a very promising research paradigm.

There is, of course, something very tempting in the hypothesis that all humans are somehow "hard wired" to learn to speak understandable languages. That is, understandable to other humans, and as far as we know, in some extent to many pets as well. Even historical documents, like stories of bishop Anskar travelling among the Vikings in the 850s, or captain Antão Gonçalves and his crew capturing black West African slaves in the 1440s appear to show that it does not take long before people apparently not sharing anything in terms of a language start to understand each other linguistically. Once you learn a language, it won't take long to learn other languages too. There must thus be something universal in learning a language.

I never contributed anything to the contemporary discussion, if we leave aside teaching at the university level. I did, however, join the work of making contemporary philosophers know about Ockham as an important medieval predecessor of the universal language hypothesis. One paper deserves special mention here. In the year 1999 Panu Raatikainen organized at the philosophy department of the University

Yrjönsuuri, M. (2025). How ideal was Ockham's universal mental language? In A. Laitinen, M. Keinänen, J. Reinikainen & A. Honkasalo (eds.). *Language, Truth, and Reality*. Tampere University Press. (pp. 162–172). https://doi.org/10.61201/h5fbdk64

of Helsinki a colloquium with the title *Universaalikieli* (in English, "Universal Language"), and I gave a talk there. Raatikainen's introduction in the resulting book discusses an impressive array of historical philosophers ranging from Raymond Lullus and Gottfried Wilhelm Leibniz to Gottlob Frege and Bertrand Russell, and further to Noam Chomsky and Jerry Fodor. As Raatikainen describes, they all put forward differing kinds of hypothesis about a universal language. My paper in that book discusses Ockham's theory of mental language as a universal language.

Disregarding the possibility that the topic might be outdated now, I wish to take it up again here. In a certain sense, Ockham clearly thought of his theory of mental language as a theory of a universal language. His claim was that all intellectual beings – humans and angels – think in this language. It is a universally shared language of thought. He quite obviously also thought that mental language is ideal in some sense. For example, he claims almost at the beginning of *Summa logicae* (Ockham 1967, 13; I, 3) that mental language does not have anything to correspond to the distinction of gender which Latin has, but which has no effect on truth values.

The substantial question I pose in this paper is how exactly we should think of Ockham's mental language as ideal. Being ideal may appear to come quite necessarily with so absolute universality as Ockham was positing for his mental language. However, being ideal is a different thing from being universal. For example, Fodor apparently thought that his mental language is universally shared by all humans, but it is based on the modularity in how the human brain works and as such based on the logically contingent developments of how the human brain became to be what it is. It is not based on transcendent logical necessity of the Kantian kind.

Here I discuss the way and extent in which Ockham's theory of universal mental language was a theory of an ideal language. I think this is not only fruitful to understand the possible universal characters of human languages, but also for seeing how the contemporary philosophical scene affects how we approach history of philosophy. The purely historical issue is how exactly we should understand what William Ockham was doing in his theorizing on mental language, and this issue has not been solved – and probably never will. Discussion on Ockham's logical and linguistic theory has of course moved on, and thus my look could be described as a retrospective account of one particular discussion in the history of philosophy. I have structured the discussion historiographically, to follow the development of the contemporary research on Ockham's theory over the last half century.

One interesting part of the story of Ockham interpretation is that it was not affected very much by the still ongoing work of producing critically edited texts from medieval manuscripts. For Ockham's *Summa logicae*, which is the crucial source here, was critically edited already in 1967. Desire for a complete and reliable translation was left waiting then, but in this case the mainlines of the development of the interpretation cannot be explained by new sources coming forward. The Latin of *Summa Logicae* was easy enough so that the edition was sufficient for what philosophers needed.

I start with presenting Ockham's theory in the crude form in which it was discussed after the critical edition of *Summa logicae* was published. Then I introduce the complications added to the interpretation as they became recognized. They effectively brought on a controversy on whether the crude form could be taken as a simplification or whether it misses the picture altogether. The development of interpretations of Ockham's theory of mental language is of course continuing still, but predicting future turns of the development is beyond my powers.

Soon after the critical edition of *Summa logicae* came out John Trentman published in *Mind* a seminal paper "Ockham on Mental" (Trentman 1970). In the ensuing discussion the cue for understanding Ockham's theory of mental language was taken from the way formality in modern logic was understood at the time. The idea is simple: there are categorematic terms which work as the material parts of sentences, and syncategorematic terms which express the formal structure. In relation to the question of ideality this approach results in claiming that Ockham maintained that humans think in a mental language that is roughly as ideal as standard predicate calculus with model theoretical semantics is. And at that time analytic philosophers took it to be quite ideal. Bertrand Russell's paper "On Denoting" published in *Mind* early in the century (1905) was much appreciated. The strategy it offered for philosophizing was to get beyond the ambiguities and other annoying features of spoken languages by finding the underlying logical structure that could be expressed by predicate calculus.

For the subpart of Ockham's mental language that is needed in basic Aristotelian syllogistic, the comparison between Ockham's mental language and predicate calculus of the 1970s works fairly well. Logical form seems to separate from the material parts of the language quite nicely. For example, the sentence 'every donkey is an animal' contains two categorematic terms, 'donkey' and 'animal', and two syncategorematic terms, 'every' and 'is'. Each of the four words have their correlates in the mental language. We can ignore the article 'an', since Latin does not use articles and there is no reason to suppose Ockham would have thought anything like it to be needed in mental language. It indeed exemplifies nicely how spoken language compromises ideality. Because of the syncategorematic terms – the universal quantifier 'every' and the copula 'is' – the sentence is evaluated from the formal viewpoint as a universal affirmative predication, and because of the categorematic terms, it is about donkeys and animals.

As a nominalist, Ockham thought that the world consists of individuals and nothing more. Furthermore, he unambiguously thought that even when it is true to attribute a relation, there is no third individual to connect the related things. Equally, quantities are nothing apart or in addition to the things that have the quantity. Also, states of affairs are not existing things. Only individual substances (for example, donkeys) and their individual qualities (for example, colors of the donkeys) exist. This kind of world appears very suitable to be described with a simple ideal language having categorematic terms to determine which individuals are spoken of and syncategorematic terms to specify what is said about them.

If given such a straightforward interpretation, Ockham's mental language will have a simple theory of truth. The function of categorematic terms is to refer to (in Ockham's terminology, to 'supposit for') individuals in the world, and the syncategorematic terms tell what exactly is claimed about those individuals. The semantics of such sentences could thus work with only one type of language–world relation: supposition. Alone, an absolute term signifies individuals, and in the sentential context this signification yields supposition for the terms. Thus, a simple sentence like 'a donkey is an animal' claims because of its syncategorematic terms sameness of the thing supposited by 'donkey' and the thing supposited by 'animal'. In other words, it will be true if and only if 'donkey' supposits for the same thing as 'animal'. This appears to be so if 'donkey' signifies donkeys and 'animal' signifies animals - and there exists at least one donkey available for supposition. With more syncategorematic terms, the logical structure becomes more complex, but the core idea remains derivable from sameness of supposition for affirmative sentences and difference of supposition for negative sentences.

For a more complex example, in the sentence 'Every donkey is an animal' the subject term 'donkey' has, in Ockham's Latin, *suppositio confusa et distributiva* ("confused and distributed supposition"), or supposits for all of the donkeys in a special conjunctive manner so that in order for the whole sentence to be true all the related singular predications must be true. That is, it must be so that it is true about each donkey to say that it is an animal, although all the donkeys together are not an individual animal. Also, the syncategorematic terms put the term 'animal' in the predicate position of a universal predication, and thus it supposits for the individuals that it signifies, but in a disjunctive manner that Ockham calls in Latin *confusa tantum* ("merely confused"). The sentence is true, because each thing supposited in this regularized manner by 'donkey' is identical with one or another of the things thus supposited by 'animal'.

The core hypothesis tested in the scholarly discussion was the assumption that Ockham thought of mental language as providing the logical form of any expression in an explicit manner. It seems clear that the theory works in the way described above for simple predications of standard Aristotelian syllogistic, but can it be expanded to cover everything that can be said in human languages? If Ockham was right that all spoken sentences have their mental correlates, one would thus need to ascertain that all mental sentences are in fact logical constructs of simple Aristotelian predications. But did Ockham think so?

In the early stages of the interpretative discussion after the critical edition of Ockham*'s Summa Logicae* was published it was thought that Ockham was indeed thinking that the mental correlates of spoken sentences do resemble formalizations. In the spirit of Russell's "On Denoting", Ockham's mental language was thus understood as an approach to analyze the perhaps misleading structures of spoken language. The ambiguities, synonymies, and opacities commonly found in spoken languages would be absent from the mental language. In this sense, mental language was looked at as being ideal in the same way as predicate calculus. Rendering a spoken sentence to mental language would make its logical form transparently visible. Also, having a clear and simple theory truth would make assignment of truth values very straightforward (if the world is known in the relevant respects).

Ockham claims mental language to be natural. It was clear already in 1970s that Ockham's understanding of 'natural' in this context was the opposite of how modern logicians speak. Predicate calculus is called artificial and English natural. That Ockham calls mental language natural means that it is shared naturally by all intellectual beings, while spoken languages are artificial, or constructed by human language users differently in different contexts. From this viewpoint, Noam Chomsky's program of generative grammar and especially Jerry Fodor's hypothesis of language of thought forcefully defended in the monograph *The Language of Thought* (1975) provided clear twentieth century analogues to what Ockham theory was taken to be. Chomsky and Fodor too claimed that the fundaments of language are beyond human control as innate structures of the mind (or brain).

Considerable amount of scholarly discussion went into figuring out whether Ockham thought syncategorematic terms to be innate (like Chomskyan grammar) or somehow learnt. This discussion was more or less abandoned, since it was found out when more texts were published in critical editions in addition to *Summa Logicae* that Ockham himself was wavering on the topic. This was thus a clear difference from the modern theories of Chomskyan vein, but not really a very significant one. The issue was simply that Ockham did not manage to complete his theory. To some of us philosophers it happens that we get involved in political interests or other endeavors of human life, and do not find the time to solve all issues opening in our theories. As is well known, Ockham got called to the Pope's curia in 1324, was living tumultuous years after that, and in the end had to escape in the darkness of night to avoid imprisonment. After having found safety at the emperor's court, he turned to political philosophy. He never returned to his proper studies and never received his doctorate.

A theoretical difference in another direction is related to Fodor's position that many or perhaps even all concepts are innate. As an extreme example, the concept 'carburetor' has got stuck in the discussions concerning Fodor's theory, since it is quite difficult to believe that evolution has produced an innate capacity to think of carburetors. Ockham for his part claimed that there are no innate categorematic terms. All concepts are acquired. He had a relatively clear theory of how we learn basic categorematic vocabulary of mental language. The most basic categorematic terms (which he called absolute terms) are learnt when encountering a significate or significates of the term. Encountering a lion, for example, any human or angelic mind will have an act of understanding or thinking about a lion, and this act yields the capability to repeat another similar act of understanding later. That is, after learning the concept we can think about lions whenever we like. Ockham spells out this capability as acquiring the mental word 'lion' to one's mental vocabulary by encountering a lion.

For Fodor, the problem was of course more complex because he could not rely on basic Aristotelian metaphysics of natural kinds, and in the way suggested by Ockham one can learn only the vocabulary for natural kinds. For Ockham, complications result from the fact that spoken human languages contain vast amounts of vocabulary that does not refer to entities that could be encountered in such a straightforward way as lions. Carburetor is one example of such a thing. It is not an Aristotelian natural kind. Not all carburetors look the same and it may be difficult to distinguish one by the looks of it. Here, though, Ockham has a clear solution. According to his account, artifacts are always signified by words that have a complex structure of meaning containing at least as a part signification of the function of the artifact. Such words are essentially complex. He calls them 'connotative terms' to distinguish them from what he calls 'absolute terms', which always signify in simple and equal manner whatever they signify. Connotative terms have nominal definitions which signify in an obviously complex manner everything that the term itself signifies in a possibly opaque manner. For example, 'saw' is nominally defined as 'metal thing with teeth used for cutting wood'. This definition makes it obvious that the term 'saw' signifies not only the metal teeth but also the human action of cutting wood. And primarily in its uses it of course means the whole tool as a tool.

Now, does mental language contain connotative terms that have such a complex structure in their signification? Interpreting Ockham's theory of mental language as a theory of an ideal universal language would suggest that there should be no connotative terms. Such terms are too messy for an ideal language. In an ideal language, each simple term signifies what it signifies in a simple and straightforward manner. In Ockham's mental language, that seems to be so for absolute terms signifying natural kinds, like 'lion'. But actual spoken languages contain vast vocabularies or terms of other types, in Ockham's terminology connotative terms.

The scholarly discussion soon noticed that Ockham makes the distinction between absolute and connotative terms in a chapter of his *Summa logicae* that belongs to the section describing the terms of mental language (Ockham 1967, 35–38; I, 10). This would not make sense if there were no connotative terms in mental language. So, the answer must be positive. There are mental connotative terms. 'Saw' may be a term in mental language and not only in English and some other spoken languages. For some time, the mainstream opinion appeared to be that connotative terms are like shorthand for their nominal definitions. That did not quite seem to work. So perhaps their analysis is in some other manner obvious? I myself suggested in a paper published 1997 that mental sentences with connotative terms need not be analyzed in a linear manner term by term, but that the sentences break into several sentences in the manner described by the medieval theory of *expositio*.

But let that be as it may. By thinking of Ockham's theory of language more from a semiotic angle, the Québécois philosopher Claude Panaccio, who had a bit more background in French philosophy, saw the theory of connotative terms differently. He claimed that mental language contains irreducible connotative terms which have one or more secondary significations despite being simple (cf. eg. Panaccio 1992, esp. 40–45). He encountered much skepticism among scholars who were approaching the problem with the hypothesis that Ockham aimed his theory of mental language as a theory of an ideal logical language comparable to the predicate calculus. For simple connotative terms of the kind Panaccio envisaged could not be acquired in the way Ockham tells us to acquire absolute terms. Also, the theory of truth as terms supporting for the same fails to give a complete account of truth in any proposition where there is a simple connotative term. And to put it simply, having simple terms that have complex signification just does not sound ideal.

One of the leading early discussants, Paul Spade, addresses the interpretative problem quite extensively in his article 'William of Ockham' in the *Stanford Encyclopedia of Philosophy* (2002, revised 2024). He describes the main insight that he defended against Panaccio as follows: "On this interpretation, if anything can be truly said about the world, it can be said *only* using absolute and syncategorematic terms in mental language." That is, anything that could be said about the world can be said without connotative terms. That would make simple connotative terms superfluous and indeed Panaccio's position superfluous. Connotative terms would simply not be needed. Whatever way they are explained away, the expressive power of mental language would remain the same. But this is not how Panaccio understands Ockham. He does not think in any manner differently about Ockham's nominalist metaphysics, but he thinks that connotative terms are needed because we can make importantly true claims about the world even when there is nothing in the world to make them true, and that there is no way to make exactly these claims without connotative terms.

Let us take the simplest possible example, the term 'white' (*albus*). For Ockham, whitenesses are existing individual qualities (comparable to tropes as spoken of in metaphysics nowadays) and 'whiteness' (*albedo*) is an absolute term. But 'white' is connotative. It primarily signifies the thing having whiteness, and secondarily the whiteness. The nominal definition of 'white' is according to Ockham 'a thing having whiteness'. Ockham uses often the example 'Socrates is white'. The sentence can be called true because the terms 'Socrates' and 'white' supposit for the same, namely Socrates. But the complexity arises because the connotative term 'white' signifies whiteness too and could not supposit for Socrates if he did not have any whiteness as a separately existing quality. In order to arrive to the sameness of supposition we must take into account also the quality of whiteness inhering in Socrates. Thus, two separate real individuals (Socrates as a substance, and whiteness as a quality) are involved as truth makers although only one of them is supposited for. That is, sameness of supposition does not give a full account of truth.

But how exactly is the whiteness involved? It is not enough for the truth of 'Socrates is white' that the individual substance Socrates exists and the individual quality of whiteness exists. The relevant individual whiteness must also inhere in Socrates rather than for example Plato. As Ockham spells out the requirement, it can be formulated as a three-part conjunction 'Socrates exists, whiteness exists, and whiteness inheres in Socrates'. According to Ockham, this conjunction is equivalent with 'Socrates is white'. With some assembly work the truth of the two first conjuncts can be spelled out as sameness of supposition ('Socrates exists' is equivalent to

'Socrates is a being'), but the third one cannot. Indeed, according to Ockham 'whiteness inheres in Socrates' has two terms which supposit for different things, and a verb joining them. Thus, sameness of supposition does not work as a criterion of truth for that conjunct, and of course is not then a satisfactory theory of truth for it.

Ockham seems to be at least almost ready to admit that 'inheres' is a syncategorematic term in the same vein as 'is'. This would solve the problem partly for this simple example, although I expect logicians' eyebrows rise at the suggestion. The claim could then be understood as containing only absolute and syncategorematic terms, as Spade's interpretative statement puts it. But truth conditions would still not be reduced to sameness of supposition. Furthermore, there are more complex examples. In his recent Oxford UP monograph *Ockham's Nominalism*, Panaccio uses the sentence 'Mary is a mother' as an example (Panaccio 2023, 122), and that surely picks out a relation that is not syncategorematic.

In Panaccio's account, Ockham sticks to sameness of supposition as the sole criterion of truth in the sense that truth of any sentence depends on sameness of supposition (or suitable logical derivative of the principle). 'Mary is a mother' is true if and only if 'Mary' and 'mother' supposit for the same thing in the sentence. The complexity is in his view at the significations of these terms. For this reason, Panaccio apparently thinks that Ockham was not presenting a theory of an ideal language. Indeed, Panaccio appears to take quite explicitly the stance that Ockham was not even trying to build a theory of an ideal language. Panaccio seems to think that nominalist metaphysics was primary for Ockham, and thus revising metaphysics to achieve a semantically ideal theory of language was not acceptable.

For Ockham as described by Panaccio, supposition builds upon a complex structure of signification of the connotative term. Thus, 'mother' signifies a female individual having at least one child. As one could say, it signifies primarily and thus typically supposits for the thing that has a child, but the supposition is really due to the secondary signification, which requires a child. Crucially, it requires a special relation between the female individual and the younger individual, and that relation cannot according to Ockham exist as a thing. In case of complex relations, the situation becomes complex because of a plurality of complex secondary significations, but Ockham's metaphysics cannot yield a status for everything signified secondarily.

Perhaps Ockham would have given up what has been called the 'truthmaker principle', if he knew about it (Cf. Panaccio 2023, esp. 49-51). In Ockham's metaphysics, there cannot be enough truthmakers to make 'Mary is a mother' or even the simple sentence 'Socrates is white' true. The truth of these sentences depends on at least two things. The relevant things are Mary and a child, or Socrates and a whiteness. Both, respectively, must exist for the sentence to be true. In that sense, we might call them partial truthmakers. But their existence is not enough, since the sentence really claims also that the child is Mary's or that whiteness inheres in Socrates. And motherhood and inherence are relations, and according to Ockham relations do not exist, they are not things. For Ockham, states of affairs were not existent things either, although soon after Ockham some philosophers changed course and suggested metaphysical existence for what they called *complexe significabile* (cf. eg. Gál 1977). Non-existents cannot be truthmakers, and thus there cannot be enough truthmakers to explain the truth of these sentences.

In the warm summer 2024 of the Italian city of Parma, at the XXIV European Symposium in Medieval Logic and Semantics with the title "Truth, Falsity, and Lying", I myself with Teemu Tauriainen and independently Milo Crimi turned out to have decided to talk about Ockham's criteria of truth for Latin sentences with terms in cases other than the nominative. For these sentences, Ockham is clear that sameness of supposition does not work as a criterion to truth. A prime example is the use of a genitive case used as expressing possession. The sentence 'this donkey is Plato's' requires, according to Ockham, that a donkey exists, that Plato exists, and that Plato owns the donkey. Ownership requires in Ockham's view that the owner and the object owned are not identical. As a good Franciscan, he rejected self-ownership. Thus, 'donkey' and 'Plato' must supposit for different things in this affirmative sentence for it to be true. However, this does not really give any criterion for truth, since Plato won't own everything other than himself, but only what he actually does own. Sameness of (or difference of) supposition is thus practically irrelevant to the truth of the sentence. Given that it is not at all difficult to find further examples in a wide variety of directions, we must judge that Ockham provides no clearcut criteria for truth for many or even most sentences, and Ockham appears to take that as no problem whatsoever.

It seems that I and Crimi would both agree with Panaccio's most important result in this respect and admit that Ockham thought that there aren't sufficient truthmakers for all true sentences. But more crucially for the purposes of this paper, it seems that the most recent work in Ockham's semantics has left behind the conception that Ockham's theory of mental language would be a theory of an ideal language. Ockham accepts serious complexities with open eyes without showing any sign of regret. Recent scholarship seems to have shown that quite clearly.

So, is Ockham's theory of mental language a theory of an ideal and universal language? It seems that there has not been serious questioning of the universality in the scholarship even though Ockham is clear about a relevant learning process. In his view, most intellectual beings only know a part of the vocabulary of the language, because they are familiar with only a part of the world. Thus, there may be intellectual beings who do not understand any given sentence of the language: Ockham even posits that as a limitation in angelic communication. In this sense, the language is not universal. But for all intellectual beings, the significations of the words that they do share are the same. And in this sense, the language is universal in his view.

At present, it seems that scholarly consensus appears to have come to accept that Ockham did not intend his theory of a mental language as a theory of an ideal language. There are indubitable aspects of ideality. As already mentioned, mental language does not contain certain aspects of spoken languages that are less than ideal. There is no gender in Ockham's mental language, for instance. As a general principle, he claims that there is nothing that has no effect on truth values. But we have encountered other dimensions in which ideality is lacking. At least a subpart of connotative terms appears to be irreducible in such a way that their signification and behavior in sentences is not ideal. This is related to Ockham's open-eyed admittance of swaths of mental language that have no adequate theory of truth.

Problems with the 1970s insight that Ockham's theory of mental language aims at ideality and not only at universality could have in itself been sufficient to shift the focus in Ockham research from logically formal aspects of mental language towards metaphysical issues and issues related to the concept of truth. Such a shift has clearly happened. But it might also be that this shift is a result of what has happened in the scene of contemporary analytic philosophy. As the situation is now, there are few defenders of the view that a simple, logically ideal language could have enough expressive power for everything true.

Ockham's actual position in our contemporary questions is hard to decide mainly because he did not pose the exact questions that we now do. There are of course genuine problems of understanding the texts as well. No philosopher has managed to write in an ideal and universal language, not even Ockham. I think that we can nevertheless conclude that Ockham either was not optimistic about a theory of ideal and universal language, or maybe he did not even try to achieve anything like that.

So, what was Ockham trying to do in constructing a theory of mental language? Why did he adopt such a theory as the starting point in his course book in logic, *Summa logicae*? What did he actually think about the relations between language and metaphysics? I think that such questions are at the core of what we do as historians of philosophy, but they are not possible to put in any manner that would avoid all anachronism. "Mental language" does not mean for us the same as "*oratio mentalis*" meant for Ockham. Even "logic" means different things for us than "*logica*" meant for Ockham.

We can and must translate old philosophical texts, and we can and must interpret them from our viewpoints. But we cannot expect the exact theories we have at our contemporary scene to have occurred centuries earlier. At most, there are interesting similarities, but similarity is always a matter of vantage point. Or in other words, similarity is just a similarity, and as such it does not exclude differences. Scholars did find interesting similarities in Ockham when Chomsky and Fodor had success in claiming universality in human and mental languages. But as the success of the latter waned, it was realized that Ockham's idea wasn't quite the same either. It may be so for all history of philosophy. Questions change. Thus, every generation must find its own answers to what exactly the past philosophers were trying to do. It is best to find them in a way that is helpful to one's own contemporaries rather than trying to uncover some eternal philosophical truths.

References

Chomsky, Noam (1965): *Aspects of the Theory of Syntax*, Cambridge, MA: MIT Press.

Fodor, Jerry (1975): *The Language of Thought*, New York: Thomas Y. Crowell.

Gál, Gedeon (1977): 'Adam of Wodeham's Question on the "Complexe Significabile" as the Immediate Object of Scientific Knowledge', *Franciscan Studies* 37(1): 66–102. URL = https://doi.org/10.1353/frc.1977.0006

Ockham, Guillelmus (1967): *Opera Philosophica, Vol. I: Summa Logicae* in *Opera Philosophica et Theologica. Opera Theologica*, St. Bonaventure, NY: St. Bonaventure University.

Panaccio, Claude (1992) : *Les mots, les concepts et les choses*, Oxford: Oxford University Press.

Panaccio, Claude (2023) *Ockham's Nominalism,* Oxford: Oxford University Press.

Raatikainen, Panu (ed.) (2000a): *Universaalikieli: Suomen filosofisen yhdistyksen ja Helsingin yliopiston filosofian laitoksen järjestämä kollokvio 16.4.1999,* Helsingin yliopisto.

Raatikainen, Panu (2000b): 'Universaalikieli-kollokvio 16.4.1999: avaussanat', in Raatikainen, P. (ed.), *Universaalikieli: Suomen filosofisen yhdistyksen ja Helsingin yliopiston filosofian laitoksen järjestämä kollokvio 16.4.1999*, Helsingin yliopisto*,* 5–11.

Russell, Bertrand (1905): 'On Denoting', *Mind* 14(56): 479–493. URL = https://www.jstor.org/stable/2248381

Spade, Paul Vincent, Claude Panaccio & Jenny Pelletier (2002/2024): 'William of Ockham', *The Stanford Encyclopedia of Philosophy* (Fall 2024 Edition), Edward N. Zalta & Uri Nodelman (eds.), URL = https://plato.stanford.edu/archives/fall2024/entries/ockham/

Trentman, John (1970): 'Ockham on Mental', *Mind* 79(316): 586–590. URL = https://doi.org/10.1093/mind/lxxix.316.586

Yrjönsuuri, Mikko (1997): 'Supposition and Truth in Ockham's Mental Language', *Topoi* 16, 15–25. URL = https://doi.org/10.1023/a:1005759617955

Yrjönsuuri, Mikko (2000): 'William Ockham ja ajattelun kieli' in Raatikainen P. (ed.), *Universaalikieli: Suomen filosofisen yhdistyksen ja Helsingin yliopiston filosofian laitoksen järjestämä kollokvio 16.4.1999*, Helsinki: Helsingin yliopisto.

Part III
Reality

15
The adventures of "ontology"

Jani Hakkarainen

Hilary Putnam observed in 2004:

> "[W]hen Quine published a famous paper entitled 'On what there is' ... [he] single-handedly made Ontology a respectable subject." (Putnam 2004, 78–79)

Since this 1948 paper, of course, other philosophers such as Saul Kripke, David Lewis and David Armstrong have done significant work in rehabilitating ontology in the analytic philosophical tradition broadly speaking after the influence of various anti-metaphysical philosophical movements such as logical positivism. It is clear, however, that Quine's influence on the rehabilitation has been significant and not only because of "On What There Is" (Egerton 2021, section 1). Indeed, he has also had a decisive influence on the way "ontology" has been and is understood in mainstream analytic philosophy.

"On What There Is" presents metaphysics primarily as an ontology. In Quine's own words, "A curious thing about the ontological problem is its simplicity. It can be put in three Anglo-Saxon monosyllables: 'What is there'? (Quine 1948, 21)" Ontological problems are, according to Quine, questions about what is there or what exists (Egerton 2021, section 2)? At the same time, he gives "ontology" a slightly new meaning: the task of ontology is to account for the various entities we assume to exist when we take certain propositions to be true. The ontological commitments of the propositions we take to be true can be expressed, according to Quine, by regimenting them in the standard first-order predicate logic and by finding out

Hakkarainen, J. (2025). The adventures of "ontology". In A. Laitinen, M. Keinänen, J. Reinikainen & A. Honkasalo (eds.). *Language, Truth, and Reality.* Tampere University Press. (pp. 175–186).
https://doi.org/10.61201/h5fbdk64

what the propositions bound by the existential quantifier are among the logical consequences of the propositions. We are committed to the existence of exactly those entities that we need in the universe of discourse for the truth of these existential quantifier-bound propositions. (Ibid.)

Indeed, the Quinean approach, perhaps most famously represented by Peter van Inwagen (2001), considers metaphysics as ontology that considers the concept of existence or being and questions about the existence of different kinds of entities such as abstract entities (Berto & Plebani 2015, 2). Since metaphysics in the Quinean approach is primarily ontology, metametaphysics in this approach is called "metaontology" (van Inwagen 2001; Berto & Plebani 2015, 2).

However, Quine was not the first 20th century philosopher to rehabilitate ontology as a legitimate field of philosophy. In post-World War I German philosophy, a renewed interest in ontology emerged after the influence of Neo-Kantianism had waned. In this paper I will make some observations on this philosophical movement, which a central figure in it, Nicolai Hartmann called "new ontology", which I discuss in the second section. Before that, I shall review the history of "ontology" from the coinage of the term and its historical background to Quine's time, including Kant's influence on the "death blow" of ontology (although Kant himself called his transcendental idealism "ontology"), in the first section. I will recapitulate four historical conceptions of ontology and summarise my own view on the subject as a conclusion. All in all, I outline, in addition to the Quinean conception of ontology, five different senses of "ontology", without claiming that these six senses constitute an all-encompassing list of the conceptions of ontology.

The origin and background of "ontology"

The first known systematic examination of the nature of metaphysics is contained in Aristotle's (384–322 BC) work, known for over 2000 years as *Metaphysics*. The term "metaphysics" is known to be a later ancient invention, not used by Aristotle himself. He refers to metaphysics as "the first science" or "the first philosophy" (*protê epistêmê, protê philosophia* in Greek) and "wisdom" (*sophia*), among other things (Politis 2004, 2). In *locus classicus,* at the beginning of the fourth book of *Metaphysics*, Aristotle says that the first science " investigates being *qua* being and what belongs to this [i.e. to being] in virtue of itself [*kath' hauto,* i.e. what belongs to being essentially]." (*Metaphysics* 4.1; translation in Politis 2004, 90). Metaphysics considers being (Greek *to on,* Latin *ens*) in so far as it is being: from the point of view that everything that is is. In metaphysics, the perspective to being is thus the most abstract and general possible, precisely in the sense that being has been separated from all other features except being and its essential features. According to the Aristotle scholar Vasilis Politis (2004, 2), Aristotle's most fundamental question in Metaphysics is, *what is it for something to be*, that is, what is being?

Aristotle's *Metaphysics* has a long and complicated history of interpretation. I will now make a few remarks on its development in the Middle Ages. The Islamic Golden Age philosopher Ibn Sīnā (Avicenna in Latin, c. 970–1037) articulated that, as a universal science, metaphysics considers *being common to* all beings from God to the creation (Goris & Aertsen 2019). In the Latin Middle Ages, he influenced the Dominican Thomas Aquinas, who argued that being in general (Latin: *ens commune*) is the subject matter of metaphysics. Hence metaphysics is a general science (Latin: *scientia communis*). (ibid) In a very similar vein, a little later, the Franciscan Duns Scotus, following Aristotle, held that metaphysics considers being *qua* being (Latin: *ens inquantum ens*) (Lamanna 2021).

At the end of the first chapter of the sixth book of Aristotle's *Metaphysics*, he famously says that the first philosophy is theology because it deals with the highest or divine things: the separate, eternal and immutable substance, or primary being (Greek: *ousia*), if there is such a thing (1026a, 15–33). Aristotle defends his position in Book 12. Knowing this well, however, Ibn Sīnā distinguished metaphysics as a universal science of the common being from theology, even though, according to him, in so far as metaphysics is the study of the first principle, the ground of being, it is theology (Lizzini 2021). This separation of universal metaphysics from theology was followed in different ways by many Latin philosophers of the Middle Ages: as a universal or general science, metaphysics considers God only in so far as He is a being (Goris & Aertsen 2019). The culmination of this distinction is the late 16th century articulation by the Jesuit Francisco Suárez of natural or rational theology as a special science, or more precisely as a special part of metaphysics, because natural theology studies a special being: God (Darge 2014, Lamanna 2014; 2021).

The concept of being seems to apply to everything that is. Such a concept was called *transcendental* by many medieval philosophers from the 13th century onwards, because such a concept transcends even the differences between the categories of being and the distinction between infinite and finite being (Goris & Aertsen 2019). They thought that the concept of being is transcendental because it applies to all beings regardless of their category (there were different views on how it applies) (ibid.). Every being is a being regardless of its category. The concept of being is transcategorical. Similarly, both infinite and finite entities are beings (ibid.). Scotus, for example, influentially called metaphysics a transcendental science (Latin: *scientia transcendens*) because it considers transcendentals (ibid.).

Among others, Scotus argued that, as concepts, transcendentals are primitive, that is, impossible to define, because of their generality. They are therefore, in this sense, the first objects of the understanding.The primacy of metaphysics as a science thus takes on a new character: it is the first science because it is about the first objects of the understanding. Aristotle had thought that the primacy of the first philosophy comes rather from the study of the first principle, that is to say, of the primary being. (Goris & Aertsen 2019.)

Nonetheless, this caused instability in the conception of metaphysics. On the one hand, metaphysics considered being as being and not, for example, just the concept of being. On the other hand, it was a transcendental science, concerned with the first objects of understanding: what is intelligible in the most comprehensive sense. However, the intelligible does not seem to be limited to being, since we can also understand non-beings, such as possibly fictional objects (e.g., centaurs) and privations such as hunger. The conception of metaphysics as a transcendental science thus had a pull to be extended to the "supertranscendentals", which transcend even the transcendentals (Goris & Aertsen 2019). It is precisely "intelligible" (Latin: *intelligibilis*, *cogitable*) that is a possible supertranscendental, as is "something" (Latin: *aliquod*) (ibid.). For example, a centaur is intelligible and something, even though there are no centaurs. One can therefore speak of the supertranscendental conception of metaphysics, according to which metaphysics is the study of the intelligible and thus transcends transcendentals and being (ibid.).

In the late 1500s, Suárez saw himself as having resolved the tension described above by arguing that metaphysics investigates the *real being* that is independent of understanding or thought, which encompasses both God and creation (Lamanna 2021). It does not, for example, study beings of reason (Latin: *entia rationis*), which are intelligible but not real (ibid.) Metaphysics is a real science (Latin: *scientia realis*) (Lamanna 2014). "Being", "something" and "thing" (Latin: *res*) are synonyms (ibid.)

The tension, however, resulted in competing conceptions of metaphysics in the German Calvinist philosophy of the early 17th century, which was deeply influenced by Suárez. It was within this philosophical development that the term "ontology" was coined, as far as we know (Lamanna 2014). Its introducer, Jakob Lorhard (1561–1609), along with Clemens Timpler (1563/4–1624) and Johannes Clauberg (1622–1665) in particular, represented a supertranscendental metaphysical approach. According to Timpler's 1604 textbook on metaphysics, the concept of the intelligible subsumes to the contradictory concepts of something and nothing (Latin: *nihil*) (ibid.) According to him, therefore, we can understand nothing; it is a signified thing. Timpler classifies the concepts of being and essence under the heading of "something", more precisely "positive something" (Smith 2022).

Two years later, in 1606, Lorhard introduced the term "ontology" in his Latin work *Ogdoas Scholastica continens Diagraphen Typicam artium: Grammatices (Latinæ, Graecæ), Logices, Rhetorices, Astronomices, Ethices, Physices, Metaphysices, seu Ontologiæ* (Eight Books of Scholastics [...])", which is effectively a repetition of Timpler's textbook (Lamanna 2014). "Ontology" comes from the Greek genitive form of "to be" (*to on*) "ontos" and "logos". "By "ontology" Lorhard designated the entire metaphysics, as we can see from the title of the work: *Metaphysices, seu* [or] *Ontologiæ*". Clauberg follows this usage in his 1647 work *Elementa philosophiae sive ontosophia*, according to which metaphysics, or ontosophy or ontology, deals primarily with the intelligible (*ens cogitable*), secondarily with something or nothing, and only thirdly with what is (*ens*) (Bardout 2002).

Lorhard's colleague at the University of Marburg in Hesse-Kassel and Timpler's philosophical rival Rudolf Göckel (1547–1628) adopted the "ontology" in 1613 but understood it differently. According to Göckel, ontology is the universal part of metaphysics, the first philosophy to consider being universally and as transcendental (Lamanna 2014). The special part of metaphysics, or "metaphysics", studies God and spirits and is thus divided into theology and angelology (ibid.). Göckel is therefore close to Suárez and rejects the supertranscendental conception of metaphysics.

The influence of these German-speaking Calvinists in the Protestant world is evidenced by the popular 1728 English-language encyclopaedia *Cyclopaedia* by Ephraim Chambers (c. 1680–1740) with the entry "ENS" (Chambers 1728, 315). Like Timpler, Lorhard and Clauberg, Chambers says that "ens" in the most general sense applies to everything that can be understood. In a slightly more specific sense, "ens" is anything that is or exists: an entity. Its opposite is the non-existent. Chambers, like Clauberg, uses the term "ens positivum" for an entity. In the proper sense, however, "ens" applies to a real being to which real attributes belong. (Ibid.) Here again we see the influence of Suárez. As regards "ontology", Chambers says that it is a doctrine or science about being (*ens*) in general, abstractly speaking (Chambers 1728, 663). Metaphysics is ontology or ontosophy in the abstract: a doctrine of being *qua* being (Chambers 1728, 543). Chambers identifies ontology and metaphysics. However, what Chambers says about being (*ens*) must be applied to ontology and metaphysics. Ontology, in his view, can therefore also be supertranscendental. Chambers tries to cover the different conceptions of ontology and metaphysics in the style of a good encyclopedist.

Göckel's terminological solutions, on the other hand, proved influential in 17th-century German Protestant philosophy. According to the Latin works of Lutheran Christian Wolff, written while he was a professor at Marburg, ontology, or first philosophy, is the science of entities in general in so far as they are entities, of being as being and its general predicates (Wolff 2022/1730, §1 and 1963/1728, §73). In his German-language Logic he calls ontology the "Grund-Wissenschaft" (Wolff 1770/1713, §14). Add to this the study of the corporeal world in general and of spirits, that is, created souls, angels and God, and, according to Wolff, the subject of metaphysics is obtained. Metaphysics thus consists of ontology, rational cosmology, psychology and theology. To his, in the German-language *Metaphysics* (1719), Wolff also includes with empirical psychology: the soul a posteriori. Metaphysics thus also considers such topics as the general structure, foundation, causality and purpose of the corporeal world, the nature of the soul, its faculties, its relation to the body and immortality, freedom of the will, and the nature, faculties and creation of God.

According to Wolff's influential conceptualization, metaphysics divides into ontology, rational cosmology, psychology and natural theology (Wolff 1963/1728, §99). Even though Wolff does not ever use these exact terms from the 17th century (Micraelius 1661, 770), it has long been common to say that ontology is general metaphysics (Latin: *metaphysica generalis*) and rational cosmology, psychology and natural theology are special metaphysics (Latin: *metaphysica specialis*) in his view. The historical roots of the distinction between general and special metaphysics can be found especially in Suarez, who considered natural theology to be a special part of metaphysics, separated from the general part of metaphysics, the study of the real being as being, as was seen above.

Wolff's disciple Alexander Gotlieb Baumgarten followed his teacher in the division of metaphysics into “ontology, cosmology, psychology, and natural theology.” (Baumgarten 2013/1757, §2) Metaphysics is the science of the first principles of human cognition (Baumgarten 2013/1757, §1). Indeed, the first principles of cognition are to be the more general predicates, which already Aristotle speaks of as the essential features of being: “ONTOLOGY (ontosophia, metaphysics (cf. §1), universal metaphysics, architectonics, first philosophy) is the science of the more general predicates of a being.” (Baumgarten 2013/1757, §4) The more general predicates of being are the first principles of human cognition (Baumgarten 2013/1757, §5). Of these, "the universal predicates [...] are in each and every single thing [*singulis*]" (Baumgarten 2013/1757, §6). For example, each of us is *one being* and *something*.

Wolff and Baumgarten are known to have had a profound influence on Kant. Kant lectured nearly fifty times on Baumgarten's *Metaphysics* over four decades. They also had a profound influence on Kant's understanding of metaphysics and ontology as fields of philosophy, although Kant is highly critical of Wolff's and Baumgarten's first-order metaphysics. In Kant's critical period, metaphysics and ontology must be understood in the context of his transcendental philosophy, which considers the necessary presuppositions of things like metaphysics and possible experience.

At the end of the *Critique of Pure Reason* (1781; 1786), in the *architectonic of pure reason,* Kant further divided metaphysics into ontology, rational cosmology, psychology and theology, alongside "rational physics" (CPR, A 847; B 875).[1] However, Kant understood metaphysics and ontology in a new way. In both, the type of cognition (German: *Erkenntnis*) is a priori: independent of experience (CPR, A 841 and 845; B 869 and 873). For Kant, metaphysics is a system of all *a priori cognition*, e.g., the categories of the understanding, and no longer the science of being and its determinations (CPR, A 841; B 869). Kant also performs a "Copernican turn" in the conception of metaphysics. Metaphysical judgments must be synthetic and a priori (Kant 1997, §4). In the *Prolegomena* (1783) Kant sums up as follows: metaphysics "is therefore cognition *a priori*, or from pure understanding and pure reason." (Kant 1997, §1)

This explains why, for Kant, metaphysics is both speculative and practical use of reason a priori, or "*metaphysics of nature*" and "*metaphysics of morals*" (German: *Metaphysik der Sitten*) (CPR, A 841; B 869). Metaphysics of nature, or metaphysics in the narrow sense, “considers everything as it is (not that which ought to be)” (CPR, A 845; B 873). The metaphysics of nature is divided in two. One is *ontology*, which Kant identifies with his transcendental philosophy: in Kant's technical parlance, ontology considers “a system of all concepts and principles that are related to objects [German: *Gegenstand*] in general” (CPR, A 845; B 873). By “objects in general” Kant means here objects which can be given but which are not assumed to be given (ibid.). Ontology

[1] Kant 2013 is cited by “CPR”, followed standardly by page numbers in the A and B editions.

is concerned especially with the concepts and principles that are related to objects that may be given in experience, such as everyday objects (e.g., stones and stumps), in particular the categories of the understanding.

Instead, the sum total of objects given to the senses or otherwise is, for Kant, "nature", which is the subject matter of the second part of the metaphysics of nature. Kant also divides it into two parts. If it is a set of objects given to the senses, that is to say, immanent, it is considered "*a priori*" by "rational physiology" (CPR, B 874; A 846). Rational physiology is divided into "rational physics" and "rational psychology" according to whether the object is "**corporeal**" (bodies), or "**thinking nature**" (souls) (ibid.).

Kant sums up his conception of metaphysics as follows: "Accordingly, the entire system of metaphysics consists of four main parts, **i. Ontology. 2. Rational Physiology** [Physics and Psychology]**. 3. Rational Cosmology. 4. Rational Theology**." (CPR, B874; A 846)

According to Kant, ontology, if it is a science, does not go beyond the limits of the understanding but remains within them. On the other hand, it is clearly not limited to being, since it is concerned in particular with the concepts and principles of the understanding. In this respect, Kant's conception of ontology could be characterized as supertranscendental perhaps, although it is conditioned by the forms of sensibility and categories of the understanding, unlike, for example, Timpler's. It is more accurate to say, thus, that Kant's conception of metaphysics and ontology is transcendental in his terms, which is clearly different from Scotus' view of metaphysics as a transcendental science and from Wolff's and Baumgarten's conceptions of ontology. Baumgarten's more general predicates such as *being one* understood as transcendental by Scotus are transformed in Kant's transcendental philosophy into the concepts of possible given objects such as empirical objects as the object of study of ontology. Whereas many scholastics thought of metaphysics as the study of being as a being and its essential attributes, Kant's ontology is essentially concerned with possible empirical objects and their constitutive concepts and principles, especially categories of the understanding. The *Transcendental Aesthetic* and *Analytic* parts of the *Critique of Pure Reason* thus consider ontology in Kant's terms.

The return of ontology in early 20th century German philosophy

Regardless of Kant's own views, under the influence of his criticism of metaphysics and ontology before him, especially that of Wolff and Baumgarten, metaphysics and ontology, as legitimate fields of philosophy, suffered a serious setback. However, alongside many other currents such as Neo-Kantianism, positivism and pessimism, 19th-century German scholarship also experienced an Aristotelian renaissance, partly caused by the strong development of classical philology (Hartung, King & Rapp 2019). One of its prominent representatives was Franz Brentano (1838–1917), whose 1862 dissertation dealt with the many meanings of being in Aristotle (Brentano 1975). Brentano was a fierce opponent of Kant's philosophy, who took his influences from the empiricist philosophical tradition, from John Stuart Mill through Hume and the Scholastics to Aristotle himself. He rehabilitated the traditional Aristotelian conception of metaphysics as the study of being as being and its categories as distinct from the categories of the understanding in Kant (Albertazzi 2016). Brentano was a popular and influential teacher at the universities of Würzburg and Vienna. His influence was thus both through publications and, in particular, through teaching.

According to Edmund Husserl (1859–1938), a student of Brentano and the founder of phenomenology, metaphysics asks, what is there? (Hartimo 2019) Ontology, on the other hand, is, according to Husserl, a more general field of study, because ontology considers the a priori essences of possible objects in themselves (German: *Objekt an sich*) (ibid.). Objects in themselves are any things (German: *Ding*) that can be the bearers of predicates applicable to them (Moran & Cohen 2012, 228, 317). Essences are, for Husserl, necessary features for the conception of an object (Belt 2021). Ontology is thus, for Husserl, the study of the essences of objects that may appear meaningfully to us. Since actual objects are also possible, ontology is also concerned with their essences.

In this respect, however, Husserl's main new opening is the introduction of *formal ontology* as a branch of ontology and its separation from material or regional ontologies. He understands formal ontology as the study of the ontological categories of possible objects in themselves (Hartimo 2019). Since the set of these categories is thus universally applicable to possible objects in themselves, formality in this context means universal applicability across the domains of possible objects in themselves (Hakkarainen & Keinänen 2023, 9). Formality, in scholastic terms, is transcendentality. In contrast, when one considers only some restricted domain of possible objects in themselves, one is, according to Husserl, doing a regional or material ontology such as the ontology of mind (Moran & Cohen 2012, 278). A formal ontology is a top-level ontology, under which each regional or material ontology is subsumed. Husserl's notion of a regional ontology comes close to more traditional notion of a special metaphysics.

Among Husserl's students, at least Edith Stein, who combined phenomenology with Thomism in the 1920s and 1930s, and the realist phenomenologists Hedwig Conrad-Martius (1888–1966) and Roman Ingarden did ontology. In 1935, the Baltic German philosopher Nicolai Hartmann (1882–1950) saw that ontology had made a comeback in philosophy, especially after the First World War, when the grip of Neo-Kantianism had loosened in Germany. He spoke of a "new ontology", to which he included, alongside himself, Heidegger, Stein, Conrad-Martius, the German theologian and philosopher Günter Jacoby (1881–1969), and the philosophical anthropologist Max Scheler (1874–1928) (Hartmann 2019/1935, 3). According to Hartmann, however, the new ontology had been more of a programmatic declaration than an actual philosophical project (ibid.). He considered himself to have realised it with his own ontological system, which he set out in *Ontology: Laying the Foundations*

(*Zur Grundlegung der Ontologie*, 1935), *Possibility and Actuality* (*Möglichkeit und Wircklichkeit*, 1938), *Der Aufbau der realen Welt* (1940) and *Philosophie der Natur* (1950). It is therefore important to understand that philosophy from the 1920s to 1940s was not in this respect simply the anti-metaphysics of the logical positivists. A rehabilitation of ontology took place in German philosophy, which played a very important role, before Quine, although Quine's rehabilitation has had a more lasting impact so far.

Hartmann was not a phenomenologist and certainly not a philosopher of existence, although he has also been read as a philosophical anthropologist (Peterson 2019). Nevertheless, a certain kind of phenomenology has its place in his philosophical method, although he was quite critical of Husserl (ibid.). Hartmann considers phenomenology to be the systematic collection of relevant evidence. It is the first stage of philosophical inquiry (ibid.). Ontology is understood traditionally by Hartmann as the study of being as being and the categories of being (Hartmann 2019/1935, 7 and 51). Metaphysics Hartmann understands as a set consisting of specific metaphysics: cosmology, rational psychology and theology (Peterson 2019).

His important insight is an understanding of the fundamental question of ontology (*Grundfrage* in German), although he is influenced by Aristotle and Hegel (Hartmann 2019/1935, 51). The fundamental question of ontology is, what is being itself (German: *das Sein selbst*)? (Hartmann 2019/1935, 54) By being itself, Hartmann means that being is in no way conditioned to the conscious subject. It is not assumed, for example, that being depends or does not depend on the subject (Hartmann 2019/1935, 57). Hartmann critisises Martin Heidegger for relativising the question of being to the human subject through the concept of the meaning of being (German: *der Sinn des Seins*) (Hartmann 2019/1935, 55–57). According to Hartmann, the starting point of ontology is neutral in relation to the distinction between idealism and realism (Hartmann 2019/1935, 51–52). However, ontological research can only proceed through the phenomenology of beings towards the study of being itself and its categories (Hartmann 2019/1935, 58–60).

Conclusion: six conceptions of ontology

On the basis of this brief historical overview, we can discern, in addition to the Quinean conception of ontology, four earlier conceptions of what ontology is. (1) Lorhard, who coined the term "ontology", represented a supertranscendental conception of ontology as the study of the intelligible. It is not limited to the study of being as being, since not everything that can be understood is a being in his view. (2) In contrast, Göckel thought that ontology is the study of being in general. His conception, then, is really that ontology is the science of Aristotle's being as being and of the essential features of being, conceived by Avicenna and Aquinas as the study of the common or general being of God and creation. They are all united by the fact that they are.

In the early 18th century, Wolff explicitly distinguished ontology, or general metaphysics, from specific metaphysics that investigate some more limited domain of being: bodies (cosmology), souls (rational psychology), or God (natural theology). (3) Kant opposed Wolff when he argued in his transcendental philosophy that ontology concerns a system of concepts and principles referring to objects in general, in particular the categories of the understanding that partly constitute empirical objects. Kant's conception of ontology thus has certain affinities with the supertranscendental conception (1).

(4) In the early phenomenological tradition, its founder Husserl saw ontology as the study of the essential structures of objects in themselves as appearing meaningfully to us. In other words, ontological research is concerned with the essence of any object that can be appear to us meaningfully. This phenomenological conception of ontology has obvious connections with Kant's conception of ontology, especially if interpreted within the framework of transcendental idealism. It was followed by Husserl's pupils Stein and Ingarden, the latter of whom, however, did not accept transcendental idealism. Immediately after Husserl, the critical realist Hartmann returned to the view of ontology as a general metaphysics concerning being as being and its categories.

(5) In *Formal Ontology*, I and Markku Keinänen have defended a view that extends the Quinean conception of ontology with considerations of grounding and fundamentality (Hakkarainen & Keinänen 2023, 57–58). According to us, ontology does not only investigate existence questions, but also possible hierarchies of grounding and fundamentality of entities (ibid.). For example, the classical question of the possible primary being, or in modern terms, metaphysically fundamental being, is in our view an ontological question (Hakkarainen & Keinänen 2023, 57).

It is also essential, we argue, that the very ontological problem settings presuppose something about formal ontology and general metaphysics, which we separate from ontology (Hakkarainen & Keinänen 2023, 59–62). Formal ontology considers categories of being by analysing them through ontological forms (Hakkarainen & Keinänen 2023, 58). For example, the forms of being of a substance can be considered as follows: ontologically independent individual entity, persistent bearer of properties. We cannot even pose the ontological question of the existence or fundamentality of substances without presupposing something about what it is to be a substance if there are substances. At the same time, we have to assume something about what it is to exist. For example, is existence the same as being? What is their relation to their opposites? Does existence modify or is it uniform? These are, according to us, the questions of general metaphysics that are presupposed by ontological and formal ontological questions and answers (Hakkarainen 2023, 138). In our view, general metaphysics is what, for example, Göckel proposed ontology to be. Our argument for our view can be summarized by saying that our view delineates the objects of study of metaphysics in a unifying manner (for more details, Hakkarainen & Keinänen 2023, 58–62).

References

Aristotle (1958): *Metaphysica*, W. Jaeger (ed.), Oxford: Clarendon Press.

Arp Robert, Smith Barry & Spear Andrew (2015): *Building Ontologies with Basic Formal Ontology*, MIT Press, Cambridge, MA.

Bardout, Jean-Christophe (2002): 'Johannes Clauberg', in Steven Nadler (ed.), *A Companion to Early Modern Philosophy*, Malden, MA: Blackwell, 29–139.

Baumgarten, Alexander (2013/1757): *Metaphysics*, C. D. Fugate & J. Hymers (trans. and ed.), London: Bloomsbury.

Berto, Francesco & Plebani, Matteo (2015): *Ontology and Metaontology. A Contemporary Guide*, London: Bloomsbury.

Brentano, Franz (1975): *On the Several Senses of Being in Aristotle*, R. George (trans.), Berkeley: University of California Press.

Chambers, Ephraim (1728): *Cyclopædia: or, An Universal Dictionary of Arts and Sciences (1 ed.),* James & John Knapton; John Darby; and others, London.

Darge, Rolf (2014): 'Suárez on the Subject of Metaphysics', in V. M. Salas & R. L. Fastiggi, (eds.), *A Companion to Francisco Suárez*, Leiden: Brill, 91–123.

Egerton, Karl (2020): 'Quine's metametaphysics', in Ricki Bliss & James Miller (eds.), *The Routledge Handbook of Metametaphysics*, New York, NY: Routledge, 49–60.

Goris, Wouter, & Aertsen, Jan (2019): 'Medieval Theories of Transcendentals', in *The Stanford Encyclopedia of Philosophy* (Fall 2019 Edition), Edward N. Zalta (ed.). URL = https://plato.stanford.edu/entries/transcendentals-medieval/.

Hakkarainen, Jani (2023): 'Learning from the Past to the Future in Metaphysics', in Jani Sinokki & Eero Kaila, *Past. Future. Philosophy*, Acta Philosophica Fennica 99, Finnish Philosophical Society, 125–141.

Hakkarainen, Jani & Keinänen, Markku (2023): *Formal Ontology*, Cambridge: Cambridge University Press.

Hartimo, Mirja (2019): 'Husserl on 'Besinnung' and Formal Ontology', in F. Kjosavik & C. Serc-Hanssen (eds.), *Metametaphysics and the Sciences*: *Historical and Philosophical Perspectives*, New York, NY: Routledge, 200–215.

Hartmann, Nicolai (2019): *Ontology: Laying the Foundations*, K. Peterson (trans), Boston: De Gruyter.

Hartung Gerald, King Colin & Rapp Christof (2019): 'Introduction: Contours of Aristotelian Studies in the 19th Century', in G. Hartung, C. G. King, & C. Rapp (eds.), *Aristotelian Studies in 19th Century Philosophy.* Leiden: De Gruyter, 1–10. URL = https://doi.org/10.1515/9783110570014-002.

Kant, Immanuel (1997): *Prolegomena, or Introduction to any metaphysics which may in future be conducted by science* (Prolegomena zu einer jeden künftigen Metaphysik), G. Hatfield (trans. & ed.), Cambridge: Cambridge University Press.

Kant, Immanuel (2013): *The Critique of Pure Reason*, (Kritik der reinen Vernunft), P. Guyer & A.W. Wood (trans. & ed.), Cambridge: Cambridge University Press.

Lamanna, Marco (2014): 'Ontology between Goclenius and Suárez', in L. Novák (ed.), *Suárez's Metaphysics in Its Historical and Systematic Context*, Berlin: De Gruyter, 135–152.

Lamanna, Marco (2021): 'Francisco Suárez's Ontology (Science of Being)', in *Conimbricenses Encyclopedia,* M. Santiago de Carvalho & S. Guidi (eds.), www.conimbricenses.org/encyclopedia/suarez-francisco- ontology-science-of-being. URL = https://doi.org/10.5281/zenodo.4934552

Lizzini, Olga (2021): 'Ibn Sina's Metaphysics', *The Stanford Encyclopedia of Philosophy* (Fall 2021 Edition), Edward N. Zalta (ed.), URL = https://plato.stanford.edu/archives/fall2021/entries/ibn-sina-metaphysics/.

Micraelius, Joh (1661): *Lexicon Philosophicum Terminorum Philosophis Usitatorum*, Stetini, impensis Jeremiae Mamphrasii, Bibliop., Typis Michaelis Höpfneri.

Moran, Dermot & Cohen, Joseph (2012*): The Husserl Dictionary*, New York: Continuum International.

Peterson, Keith R. (2019): 'Translator's Introduction: Hartmann's Realist Ontology', in N. Hartmann, *Ontology: Laying the Foundations*, Boston: De Gruyter, xv–xxxix. URL = https://doi.org/10.1515/9783110627350-001.

Politis, Vasilis (2004): *Routledge Philosophy Guidebook to Aristotle and the Metaphysics*, London: Routledge.

Putnam, Hilary (2004): *Ethics Without Ontology*, Cambridge, MA: Harvard University Press.

Quine, Willard (1948): 'On What There Is', *Review of Metaphysics* 2(5): 21–38. URL = https://www.jstor.org/stable/20123117.

Tahko, Tuomas (2015): *An Introduction to Metametaphysics*, Cambridge: Cambridge University Press.

Smith, Barry (2022): 'The Birth of Ontology', *Journal of Knowledge Structures and Systems,* 3(1): 57–66.

Van Inwagen, Peter (2001): *Ontology, Identity, and Modality*, Cambridge: Cambridge University Press.

Wolff, Christian (1719): *Vernünftige Gedanken von Gott, der Welt und der Seele des Menschen, auch allen Dingen überhaupt*, Renger, Halle.

Wolff, Christian (1770): *Logic, or Rational Thoughts on the Powers of the Human Understanding with their Use and Application in the Knowledge and Search of Truth*, L. Hawes, W. Clarke, and R. Collins London.

Wolff, Christian (1963): *Preliminary Discourse on Philosophy in General* (*Discursus praeliminaris de philosophia in genere*), R. J. Blackwell (trans.), Indianapolis: The Bobbs Merrill Company.

Wolff, Christian (2022): *First Philosophy, or Ontology Treated According to the Scientific Method, Containing the Principles of All Human Cognition Part 1*, K. Ottmann (trans.), Thompson: Spring Publications, §§1–78.

16
Ten queries about Hasok Chang's pragmatic realism

Ilkka Niiniluoto

Scientific Realism vs. Pragmatism

Hasok Chang works as Professor at the Department of History and Philosophy of Science at the University of Cambridge. In his earlier works he has studied the measurement of temperature (*Inventing Temperature*, 2004) and progress in the history of chemistry (*Is Water H_2O?*, 2012). Chang's new philosophical book *Realism for Realistic People: A New Pragmatist Philosophy of Science* (2022) is a rich and impressive improvement of neopragmatist accounts of science. With inspiration from William James, it presents a coherence theory of reality and truth as a basis of pluralism. Friendly but critical references to my *Critical Scientific Realism* (CSR, 1999) give an opportunity for dialogue and fruitful confrontation.[1]

The debate between realists and pragmatists started already as a divide between Charles S. Peirce and his "kidnappers" (cf. Haack, 2024) and has grown to one of the most important themes in contemporary philosophy. Several philosophers have developed systems which combine elements from realism and pragmatism—among

[1] The first version of this paper was presented in a workshop "Can Realism Allow Pluralism and Contingency?", Université Lorraine, Archives Henri-Poincaré – Philosophie et Researches sur les Sciences et les Technologies, Nancy, France, October 10, 2022. I am grateful to Hasok Chang for an instructive and lively discussion. A later version was presented in the XXV World Congress of Philosophy, Rome, August 6, 2024.

Niiniluoto, I. (2025). Ten queries about Hasok Chang's pragmatic realism. In A. Laitinen, M. Keinänen, J. Reinikainen & A. Honkasalo (eds.). *Language, Truth, and Reality.* Tampere University Press. (pp. 187–195).
https://doi.org/10.61201/h5fbdk64

them Hilary Putnam's (1981) internal realism, Sami Pihlström's (1996) pragmatic realism, Philip Kitcher's modest or real realism (cf. Gonzalez, 2011), Hasok Chang's (2018) pluralist realism, and Rein Vihalemm's practical realism (cf. Mets et al. 2024).[2] With my own sympathies in realism, I have always thought that pragmatism is more interesting than naive realism (Niiniluoto, 1986, 67). Critical realism is shared by Panu Raatikainen (2004a, 2014), to whom it is a pleasure to dedicate the thoughts of this article.

Chang on reality

According to Chang's operational conception, an entity is *real* if it has "the capacity to support coherent activities" (p. 122), "to the extent that there are operationally coherent activities that can be performed by relying significantly on its existence and its properties" (p. 121), where the "various parts of an activity come together in harmonious way towards the achievement of its aims" (p. 24).

Chang's ontology is directed against "standard" or metaphysical realism with the fallacy of prefigured or ready-made Kantian things-in-themselves: realities are *mind-framed* yet *mind-uncontrolled* entities. "We can make concepts as we like, but whether the entities they specify turn out to be *real* is not up to us" (p. 125); it is "best to have no word to call the noumenal Being independent of our conceptions" (p. 77).

Chang on truth

Chang's pragmatism is not interested in propositional knowledge-as-information but rather in knowledge-as-ability to do something. He defines primary truth-by-operational coherence: "a statement is *true* to the extent that there are operationally coherent activities that can be performed by relying on its content" (p. 167). Inspired by James, this is indeed intended as a *definition* of truth, even though some neopragmatists (Putnam and Pihlström) have suggested that when James asked about "the truth's cash-value in experiential terms" he did not try to define truth at all. Thus, for Chang coherence is not a *criterion* of truth as the Marxists state about practice.[3] It follows that for Chang truth does not *explain* success. This excludes the abductive no miracle argument for scientific realism, which claims that realism is the best (or even the only) explanation of the success of science (p. 109).

Chang also defines a secondary concept of truth as a derivation of a statement from other truths (p. 165). But it is clear that Chang's primary and secondary notions

[2] For comments, see Niiniluoto (2019).

[3] Rein Vihalemm's practical realism, which is inspired by Marxism, uses a deflationary notion of truth (see Mets et al., 2024, p. 3).

of truth differ from the classical correspondence theory of truth as a relation between language and reality.

I will present ten queries and potential objections to Chang's operational definition of reality and truth.

First query: noumena

Immanuel Kant argued that human knowledge has subjective and objective elements: the world of *phenomena* or things-for-us is constituted by the transcendental subject (the mind supplies time and space as forms of sensible intuition as well as categories of understanding like causality), but it is also caused by the *noumenal* world of things-in-themselves. This was a contradiction by Kant's own standards, since causality should apply only to phenomena. The idealists reacted by rejecting the things-in-themselves, and the phenomenalists eliminated the transcendental subject as well. Kant combined his transcendental idealism with empirical realism but allowed no knowledge about things-in-themselves. The critical realists reacted by arguing that we have knowledge about the noumena by science. Kant also approved universalism by assuming that we are bound to only one conceptual scheme. The pragmatists reacted by allowing a plurality of alternative conceptual frameworks. James's radical empiricism was close to phenomenalism.

But does the mind-independent noumenal WORLD exist? A negative reply has been given in three different forms:

(a) false metaphysical assumption (idealism)
(b) presupposition not needed, left open (Husserl's *epoche*)
(c) meaningless metaphysical claim (young Carnap's logical positivism).

The position of many classical and neopragmatists is ambiguous between alternatives (a) – (c) (e.g. Putnam, Rescher, Rorty, Margolis, Pihlström, Vihalemm) (cf. Niiniluoto, 2019). When Chang does not assume Kantian things-in-themselves in his ontology, is his position (a), (b), or (c)?

Second query: grounding

Chang's notion of "mind-framing" has a Kantian flavor. He gives a fine emphasis on the lack of mind-control of real entities, but how can he explain this without assuming something *ontological* about the mind-independent WORLD?

In CSR, the WORLD is a lawlike flux of causal processes, where physical objects and their kinds and other entities (fields) are identifiable by their physical properties and spatio-temporal continuity; the conceptualized world-versions or L-worlds W_L (for various languages L) are mind-framed by L and mind-uncontrolled by the

WORLD,[4] i.e. the WORLD determines which sentences of a language L are true or false in $W_{L.}$ In modern analytical metaphysics, this non-causal dependence relation between a L-world W_L and the WORLD is called *grounding* (cf. Niiniluoto, 2024).

Third query: too positive

Is Chang's operational definition *too positive*? We often stumble on hard realities which *prevent* our activities or make them *less* successful. Peirce called this "brutal" ability of reality to resist us and our will "factuality" or "secondness". Such hard realities may belong to the natural world: stone walls, heavy rains, heat waves, earthquakes, and other calamities. They may also show the reality of other people: prejudice, hostility, aggression, and cheating.

Fourth query: approximate truth

For Chang realness and truth have degrees ("true enough"), but the notion of *approximate truth* is a "watering-down move" for him (p. 250). One may wonder why? In CSR, truthlikeness is a tool of critical realism against absolute and naive realism. It presupposes objective truth as the target to be approached, but some philosophers have defined epistemic truthlikeness without objective truth.

Fifth query: too restrictive

Chang is sensitive to traditional objections to pragmatism, but is his account *too restrictive*?

The argument from the past (p. 73) points out that dinosaurs existed on the earth long before they were identified as dinosaurs. For Chang such pre-human or "past entities must be framed by our current conceptions if we are to consider them at all", but one may wonder whether such talk about past entities is allowed at all, if "all entities are mind-framed" (p. 77, 133). To apply Chang's operational definition, what present coherent operational epistemic activities could rely on the past existence of dinosaurs? Perhaps the classification of dinosaur fossils (p. 123), but then there is the danger of losing the distinction between fact and fiction, since also pictures of unicorns can be classified coherently.

[4] A similar construction of ”ontic domaims” from categorical-conceptual frameworks and the independent noumenal reality is given by Lombardi (2024). Like my L-worlds, her ontic domains serve as objective truth-makers of linguistic statements.

The inaccessibility argument (p. 123) acknowledges that past objects and events are abductively accessible to us by the causal effects or traces that they have left. But most of them were real but left no traces.[5]

One may add here the argument from irrelevance: could all particular grains of sand, raindrops, mosquitos, stars, black holes, atoms and unobservable objects in distant galaxies be relevant to human practices?

Sixth query: too permissive

The objection from effective false beliefs (p. 189) asks: Is the coherence theory of reality *too permissive*?

As belief in God's existence may have effects in a person's actions, does it follow that a supernatural Being exists? This was a traditional debate about James's "will to believe" doctrine.

Similarly, many cultures have led harmonious and successful lives with animistic beliefs in angels, fairies, brownies, witches, and evil spirits. These queer entities have been discussed by cultural relativists, but do they really exist?

Examples of effects without realities might also include witchcraft and placebo in medicine.

According to Chang, "pluralist ontology becomes easily acceptable when we move away from the fallacy of pre-figuration"; it is beneficial to encourage multiple ontologies, each of which can facilitate coherent epistemic activities (p. 148). In science, phlogiston had some successful applications, and thus it is real for Chang, but as a description of the process of combustion it cannot be accepted in our world view. The pluralist ontology with conflicting postulated objects (e.g. phlogiston and oxygen) leads to a too cumulative *conservationist* model of scientific change: as most historical theories in science were to some extent successful, Chang's advice to the scientists is not to discard them – even though they were later surpassed by more powerful and truthlike theories. This view resembles Paul Feyerabend's anarchism without Popperian falsification: knowledge is "an ever-increasing ocean of mutually incompatible (and perhaps even incommensurable) alternatives", where "nothing is ever settled, no view can ever be omitted from a comprehensive account" (Feyerabend, 1975, p. 30).

[5] This is a variant of Bertrand Russell's famous argument (the unknown but true number of Churchill's sneezes in 1940) against John Dewey's notion of truth as warranted assertability.

Seventh query: conceptual pluralism with correspondence truth

Chang notes my *conceptual pluralism* in CSR: the WORLD can be approached and described by alternative conceptual systems L. But does conceptual pluralism really require the operational notion of truth?

CSR defends conceptual pluralism which is compatible with the correspondence theory of truth, against Hilary Putnam's (1981) internal realism which combined conceptual pluralism with an epistemic notion of truth as ideal acceptability.[6] The main point is that world-versions W_L are L-structures, and truth in W_L is defined by Tarski's model theory (cf. p. 79). But as conceptualizations of the same reality these world-versions cannot contradict each other (even though beliefs in different languages can); and truth about W_L is also truth about the WORLD.

When scientific realists argued against Kant that *Dinge an sich* are knowable, they are claimed to be inexhaustible rather than inaccessible. For CSR it is important that there is no single ideal language L (Wilfrid Sellars's "Peirceish") which captures all of the WORLD. A similar formulation can be found in James's pluralism: "There is no where extant a complete gathering up of the universe in one focus".[7]

Eight query: humility

Does Chang's coherence account overlook important aspects of the world? Pragmatism is interested in the human world constituted by our practices, and we bear responsibility for this world-for-us (Pihlström, 2022). But human existence is only a tiny fragment of the long history of the universe. Cosmologists study the first seconds after the Big Bang, and it took almost 15 billion years before the human period began. The conception of the WORLD expresses this *humility* with respect to the mind-independent reality. We can investigate it by introducing conceptual schemes as mediating steps in our search for objective knowledge. And we are also responsible for the (often unintended and non-conceptualized) causal effects of our actions on nature (pollution, climate change, loss of biodiversity).

Ninth query: human action

Does Chang's coherence account overemphasize the role of human action? He suggests that his account gives "a positive view that you can use for your own purposes" (p. 9). In my view, however, pragmatism is limited as a philosophy of natural science or basic research, but it is more promising as a philosophy of human

[6] For Kitcher's similar position, see Gonzalez (2011), p. 178. However, Kitcher thinks that James did not reject the idea of correspondence truth but rather demystified it (p. 176).

[7] In a letter to Minot Judson Savage in 1910.

practical action, including applied science, technology in the broad sense, and engineering.[8] If human action includes Aristotle's *praxis*, in addition to *poiesis* and *techne*,[9] pragmatism can be developed also as a philosophy of the humanities (see Pihlström, 2022).[10]

Tenth query: tools and pluralism

Should Chang acknowledge the difference between truthful statements and tools? I agree with Alfred Tarski that alternative epistemic conceptions (e.g. credibility, confirmation, assertability, operational coherence), which do not satisfy the T-equivalence ('p' is true iff p),[11] may be highly valuable as *criteria of reality and truth* but they should not be called by the name "truth" (CSR, 100).[12] So perhaps scientific realism and pragmatism can live in peaceful co-existence, if the revitalized Jamesian activity-oriented coherence conceptions are understood to express notions that are different from the classical realist's reality and truth - such as intellectual *tools* and their *effectiveness* in Dewey's instrumentalism. As material artefacts, tools are mind-framed and controlled by mind-independent laws of nature.

Based on such reading, we can learn a lot from Hasok Chang's magnificent book. For example, we have a clear motivation for pluralism: Newton's mechanics is still used by engineers; we ride bikes despite cars and trains; chalkboards and printed books can be used in classrooms despite computer technologies and the internet; tool-like realities deserve to be conserved if they help us to do important things in certain contexts.

[8] See Chang's (2024) recent work on battery science, which is based on the interaction between the theory of electricity and new technologies.

[9] Dewey's instrumentalism was interested in useful problem-solving. Even though he had difficulties with the notion of intrinsic value, he had a place for future-oriented ”ends-in-view” or ”plans”, with discussion of topics like democracy, education, and the fine arts.

[10] On the other hand, as Raatikainen (2004a), 83, acutely observes, critical realism can give an account of the human sciences by treating ”ideal entities” like beliefs, values, meanings, and conceptual frameworks as unobservable theoretical concepts. They are not independent of the human mind in general, but of the mind of the researcher.

[11] Cf. the fifth and sixth queries.

[12] Chang disagrees when he states that Dewey's move to talk about ”warranted assertability” instead of truth was ”unwise” (p. 206). Raatikainen (2004b, 2014) gives good reasons to reject Michael Dummett's verificationist notion of truth as provability or assertability, which had a profound influence in Putnam's (1981) conversion from classical realism to internal realism. For Raatikainen's comments on Tarski's semantic concept of truth, see Raatikainen (2023).

References

Chang, Hasok (2018): 'Is Pluralism Compatible with Realism?', in Saatsi, J. (ed.), *The Routledge Handbook of Scientific Realism,* London: Routledge, 176–186.

Chang, Hasok (2022): *Realism for Realistic People: A New Pragmatist Philosophy of Science,* Cambridge: Cambridge University Press.

Chang, Hasok (2024): 'Practice-Oriented Realism in the Tradition of Rein Vihalemm', in Mets et al. (eds.), *Practical Realist Philosophy of Science. Reflecting on Rein Vihalemm's Ideas,* Lanham: Lexington Books, 121–141.

Feyerabend, Paul (1975): *Against Method: Outline of an Anarchistic Theory of Knowledge,* London: New Left Books.

Gonzalez, Wenceslao (ed.) (2011): *Scientific Realism and Democratic Society: The Philosophy of Philip Kitcher,* Amsterdam: Rodopi.

Haack, Susan (2024): 'Ugly Enough to be Safe from Kidnappers: 'Pragmatism' and 'Pragmaticism', and the Ethics of Terminology', *Transactions of the Charles S. Peirce Society* 60(1): 1–22. URL = https://doi.org/10.2979/csp.00017

Lombardi, Olimpia (2024): 'Pluralist Realism: Where Onticity and Practice Meet', in Mets et al. (eds.), *Practical Realist Philosophy of Science. Reflecting on Rein Vihalemm's Ideas,* Lanham: Lexington Books, 93–120.

Mets Ave, Löhkivi Endla, Müürsepp Peeter & Eigi-Watkin Jaana (eds.) (2024): *Practical Realist Philosophy of Science: Reflecting on Rein Vihalemm's Ideas,* Lanham: Lexington Books.

Niiniluoto, Ilkka (1986): 'Pragmatismi', in Niiniluoto, I. and Saarinen, E. (eds.), *Vuosisatamme filosofia,* Helsinki: WSOY, 40–73.

Niiniluoto, Ilkka (1999): C*ritical Scientific Realism,* Oxford: Oxford University Press.

Niiniluoto, Ilkka (2019): 'Queries about Pragmatic Realism', in Rydenfelt, H., Koskinen, H. J. & Bergman, M. (eds.), *Limits of Pragmatism and Challenges of Theodicy: Essays in Honour of Sami Pihlström,* Acta Philosophica Fennica 95, Helsinki: Societas Philosophica Fennica, 31–43.

Niiniluoto, Ilkka (2024): 'Abductive Arguments for Ontological Realism', in Angelucci, A. et al. (eds.), *Realism and Antirealism in Metaphysics, Science and Language: Festschrift for Mario Alai,* Milano: Franco Angeli, 41–50.

Pihlström, Sami (1996): *Structuring the World: The Issue of Realism and the Nature of Ontological Problems in Classical and Contemporary Pragmatism*, Acta Philosophica Fennica 59, Societas Philosophica Fennica, Helsinki.

Pihlström, Sami (2022): *Toward a Pragmatist Philosophy of the Humanities,* Albany: State University of New York Press.

Putnam, Hilary (1981): *Reason, Truth and History,* Cambridge: Cambridge University Press.

Raatikainen, Panu (2004a): *Ihmistieteet ja filosofia,* Helsinki: Gaudeamus.

Raatikainen, Panu (2004b): 'Conceptions of Truth in Intuitionism', *History and Philosophy of Logic* 25(2): 131-145. URL = https://doi.org/10.1080/014453401625669

Raatikainen, Panu (2014): 'Realism: Metaphysical, Scientific, and Semantic', in Kenneth R.Westphal (ed.), *Realism, Science, and Pragmatism,* London: Routledge, 139-158.

Raatikainen, Panu (2023): 'Varieties of Ideal Language Philosophy', in Panu Raatikainen (ed.), *Essays in the Philosophy of Language,* Acta Philosophica Fennica 100, Helsinki: Societas Philosophica Fennica, 23-53.

17
Defining realism in social ontology

Arto Laitinen & David P. Schweikard

Introduction

Social ontology studies first-order metaphysical questions about social reality. It studies, for example, the nature and existence of group agents, of social kinds related to race, gender, class and disability, of institutions like money or marriage, of organisations like FIFA or the United Nations, and more generally it inquires into the nature of social facts, properties, and relations. Here are some good questions social ontologists ask about entities in any of these categories: Are the entities irreducible to their constituent parts? Are they grounded in something more fundamental? Are they somehow dependent on human minds - are they constructed, conferred, projected? Are they ultimately eliminable? Does talking about them actually refer to anything? Should we take them into account in providing causal explanations, in seeking normative guidance in the social world, or when we engage in social criticism?

In this paper, we are not concerned with debates on these substantive first-order questions. We are interested in the second-order question as to *what counts as realism* in social ontology. Our aim is to make progress with regard to how realism in social ontology should be defined and what defending it requires.

Rival definitions of ‘realism’ are more or less independent from substantive answers to different first-order questions. Two authors may agree in their answer to a first-order question - for instance, they may agree that social kinds like gender are

Laitinen, A. & Schweikard, D. P. (2025). Defining realism in social ontology. In A. Laitinen, M. Keinänen, J. Reinikainen & A. Honkasalo (eds.). *Language, Truth, and Reality*. Tampere University Press. (pp. 196–226).
https://doi.org/10.61201/h5fbdk64

mind-dependent - but disagree on whether this commitment marks their view as realist (or in this case, anti-realist). Conversely, two authors may agree which second-order question targets the mark of realism - for instance, that realism about group agents consists in claiming that they are irreducible to their individual members - but disagree as to what is the correct answer to the respective first-order question whether group agents are reducible or irreducible (cf. the section "Two debates about realism," below).

Our overall aim then is to map and assess rival usages of 'realism' in social ontology. We set in by proposing definitions of a variety of realisms (and, by implication, anti-realisms) in social ontology, and thereby seek to structure present debates. We first introduce four guiding questions which we use to expound a new systematic way of mapping four kinds of realism and four corresponding kinds of anti-realism. (Cf. the section "A basic map of realisms and anti-realisms," below)

We then turn to argue which of the four proposed definitions of realism would be most appropriate in social ontology. We first discuss irreducibility or non-redundancy as a definition of realism, making three points. We point out that the distinction between 'really real' and 'real' may be hard to defend; that fundamentality as a definition of realism would consider all things social to be unreal; and that while causal and normative relevance are the main motivation to argue that things can be mind-dependent but real, independent arguments would be needed for the case that that causally inert or epiphenomenal entities do not exist at all, or do not 'really exist' at all.[1] Thus, we argue that while X's causal efficacy is a good reason to believe that X is real, it need not be a special definition of what it is to be real. Then, in the next section, we take a closer look at variations, degrees and kinds of mind-independence. We shall suggest that some forms of mind-dependence are more clearly anti-realist than others. Subsequently, we shift the focus to the more minimal realism characterised as cognitivist success theory, and argue that it is the best definition of realism for social ontology. (The second best is non-redundancy understood as causal and normative relevance, but it is better conceived as a reason to believe something is real, rather than a definition of what it is to be real). Mere cognitivism seems too minimal. (Cf. the subsections under "contested issues" below). The final section presents the conclusions.

[1] Causality is often regarded as a mark of being real. If something makes a causal difference in the world, we have reason to think it is real. Arguably, normativity can be regarded as analogous to causality in this respect: if something makes a normative difference, we have reason to think it is real. For example, there is real oppression. The analogy does not do central work in this paper, nor do we argue for it, but we occasionally mention normativity side by side with causality to draw attention to the possible analogy. We thank an anonymous referee for a comment on this.

Two debates about realism

The aim of this section is to highlight the difference between two kinds of disagreements: one kind of disagreement or debate concerns substantive questions in social ontology (say, is race something socially constructed?) and the other kind of disagreement concerns the meta-debate about realism (say, is social constructivism about race a form of anti-realism?). The substantive debates, as we elaborate in the first subsection, target first-order issues concerning the nature and status of social entities. The meta-debate to which we turn in the second subsection and the rest of the paper is about what warrants labelling a view taken within a substantive debate as 'realist.'

Some substantive debates in social ontology

Before turning to rival definitions of realism, it will be helpful to demonstrate that the first-order questions cited in popular definitions of realism are genuine questions at the heart of substantive social ontological debates. In outlining four such debates, we shall use 'race' and 'group beliefs' as examples.

(1) Views that would be analogous to instrumentalism in philosophy of science or non-cognitivism in metaethics, would not take race-talk or group-belief-talk at face value. They would hold that 'even if race is an illusion, racial discourse might serve some important interests.' (Glasgow 2008, 11).[2] Most emancipatory race activists agree that literally there are no biological races, but race-talk is nonetheless called for by the aim of correcting past injustices. In the same vein, Daniel Dennett (1987), Raimo Tuomela (2013) or Deborah Tollefsen (2015) may think it is possible and for normative or explanatory purposes important to relate to groups as if they are believers, while thinking that they are not literally agents or subjects of intentional attitudes. In general, instrumentalism or non-cognitivism would hold that the point of race-talk or group-belief-talk is not to describe reality: it may be useful for certain purposes, but talking in these ways is merely instrumental and does not commit one to thinking there are races, or group beliefs. It may be useful for addressing existing oppression or for holding collectives responsible, but on this kind of view the discourse is not to be taken literally. It is taken to be metaphysically non-committal because the respective language is to be interpreted non-cognitively or instrumentally. In physics, one may regard talk about 'centre of gravity', or 'holes', or 'quantum strings' as instrumentally useful even while thinking such theoretical entities do not exist. And in social science, one can hold that socio-economic structures are similarly only theoretical entities. It is fair to say that

[2] Pace Charles Mills (1998, 49), who seems to think there are no non-cognitivists about race.

such instrumentalist or non-cognitivist views are widespread in various branches of philosophy, in philosophy of science including the philosophy of the social sciences as well as in metaethics. And it is a position that can be taken in metaphysics and social ontology, too, albeit one that assigns only a secondary role to metaphysical commitments. Unsurprisingly, many do not take this line but hold that race-talk or group-belief-talk - or the talk of holes or social structures - is to be taken literally, i.e. they advocate non-instrumentalism or cognitivism.

(2) But even if one adopts a non-instrumentalist or cognitivist view, there is room for further debate. Understanding the respective discourses in this way does not answer the questions whether there *are* races or group beliefs. It is possible to hold that race-talk is literal and descriptive but that there are no races, and likewise for group-belief-talk and group beliefs. This amounts to eliminativism concerning races and group beliefs, and to error theory concerning the talk or theorising about them. As Haslanger (2012, 198) notes, error theories about race are common.[3] Similarly, a garden variety individualist would hold that there are no group minds and *a fortiori* no group beliefs.[4] Whereas non-cognitivism need not advocate the abolishing of all talk about race or group belief, eliminativism is likely to come with that advice (see Glasgow 2008, 114). Eliminativists may argue that there have been inappropriately racialised groups, and there may be stringent duties of restorative justice towards them, but that it is a massive mistake to think that there are races. Behind all race-talk (or talk about group beliefs), they may proceed, lies a systematic erroneous presupposition that is to be abolished. Again, this is a reasonable sort of view, and almost everyone is likely to be an eliminativist about something, if not about races or group beliefs, then at least about phlogiston or witches in the discourses in which they are claimed to exist.

(3) Another debate concerns the mind-dependent, language-dependent, constructed, projected, or conferred status of the entities under discussion. Perhaps there are races (or racialised groups), genders (as opposed to sexes), disabilities (as distinguished from physical impairments), as well as group beliefs, plans and policies, but they are social constructions. Variants of social constructionism may well be the dominant positions regarding many issues in social ontology. They oppose objectivisms or non-constructionisms or naturalisms of different

[3] See, for instance, Appiah (1996) and Zack (2002), and Glasgow (2008) on Haslanger's account.

[4] Whereas defenders of group beliefs such as Rovane (1998) or Pettit (2003) would argue in the other direction that as groups are able to form beliefs and act, they have 'minds' in some restricted sense while not having phenomenal, experiential minds.

kinds. Social constructionists stress that qua social constructions races, genders, and disabilities differ from related natural or non-constructed phenomena such as having certain ancestry, or having certain female or male characteristics, or having bodily impairments.[5] With respect to questions about group beliefs, institutions, or organisations, it may be less tempting to claim that they are not social constructs in some sense (and concerning them, social constructionism may seem uncontroversial), but arguably one could hold that, say, 'marriage' refers to certain patterns or functions,[6] or that 'group belief' refers to certain dispositions of the individuals in a population etc. that it is the task of social sciences to find out.[7] Even if no-one regarded institutions or group beliefs as independent of social construction, it would not follow that social constructionism is an uncontroversial position: instrumentalists and non-cognitivists, and eliminativists and error theorists would oppose either the cognitivism or the non-eliminativism it entails. Yet, social constructionism is clearly a wide-spread view, or family of views, concerning many questions in social ontology, and there are many important further substantive questions on which there are family disputes within this approach.

(4) Consider finally the set of questions surrounding reducibility, fundamentality, groundedness, or the possible emergent or *sui generis*-nature of social entities. These questions concern the dependence of social or institutional entities not so much on minds or observers or conferrers, but on their constituent parts, or on more fundamental layers of reality such as the natural, the physical, or the material, or (in the cases of groups) the individuals that the group in some sense consists of. Reductionist views hold that there are group beliefs, but that they are reducible to individuals' attitudes, whereas non-reductionists treat them as irreducible. Whether or not group beliefs, or social properties and social entities more broadly, are reducible is a matter of substantive debate between individualists and their opponents in philosophy of social science.[8] And similar debates about irreducibility are ubiquitous in metaphysics, where they are closely related to debates about grounding, fundamentality, and dependence.[9]

[5] Constructionists typically advance substantive arguments against views which claim that races, gender, disabilities are equally non-constructed phenomena as ancestries, genitalia or impairments. (For example, Ásta (2018) holds that the social properties like being disabled are conferred, and the conferrers try to track the base properties like having impairments).

[6] Guala 2016.

[7] See e.g. Thomasson 2019 for discussion.

[8] See, e.g. Zahle & Collin (eds.) 2014; Ylikoski (2017).

[9] See, e.g. Schaffer (2009), Sider (2011), Barnes (2014), Tahko (2015), Mikkola (2017).

The consensus in social ontology (apart from possible radical social constructionists) is that the entirety of social reality, if it exists at all, is non-fundamental: social reality presupposes the existence of social animals, and even they are hardly included in the inventory of fundamental entities, for their existence is dependent (in some way) on something more fundamental.[10] The fundamental entities are not ontologically dependent. Yet, non-fundamental entities may have features that are not fully reducible to the fundamental entities but are for example emergent. Ontological dependence need not entail full reducibility, and social entities may be non-fundamental and yet irreducible.

Regarding each of these issues, there are different substantive metaphysical, first-order debates.[11] The related meta-debate, to which we now turn, concerns which of the positions adopted on those issues count as 'realist' and why. This meta-debate is in dire need of mapping and clarity.

The meta-debate: rival definitions of realism

We can introduce the meta-debate about conceptions of realism with the help of these sets of substantive questions. Independently of which of the theories are true (in some domain), the second-order question is what it takes to be a realist (in that domain). What is the role of cognitivism, non-eliminativism, mind-independence, or irreducibility, in defining realism? We next seek to demonstrate that quite different definitions of realism have been proposed in the literature, but they have not always been clearly distinguished from one another where scholars have often understood their proposal as the only (and sometimes obviously correct) definition of realism.

According to the received view of metaphysical realism only mind-independent things are real.[12] This definition of realism is used in social ontology as well: 'A 'racial realist' ... will be somebody who thinks it is objectively the case, - *independent of human belief* - that there are natural human races; in other words, that races are natural kinds' (Mills 1998, 46; italics added). On this definition, socially constructed entities and properties are not real and social constructionism (of any kind) is an anti-realist view. Instead of accepting this verdict, it has been argued forcefully that this definition of realism, standard as it may be in general metaphysics, will not do for social ontology.

At any rate, realism in social ontology faces a special tension. On the one hand, ordinary people in their everyday lives normally take for granted that social structures and entities are in some relevant sense something real. Institutions, practices, behavioural expectations, racialised and gendered oppression are

[10] We thank an anonymous referee for urging us to note the possibility of radical social constructionism.

[11] This list of substantive questions is naturally far from exhaustive. In particular, it doesn't include systematic examination regarding particular modes in which entities might exist. In setting this aside for now, we attend only to discussions about *whether* an entity (or kind of entities) exist, and not to those about *how* they exist, with the exception of mind-(in)dependence and (ir)reducibility; as these have figured in debates on how to define realism.

[12] See, e.g. Devitt (1984), Thomasson (2003), Barnes (2017), Haukioja (2021), Khlentzos (2021), and Miller (2022). Realism in this sense is traditionally opposed to 'idealism' and 'phenomenalism' (cf. Raatikainen 2014).

something we really do encounter - they are not mere figments of imagination. They often are 'all too real,' as Haslanger puts it (2012, 5). Moreover, they are causally and normatively effective: they cause events and generate oughts. On the other hand, these institutions, expectations and wrongs seem equally obviously to be somewhat dependent on human minds, actions, and practices, which suggests that they aren't real after all or that they are less real than the mind-independent, more fundamental natural facts.[13] However, there are rival definitions available.

Some scholars have argued that there are other criteria than the commitment to mind-independence for a view to count as realist, and these are related to the terms of debates (1) and (2) we outlined in the previous subsection. For example, Haslanger and Sayre-McCord suggest the following:

> 'A realist about a domain D maintains that claims purporting to describe D are truth-apt, that is, the claims are the sort of thing to be either true or false, and at least some of them are true' (Haslanger 2012, 198).
>
> 'Realism involves embracing just two theses: (1) the claims in question, when literally construed, are literally true or false (cognitivism), and (2) some are literally true.' (Sayre-McCord 1986, 2).

Thus, one question is whether the relevant claims have truth-value. Another question is whether any of the claims in that discourse are true.[14] In any domain, including the social and institutional world, there are correspondingly two ways of being anti-realist, that is by subscribing to non-cognitivism or instrumentalism, or by adopting an error-theory or eliminativism,[15] whereas realism would entail cognitivism and non-eliminativism.

These two questions can be seen as providing a relatively minimal answer to what realism is, so that we arrive at realism before even asking about mind-independence. This is one sense in which one can be a realist about mind-dependent entities.[16]

[13] See Fine (2001) for a suggestion that all realism versus anti-realism debates face something like this tension.

[14] See Devitt (2010) for an argument (against e.g. Sayre-McCord) that realism should be defined independently of questions of truth and merely focus on the existence and independence – conditions; more or less our criteria 2 and 3.

[15] Haslanger (2012, 198) notes that plausible anti-realism in the social domain can take especially 'the form of error theories. Error theories about race are common (e.g., Appiah 1996; Zack 2002; Glasgow 2008). On this view, because there are no races, statements involving racial terms, although they purport to be true or false, are all false, since racial terms do not refer.' (Haslanger 2012, 198). See also Mills (1998).

[16] Sayre-McCord is explicit that 'Realism is not solely the prerogative of objectivists [defenders of a mind-independence view].' (1986, 12). Joyce (2016) holds that Sayre-McCord does not give sufficient reason to exclude mind-independence as a criterion of realism. 'Perhaps all that is needed is a more careful understanding of the type of independence relation in question.' For a distinction between four kinds of mind-independence, see Page (2006). For a recent suggestion that what matters is not mind-dependence but mind-groundedness, see Cohen (2022).

But an even more minimal definition of realism is available, on which even an error theory or eliminativism count as realist, and on which all that is needed for a view to be realist is a commitment to non-instrumentalism or cognitivism.

For Dummett, a sufficient mark of realism is

> '. . . the belief that statements of the disputed class possess an objective truth-value, independently of our means of knowing it: they are true or false in virtue of a reality existing independently of us.' (Dummett 1978, 146)

In a similar vein, Putnam does not invoke non-eliminativism as a requirement either, though he does suggest linking up cognitivism and mind-independence:

> 'A realist (with respect to a given theory or discourse) holds that (1) the sentences of that theory are true or false; and (2) that what makes them true or false is something external-that is to say, it is not (in general) our sense data, actual or potential, or the structure of our minds, or our language, etc.' (Putnam 1979, 69-70)[17]

Further along on the spectrum spun up above, many theorists think there is more to the question of realism in social ontology than the three questions of cognitivism, non-eliminativism, and mind-independence.[18] There is a fourth question which distinguishes between some *Xs* that admittedly may exist in some weak or redundant sense (like shadows or holes, or heat, or random aggregates of individuals), and something else, *Ys* and *Zs,* that are 'really real'. The non-redundancy can be cashed out with the help of either of two related distinctions, ontological irreducibility or explanatory relevance. The first is a distinction between *Xs* that are reducible to (they are 'nothing but') something else, *Ys,* on the one hand, and those more fundamental *Ys,* that ground or constitute or help compose the reducible *Xs*; or also things that are otherwise ontologically irreducible (*Zs*) (while perhaps not grounding anything else), on the other.[19] The second is a distinction between things, *Ys,* that are non-redundant in making an explanatory difference, causally or normatively, on the one

[17] Cf. also Uskali Mäki (2005, 231), who defends different definitions of realism for different domains, suggests that on some domains at least cognitivism is enough (while in other domains we may need a success-theory): 'It is sufficient for a realist to give the existence of an entity (and the truth of a theory) a chance, while in some areas we may be in a position to make justified claims about actual existence (and truth).'

[18] See Baker (2007), Barnes (2014 and 2017), Enoch (2017), Fine (2001), Haslanger (2012), Himmelreich (2019), Hindriks (2006), List and Pettit (2011), Mikkola (2018), Mäki (2008), Pettit (2009), Schaffer (2009 and 2017), Sider (2011), and Taylor (1989).

[19] See e.g. Baker (2007) for a view that links realism and non-reductionism. Similarly, Pettit 2009 and List & Pettit 2011 ask whether group-level-talk is 'readily reducible' to individual-level talk. If yes, we have 'thin, redundant realism', but if not, 'non-redundant realism'. See Section 4.1 for the distinction between existing in some thin sense or 'really existing' (Fine 2001, Himmelreich 2019).

hand, and things, *Xs,* that are inert, epiphenomenal, or merely idle wheels, on the other hand.[20]

A group agent, for example, is real in this more demanding sense if it cannot be metaphysically reduced to its constituents and so is irreducibly real, or if it (rather than its component parts) is causally efficacious and non-redundant. But there is also room for debate on how best to understand this question of non-redundancy, whether as metaphysical fundamentality or as causal or normative relevance.[21] On the latter reading something is real even if it is not fundamental (and even if it is mind-dependent).

This malaise of definitions of realism has not gone unnoticed. One response to the observation that, say, mind-dependence or fundamentality are more appropriate characteristics of what is real in some domains than in others would be to adopt *local realisms*. For what being real amounts to may differ from one domain of reality to another.[22] In physics and chemistry, entities' being real may well consist in them being mind-independent. In social ontology, however, we may need to refer to a different definition of realism to make sense of the idea that social reality is indeed social *reality.*

As is standardly understood, commitment to realism is always commitment to realism-about-some-domain, where one can be a realist in one domain (say, physics) and an anti-realist in another domain (say, ethics). And within any domain, one can be a realist concerning *Xs* (say, quarks) and an anti-realist concerning *Ys* (say, phlogiston). (see e.g. Miller 2021, Miller 2022, Mäki 2008). Within social ontology, one can coherently be a non-cognitivist concerning, say, hurray-talk of football fans, an eliminativist concerning group minds, a reductionist concerning group agency, and yet a non-reductionist concerning groups per se. What interests us is what makes one

[20] As Barnes writes, some 'metaphysical realists go further. They think that among the things that exist, some things are more explanatorily important than others. Maybe it's true that both holes and electrons exist. Nevertheless, electrons are explanatorily more significant than holes. Metaphysicians trying to give a good theory of the world should care about electrons more than they care about holes.' (Barnes 2017, 2418).

[21] Barnes and Mikkola illustrate this with reference to Haslanger and Sider. Barnes adds that e.g. on Sally Haslanger's view '*social categories* are among the most explanatorily important things that there are.' (Barnes 2017, 2418). 'Theodore Sider, in his *Writing the Book of the World*, gives the perhaps the most detailed defense of metaphysical realism in the contemporary literature—one that attempts explain both what such realism consists in and how such realism can lay the groundwork for distinguishing between 'substantive' and 'non-substantive' ('shallow', 'terminological') disputes' (Barnes 2017, 2425). 'on Sider's construal of ontological realism, Haslanger is not an ontological realist about social kinds. To me, this result suggests that Sider's construal of ontological realism is impoverished, rather than that Haslanger is misdescribing her view.' (Barnes 2017, 2430). Mikkola (2017, 2442) notes that Sider and Haslanger 'disagree on what counts as reality. For mainstream metaphysicians like Sider (and not all metaphysicians agree), only that which is fundamental does; for feminists like Haslanger, whatever has causal efficacy counts as real.' Barnes sides with Haslanger; ''The debate over gender realism isn't a debate about how/whether genders are grounded. It's a debate about what (if anything) they do, and what (if anything) they explain.' (Barnes 2017, 2433) What Barnes's, Haslanger's and Mikkola's views have in common with Sider, however, is the attempt to distinguish between redundant and non-redundant existences. Their debate is on whether this is to be cashed out with reference to fundamentality, or to explanatory relevance.

[22] Mäki (2008).

a ‘realist’ or ‘anti-realist’ in social ontology: is, say, rejection of non-cognitivism, or of eliminativism, sufficient for being a realist?

In the remainder of this paper, we map out and discuss four candidates for such local realisms in social ontology and defend one of them. Yet, each candidate for a monist local realism (one for social ontology, one for mathematics, one for physics, etc.) is at the same time one constituent of a pluralist global realism (several definitions of realism in each of the domains). A global pluralist map of different definitions would be relevant for potential debates in any domain (cf. Sayre-McCord 1986, Haslanger 2012, and Fine 2001). This response has the advantage of not relying on a prior metaphysics of domains, and of treating mind-independence as one, but only one, criterion of realism. We shall adopt this approach in the following.

A basic map of realisms and anti-realisms

We now proceed to mapping out and defining a variety of views that may warrant the label 'realism.' Our proposal is based on four guiding questions, answers to which we take to define realist and anti-realist views. In the next two subsections we formulate the questions in terms of statements or sentences that can be truth-apt or true and are made true by social entities. In an ensuing, shorter subsection we briefly discuss how the questions could be reformulated in terms of entities that exist, are (ir)reducible or mind-(in)dependent, and in virtue of which the relevant sentences are true. No specific account of truth or truth-making is presupposed by the analysis.

Four Questions

Statements such as 'the bank closes at 4 pm,' 'the Kaizer Chiefs scored a goal,' 'the prize committee is a group agent,' 'there is racial discrimination,' and 'Alex is a woman' capture some of the central issues in social ontology. These statements are about institutions, group agents, and the social kinds of race and gender. As indicated above, we can ask a variety of questions about such statements (referred to as 'S' below). The following sequence of questions is apt to capture definitions of realism and anti-realism:

(Q1) Does S have a truth value?
(Q2) Is (a statement like) S ever true?
(Q3) Is S true in virtue of mind-independent facts?

In line with Sayre-McCord (1986) and Haslanger (2012), we take (Q1) to mark the distinction between *cognitivist* and *non-cognitivist* views. Cognitivists’ answer to (Q1) is 'yes,' non-cognitivists’ answer is 'no.'[23] Whereas this contrast is familiar from

[23] There are rival ways to characterise the question distinguishing cognitivism or representationalism from non-cognitivism or expressivism. Not only are there detailed hybrid positions such as fictionalism and quasi-realism which would call for a more sophisticated map (see Van Roojen 2015), but also the initial question

discussions in metaethics, some debates in philosophy of science feature a rather similar contrast to distinguish realist from instrumentalist views.[24] The latter hold that theoretical concepts (say, 'inflation', 'social structure', or 'male domination') are mere tools for organising our observations, and do not imply ontological commitment.[25] By contrast, realists in this sense claim these terms refer to the reality under investigation.

(Q2) in turn serves to distinguish between *error theories* that maintain that S-type statements are truth-apt but never true, and *success theories* according to which such statements are sometimes true. For success theory, mere cognitivism is not enough, as it holds that at least some of the statements in question are true. Eliminativism about race, for example, construes race-claims in a cognitivist fashion, but states that there are no races (see the discussion above). Realism that contrasts with eliminativism or error-theory takes at least some of the relevant claims to be true (Haslanger 2012, 198, Sayre-McCord 1986, 3).

(Q3) prompts leaving the binary schema. It invokes the contraposition between views that affirm mind-independence (often labelled as objectivism) and those that affirm mind-dependence (idealism, phenomenalism, subjectivism, intersubjectivism, constructionism). Many authors in social ontology reject this definition of realism in claiming that something *can* be both socially constructed and real.[26] On this view

> 'social structures are real - as real as anything - but they are *made*. They aren't 'joints in nature', they're joints in the social world. We created them, and our collective social activity is responsible for their continued existence, but they're no less real as a result.' (Barnes 2017, 2423).

Following Sayre-McCord (1986, 10ff.), we suggest distinguishing *objectivism* in social ontology from two variants of non-objectivism, namely *subjectivism* and

can be formulated without the notion of 'truth', leaving room for the possibility that for expressivists the sentences do have truth-value, on some suitably minimalist notion of truth (Dreier 2004). Especially in metaethics, the question can be posed in terms of moral language deriving its contents from the world, as representationalists would have it (from the truth-conditions of the statements, or from the moral properties referred to) or from our minds, as expressivists would have it (from the states of mind expressed in such statements). Similar shifts in the definition of non-cognitivism are possible and even foreseeable in social ontology, but here we stick to the current usage, e.g. in Haslanger (2012). Again, the main point is that, as we will see in due course, one of the four main variants of anti-realism in social ontology is non-cognitivist expressivism, and it is to be expected that there are many rival formulations and hybrids possible concerning this question, like with other questions. We thank Teemu Toppinen for comments on this.

[24] The heyday of instrumentalism was in the first half of the 20th century, cf., however, a more recent proposal in (Rowbottom 2011).

[25] Contemporary scientific realism emerged largely in opposition to 'instrumentalism', the view that 'it is not possible to eliminate theoretical concepts from science, or define them in terms of observational concepts, but these theoretical concepts do not refer to anything real; they are only practically useful fictions which enable one to systematise observations and predict new observations on the basis of old ones.' (Raatikainen 2014, 5)

[26] See Barnes (2017), Haslanger (2012), Mills (1998), and Mason (2020).

intersubjectivism. Objectivists hold that the truth of S-type statements is borne out by facts that are objective in the sense that they are altogether independent of attitudes or practices. Intersubjectivists, by contrast, hold that S-type statements are true in virtue of a particular social practice or attitudes shared within a particular community. And subjectivists hold that S-type statements are true in virtue of facts about a particular subject, e.g. the one uttering S. Only objectivism thus characterised claims that S-type statements are true in virtue of mind-independent facts, whereas subjectivists and intersubjectivists embrace mind-dependence.[27]

To do justice to the social ontological debate about reducibility, we suggest including a fourth question:

(Q4) Is S about some irreducible entity?

This question reflects the longstanding concern with fundamentality debated in metametaphysics. In relation to questions of existence and reality, the issue is whether only fundamental entities exist and are real, where being fundamental can be understood as being irreducible. The general issue here is whether certain statements refer to irreducible entities or whether they ultimately refer only to their constituents or grounds (see e.g. Mäki 2008, Pettit 2009, List & Pettit 2011, Fine 2001, Schaffer 2009). Kit Fine (2001, 27) suggests 'a general presumption in favour of the grounded not being real.' This view holds, to use Fine's example, that if a war between nations is grounded in military activity of their citizens, then the citizens and their activity are real, but the nations and war are merely apparent, unreal or grounded, not part of fundamental reality. Lynne Rudder Baker (2007, 3) seems to agree with Fine and others that the issue of reducibility is relevant to realism, but she disagrees concerning what is reducible: 'The aim of *The Metaphysics of Everyday Life* is to present a theory that focuses on the familiar objects that we encounter every day – flowers, people, houses, and so on – and locates them irreducibly in reality.' (Baker 2007, 3).

In social ontology, a familiar context for asking (Q4) is provided by statements referring to social groups or group agents: Are statements about social groups or group agents reducible to statements about their members or are they about some irreducible entity? With regard to social groups in general the question becomes salient whenever S ascribes a property to a group. With regard to group agents in particular the issue arises whenever a belief, an intention or an action is ascribed to a group. In line with the received terminology, an affirmative answer to (Q4) makes one a *non-reductionist*, a negative answer makes one a *reductionist*. Answering (Q4) either way is independent of how one answers (Q3). We thus need to make room for

[27] In Mills' (1998, 46-49) terminology, something mind-dependent but existing and causally efficacious is 'objective' but not 'real'; we follow Sayre-McCord in calling it 'real' but 'not objective'. When mapping rival usages, it is important to note that despite the terminological difference with respect to 'objectivism' and 'realism', Mills draws the same conceptual distinctions as Sayre-McCord (1986) and Haslanger (2012).

non-reductionist and reductionist variants of objectivism, intersubjectivism, and subjectivism.

Taken together, the main thrust of these questions is metaphysical: Do *Xs* exist, are *Xs* ontologically dependent on minds, and are *Xs* reducible to some *Ys*? Our focus here is on metaphysical realism in social ontology. We have formulated these metaphysical questions semantically in terms of statements S about *Xs* and *Ys* to link them to the prior question of how to analyse the statements. According to this framework, one way to fail to be a metaphysical realist about *Xs* is to think that the statements about *Xs* are not truth-apt.[28]

The distinctions introduced so far yield the following map which we will use to individuate different types of realism and anti-realism in the following subsection (see figure 17.1).

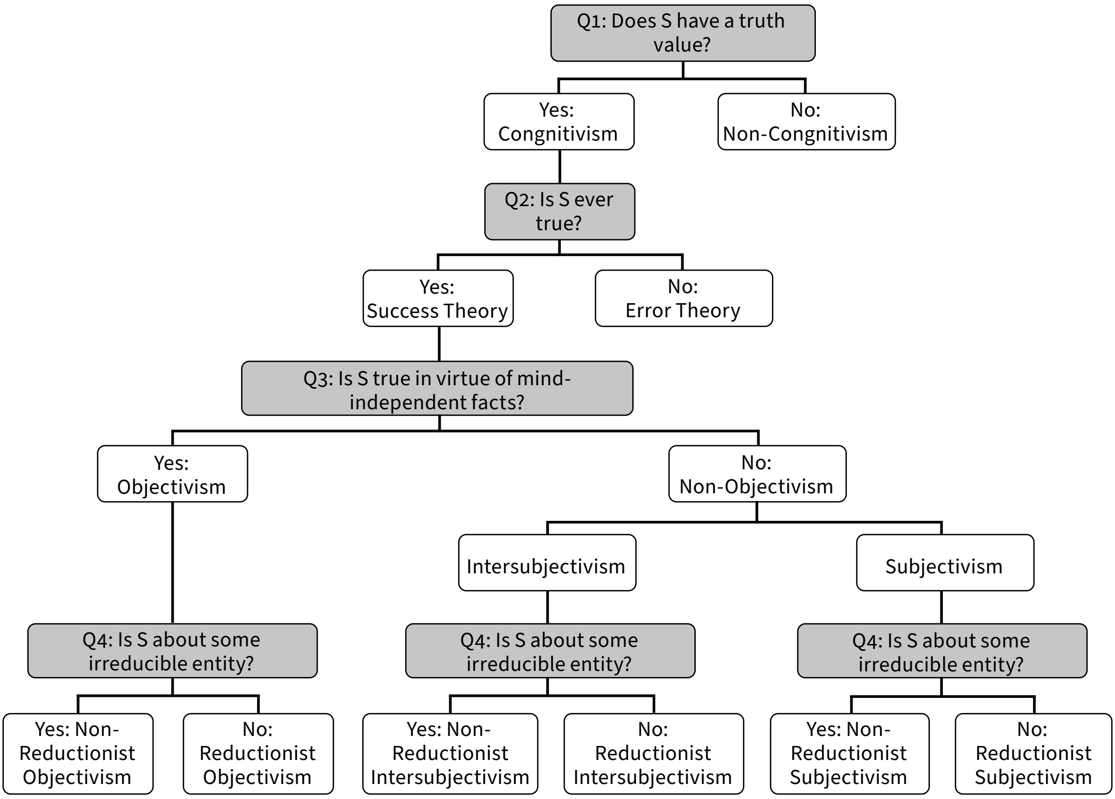

Q1: Does S have a truth value?
Yes: Congnitivism
No: Non-Congnitivism
Q2: Is S ever true?
Yes: Success Theory
No: Error Theory
Q3: Is S true in virtue of mind-independent facts?
Yes: Objectivism
No: Non-Objectivism
Intersubjectivism
Subjectivism
Q4: Is S about some irreducible entity?
Q4: Is S about some irreducible entity?
Q4: Is S about some irreducible entity?
Yes: Non-Reductionist Objectivism
No: Reductionist Objectivism
Yes: Non-Reductionist Intersubjectivism
No: Reductionist Intersubjectivism
Yes: Non-Reductionist Subjectivism
No: Reductionist Subjectivism

Figure 17.1: Basic map of views in social ontology

[28] For a more detailed analysis of the interplay between scientific, metaphysical, and semantic realism cf. Raatikainen (2014).

Four Realisms and Four Anti-Realisms

Let's begin with the realisms. Using the order and numbering of the questions (Q1) through (Q4), we can list and label the following four types of realism by giving their respective answers to those questions as identifying commitments regarding statement S:

(R1) *most minimal realism*: 'S has truth value.'
(R2) *minimal realism*: 'S is (sometimes) true.'
(R3) *objectivist realism*: 'S is true in virtue of mind-independent facts.'
(R4) *non-reductionist realism*: 'S is about some irreducible entity.'

We do not intend to suggest here that any of these positions can count as a comprehensive view. But differentiating between them facilitates identifying a variety of realist commitments. Especially when read in this order it becomes clear that some of these realisms build on others in the sense that those with higher numbers imply commitments to those with lower numbers; thus minimal realism implies a commitment to most minimal realism, and objectivist realism implies a commitment to both most minimal realism and minimal realism.

Realism can be defined in minimal fashion so that one is a realist in the sense of (R1) if one believes that a statement of the relevant kind can be true (whether or not it is). The attitude Uskali Mäki (2008, 340) calls 'a realist attitude' implies that 'there is a fact of the matter concerning whether or not X exists and whether or not [S] is true. It is an attitude that will give real existence and objective truth a chance, but one that at the same time is prepared for concluding that X does not exist or [S] is not true, after all.' In this sense, one can be a realist about phlogiston and deny that there is any.

The view here labelled as 'minimal realism' (R2) is demarcated by the thought that error-theory, eliminativism or nihilism are forms of anti-realism. As illustrated with regard to Haslanger's and Sayre-McCord's frameworks one must thus meet the two conditions that define a success theory to be a minimal realist.

A more demanding form of realism is objectivist realism (R3), adoption of which requires subscribing to the idea that the statements in question are true in virtue of something mind-independent. This definition is congruent with Mallon's (2016, 138 ff.) characterisation of 'basic realism' as views that fulfil three requirements (literalness, success, and objectivity), i.e. give affirmative answers to our first three questions. The definition is akin to one of the elements Thomasson (2003, 580) identifies as belonging to the 'realist philosophical world-view'. According to this 'ontological view [...] there are kinds of things that exist and have their nature independently of human beliefs, representations, and practices' (ibid.). Sayre-McCord (1986, 11) polemicises against treating (R3) as the only form of realism and deems all three variants (objectivism, intersubjectivism, subjectivism) 'quite clearly' realist.

With regard to (R4) it needs to be highlighted that this, too, picks out a variety of realist views. As mentioned before, many take irreducibility or fundamentality respectively to be crucial to being real. On this view, one is not a realist about *Xs* if one thinks *Xs* are reducible to something else. Mäki (2008, 335) refers to reductionists as anti-realists and thus treats a commitment to irreducibility as a mark of realism. Similarly, Pettit's (2009) argument for the reality of group agency can be read as targeting the question of reduction, even though it is couched in the debate about whether certain groups can be real agents. Responses to the associated question (Q4) do not, however, build on responses to the other questions, and thus (R4) does not build on other forms of realism in the way, say, (R3) builds on (R2). Correspondingly, our map indicates that one can adopt a reductionist or a non-reductionist account irrespective of one's commitment to either objectivism, intersubjectivism, or subjectivism. Only for the view on which statements about social entities are true in virtue of irreducible objective facts is it the case that all the mentioned realisms build on each other.[29]

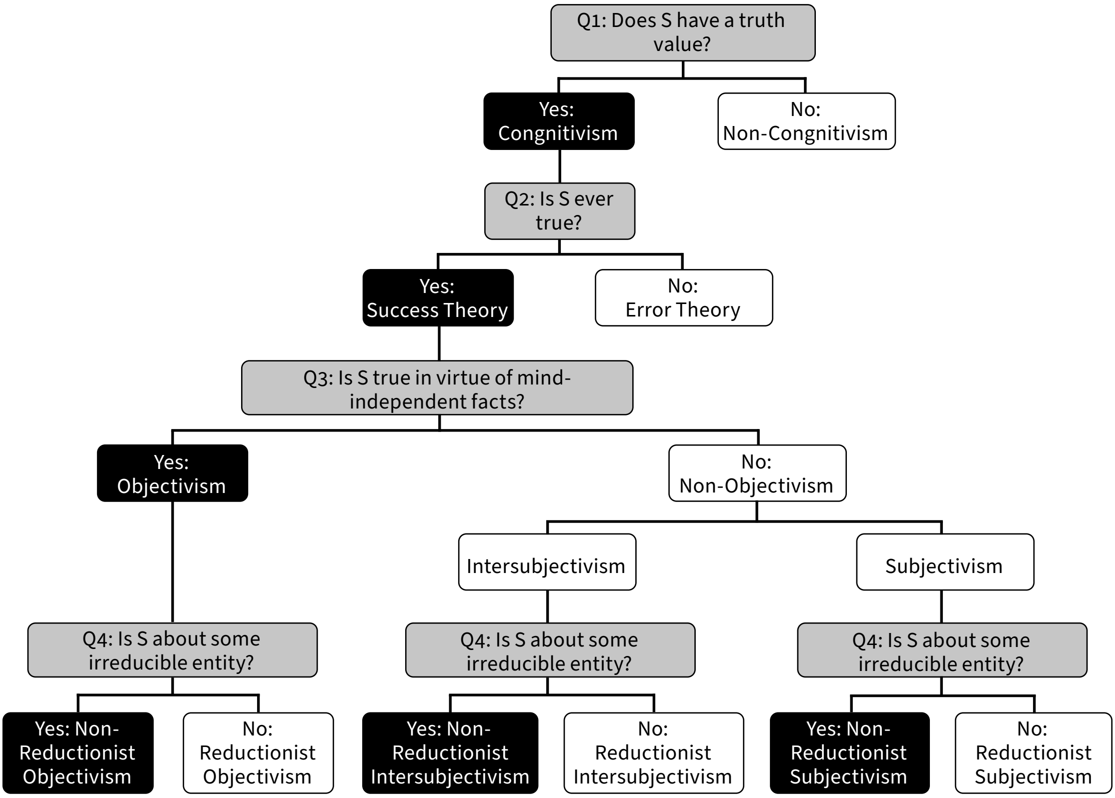

Q1: Does S have a truth value?
Yes: Congnitivism
No: Non-Congnitivism
Q2: Is S ever true?
Yes: Success Theory
No: Error Theory
Q3: Is S true in virtue of mind-independent facts?
Yes: Objectivism
No: Non-Objectivism
Intersubjectivism
Subjectivism
Q4: Is S about some irreducible entity?
Q4: Is S about some irreducible entity?
Q4: Is S about some irreducible entity?
Yes: Non-Reductionist Objectivism
No: Reductionist Objectivism
Yes: Non-Reductionist Intersubjectivism
No: Reductionist Intersubjectivism
Yes: Non-Reductionist Subjectivism
No: Reductionist Subjectivism

Figure 17.2: Realisms in social ontology

[29] It is important to note that the views in question here needn't rely on a single feature such as those picked out by (R1)-(R4) in defining realism. There are a number of multi-feature views which are worth discussing; such as those that combine (R2) and (R3) (e.g. Haukioja 2021), (R2) and (R4) (e.g. Pettit 2009), or (R1) and (R4) (e.g. Mäki 2008). We set these more complex views to the side for the time being and focus on the tenability of their components.

Now for the anti-realisms. Juxtaposed to (R1) through (R4) above, they can be labelled and formulated as follows:

(AR1) *non-cognitivism*: 'S does not have truth value.'
(AR2) *error theory*: 'S is never true (always false).'
(AR3) *the mind-dependence view*: 'S is true in virtue of mind-dependent facts.'
(AR4) *reductionism*: 'S is about some reducible entity.'

The contours of (AR1) and (AR2) are relatively clear. Defenders of (AR1) are non-cognitivists about the kind of statement in question. A detailed analysis of non-cognitivist accounts in social ontology - which we are not undertaking here - would have to pay close attention to the question whether at least some variant of what is known as constructionism in social ontology could be given, or is indeed taken to have, the form of 'collectivist expressivism.'[30] Defenders of (AR2), on the other hand, embrace an error theory about the kind of statement in question, i.e. they hold that although truth-apt the statements in question are never actually true as there is nothing that makes them true.

(AR3) is the 'mind-dependence view' according to which statements about social entities are truth-apt, sometimes true, and true in virtue of mind-dependent facts. Given the threefold distinction we have invoked, the options here are subjectivism and intersubjectivism, depending on whether the mind-dependent facts in virtue of which some statements are true are taken to be facts about individuals or facts about collectives or communal practices. Both of these views reject the objectivist response to (Q3) and can thus be called 'non-objectivist.' If (Q3) is understood as targeting the question of realism in terms of mind-independence, then such non-objectivist views are *ipso facto* anti-realist. Below we will suggest rejecting the criterion of mind-independence as a marker of realism. There are, we will argue, good reasons to label at least some non-objectivist views 'realist.'

Given how (Q4), the question concerning reducibility, departs from the cascading provided by (Q1)-(Q3), it is possible to combine (AR4) with any view regarding the truth-aptness of S-type statements and regarding in virtue of what they are true. If a commitment to irreducibility is taken to be the mark of realism, then a view that subscribes to (R3) and (AR4) - call it 'reductionist objectivism' or, more precisely, 'reductionist objectivist realism' - would count as a form of anti-realism. It is, however, debatable whether responses to (Q4) mark off realist and anti-realist views in a satisfactory manner, not least because the issue of reducibility is of considerable complexity. We shall return to this below, in the subsection on reducibility.

[30] Interestingly, we can ask a question analogous to (Q3) regarding the non-cognitivist branch as well. S-type statements could express individuals' attitudes (subjectivism) or collective attitudes (intersubjectivism). The latter, 'collective expressivism' is an underexplored possibility in social ontology, but perhaps something like Sellarsian (1968) views about collective intentionality tied to Gibbardian (1990) views about norm-expressivism could be used as a starting point for developing such a position.

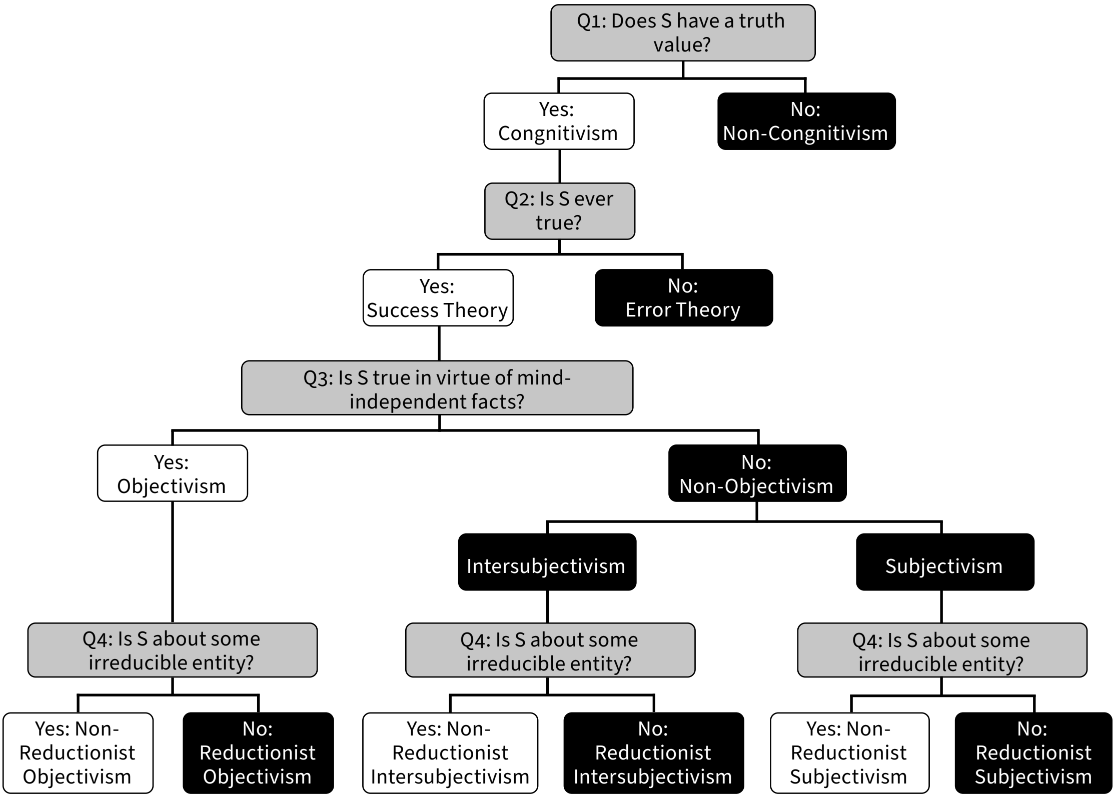

Q1: Does S have a truth value?
Yes: Congnitivism
No: Non-Congnitivism
Q2: Is S ever true?
Yes: Success Theory
No: Error Theory
Q3: Is S true in virtue of mind-independent facts?
Yes: Objectivism
No: Non-Objectivism
Intersubjectivism
Subjectivism
Q4: Is S about some irreducible entity?
Q4: Is S about some irreducible entity?
Q4: Is S about some irreducible entity?
Yes: Non-Reductionist Objectivism
No: Reductionist Objectivism
Yes: Non-Reductionist Intersubjectivism
No: Reductionist Intersubjectivism
Yes: Non-Reductionist Subjectivism
No: Reductionist Subjectivism

Figure 17.3: Anti-realisms in social ontology

Sentences or entities?

We have plotted the maps of realisms and anti-realisms with the help of sentences (*Ss*) being truth-apt or true or reducible. One might object to this approach that instead of using sentences about social entities a systematisation of kinds realisms should be about entities (*X*s) and their reducibility. Does this difference in formulation matter? Not really, we argue in this subsection.

Starting with sentences being truth-apt brings to fore well-known questions about the nature of truth. We do not presuppose or expound any specific substantive theory of truth, and we hold that what we say here is neutral with respect to different understandings of truth. As it might turn out that for example the difference between cognitivism and non-cognitivism should be defined differently depending on whether one holds a deflationary, a minimalist, a perspectivalist or a correspondence theory of truth, in this section we try to shed light on how, roughly, the four guiding questions could be rendered in 'entity-talk' rather than in 'sentence-talk.'

The first question can start, instead of sentences and their truth-aptness, with the entities that figure in theories and discourses about a domain. Do the entities in question exist? For example, do the theoretical (unobservable) entities that figure in scientific theories exist? Are there entities to which the theoretical terms refer? 'Scientific realism' is the traditional view that yes, there are such entities, whereas 'instrumentalism' holds that the theoretical concepts are merely useful instruments and should not be taken to refer to entities.[31] For example, even if there is no such thing as a 'centre of gravity,' it is a useful concept. Many authors (such as Sayre-McCord 1986) treat the debates between scientific realists and instrumentalists on the one hand, and between cognitivists and non-cognitivists on the other hand, as centred around the same question. While questions remain about the exact relationship of sophisticated non-cognitivism and sophisticated instrumentalism, they share the idea that some seemingly referring aspects of discourse can be interpreted as having some other function. The first way of being anti-realist then is to argue that contrary to appearances, our discourse is non-committal with regard to the existence of such and such entities. This way of formulating the question does not refer to the truth of sentences. Yet it does not take a stand on the question whether or not *Xs* exist, but remains non-committal.

The second question can be understood as distinguishing between eliminativism about *Xs* and non-eliminativist realism about *Xs*. Eliminativists hold that there are no *Xs*, i.e. that *Xs* don't exist, whereas the non-eliminativists would include *Xs* in the inventory of what exists. This formulation wears its metaphysical character on its sleeve, as it were. (By contrast, error-theory and success-theory were formulated as divided over the issue of whether certain sentences are ever true).

The third question can then be recast as about whether the relevant *Xs* are mind-independent. Again, this question is straightforwardly about entities and its formulation avoids reference to 'truth.'[32] Our (Q3) above is about the mind-independence of what makes sentences true. If it turns out that for independent reasons the formulation in terms of entities is better than the formulation in terms of sentences, the translation should be pretty straightforward.

The fourth question concerns reducibility. In debates about reducibility, entity-reduction is often distinguished from theory-reduction. For our concerns, it is the reducibility of some *Xs* to some *Ys* that is central. We discuss this in more detail below.

Contested issues

Providing the basic map of realisms and anti-realisms in social ontology is the main aim of this contribution. Maps in general may be found wanting in two important respects: their resolution may not be sufficient for your purposes, and they don't tell you where you are on the terrain they depict. A map of the sort we have plotted in the previous section might warrant reactions of this kind in that, firstly, it may not be

[31] See e.g. Raatikainen (2014).

[32] For an overview on how different theories of truth relate to theories of realism and anti-realism, see section 4 in Glanzberg (2021); and section 6 in Miller (2021).

sophisticated enough to locate a specific view about social phenomena. For instance, the map as given doesn't distinguish between and define variants of non-cognitivism in social ontology which may be of interest to accounts according to which collective intentionality is best analysed within an expressivist framework. Secondly, although the map helps locate different views using the four questions to determine the respective coordinates, it doesn't offer guidelines as to which definition of realism is best suited for social ontology.

With respect to the first point, which we accept as valid, we can here only defer to future work towards a more sophisticated map of realisms and anti-realisms. Working this out would go beyond the scope of the present contribution. With respect to the second point, we want to use this section to at least briefly discuss some contested issues that will help specify which definition of realism might plausibly be adopted for social ontology.

In doing so, we take up the issues targeted in our guiding question in reverse order, turning to the question of fundamentality and reducibility first and the question of mind-independence second. We argue that insistence of these criteria as marks of realism in social ontology is unconvincing and sketch an argument for success theory as a useful definition of realism.

Fundamentality and irreducibility

The issue of the reducibility of collective or supraindividual entities to individuals, has dominated the social ontological discussion in the philosophy of the social sciences, especially in interplay with debates about ontological and methodological individualism.[33] Unsurprisingly perhaps, those who defend the irreducibility of the supraindividual entities, have come to be called 'realists' about those entities. However, should the definition of realism in social ontology be tied to a requirement of fundamentality and irreducibility?

Within general metaphysics, the relevance of reducibility to being 'real' has been influential in debates about fundamentality and grounding. The fundamentality-approach can be understood as supporting the idea that only the fundamental entities are real. A broader view holds that there are two kinds of real entities: in addition to the fundamental ones, those that are non-fundamental and irreducible.

One general challenge for any such view is to make sense of the distinction between eliminativism and reductionism. For how are we to distinguish between eliminativism and reductionism about *Xs*, if *Xs* indeed are not fundamental and irreducible? Current scientific understandings advise us to be eliminativists concerning phlogiston and witches, for instance, and to be error-theorists concerning astrological claims on the whole. If it turns out that football teams, corporations and states can be reduced to individuals (or to some further fundamental entities

[33] See Epstein (2009), Ylikoski (2017), Van Riel and Van Gulick (2019).

all the way down to the micro-physical),[34] should we then regard the Kaizer Chiefs, Supercell, or Estonia with the same kind of suspicion advised vis-á-vis phlogiston, witches, or astrological energies? Fine (2001), for instance, discusses attempts to make conceptual room for the distinction between downright eliminativism and non-fundamentality. In Himmelreich's words (2019, 6) this amounts to the attempt to distinguish 'what exists from what really exists.'

The definition of realism (R4), according to which, by implication, only fundamental, ungrounded, irreducible, or *sui generis* entities are real, is indeed accepted by many as one feature of realism (e.g. Baker, Fine, Pettit, Mäki, Schaffer, Sider, *op.cit.*).[35] The credentials of this view will partly depend on whether there is a non-ad-hoc account of the relevant sense in which tables, football teams, states, which are held to not 'really exist,' nonetheless 'exist' in some relevant sense. It is not entirely clear that a good, disciplined account along these lines has been provided. If one wishes to bite the bullet and be an eliminativist about these entities, one actually defends anti-realism of the (AR2) kind, which amounts to there being no fourth sense of anti-realism (AR4) after all.

The alternative is to hold that even reducible, non-fundamental, grounded entities are real. On this view, groups would be real entities despite being reducible to individuals. The reducible, non-fundamental, grounded entities nonetheless exist, and no distinction between 'really existing' and 'existing' is needed on this view.

In addition to such general considerations, the relevance of non-reductionism (R4) in social ontology can be questioned. Various authors - among them Haslanger (2012), Barnes (2014 and 2017), and Mikkola (2017) - have suggested that social (and feminist) ontology must go beyond the fundamental. If fundamentality is required for reality, then nothing in the social or institutional domain is real.

To save the intuition that the study of non-fundamental aspects of the social world can nonetheless be about something 'real', the definition (R4) requires that the entities are shown to be irreducible. A central motivation for classifying some non-fundamental things as irreducibly real is that they can make a causal difference, or a normative difference.

Causality is very commonly understood to be a mark of the real.[36] If something causes something, then it presumably exists and is real. This is a powerful reason

[34] There is an important class of arguments that suggest that the reduction fails. They aim to show that groups are not reducible to individuals, for example because they have causal powers that individuals lack, and that they are thus indispensable in best explanations. Or it may be that they have irreducible deontic or normative features, and thereby are indispensable in best deliberation, or in living everyday lives (Taylor 1989, 58; Enoch 2007, 22; Thomasson 2019). The most permissive and less permissive views would disagree on whether it matters that the social entities simply have empirical features that are irreducible to the features of individuals (say, a team may have 11 members, but none of the members has 11 members). These arguments try to show that some social entities are irreducible in the relevant sense, or sui generis, or emergent, even though some aspects of the social entities are partially grounded in facts about individuals.

[35] For the view that this question is orthogonal to realism, see Miller 2021, section 4.

[36] See e.g. Barnes (2014), Haslanger (2012), and Psillos (2011) for critical discussion. For argumentative use of causal efficacy as a criterion of reality, see e.g. the debate between Hindriks (2017) and Tuomela (2017), both of whom accept that the reality of group agents hangs on their causal efficacy.

to regard something as real (even in cases of mind-dependence; see the next subsection). While less often noted, having normative roles or normative significance can be equally important as having causal roles or significance: for something to make something wrong, or good, it also is a strong indicator of its being real. And as causal relations are typically contrasted with constitutive relations, it is also possible to hold 'playing a constitutive role' or 'having constitutive relevance' as an indicator of being real. For our purposes, theorists stressing the causal role only or also admitting normative and constitutive roles are on a par, they are just varieties of 'non-redundancy' as a mark of irreducibility - the main contrast is with 'epiphenomenal' properties or entities that do not play such roles.

We appreciate the emphasis on such 'non-redundancy,' and the problem with (R4) is not with the irreducible entities that it classifies as real. The problem is with entities that are reducible: is the view really that reducibility amounts to non-reality? Would that not make reducibility amount to elimination?

It is possible to argue that something is real even though it is 'epiphenomenal,' i.e. lacks causal relevance. Indeed, presumably only real, existing things can be epiphenomenal. On the other hand, one could apply Occam's razor in a broadly pragmatist spirit, so that in thinking about whether to regard something as really existing or not, one ends up holding the view that something deserves a place in the one's inventory of the world's furniture *only* if it is needed in causal explanations, or normative explanations, or constitutive explanations. The former view might hold that the existence of some causally, normatively or constitutively inert *Xs* is not dependent on whether we should add those *Xs* to our inventory - the emphasis on our explanatory needs would be to put the cart before the horse, as it were. Perceived causal, (or constitutive, or normative) significance may be a reason to believe in the reality of something rather than a suggested analysis of what it means to be real: perhaps causally inert entities exist and are real as well.

This subsection has made three points: first, it is implausible to define reality in terms of fundamentality alone as that would lead to forced anti-realism about everything social and institutional. Second, any definition linking irreducibility and reality must come with an account of how to distinguish reduction and elimination, and a related account of the distinction between 'existing' and 'really existing'. Third, while it is a good idea to focus on entities that have causal, normative or constitutive significance, it is less clear whether we should deem the epiphenomenal, reducible aspects of reality as 'not real'. It seems to be more faithful to the spirit of realism to acknowledge that some aspects of reality are not playing those roles, and that the question of reducibility is in the end orthogonal to the question of realism.[37] Overall, there is strong reason to reject versions of (R4) that appeal to fundamentality, and reason to feel some unease with versions of (R4) that appeal to irreducibility. Yet, the indubitable importance of causal (and relatedly, normative and constitutive) roles as marks of real gives a reason to acknowledge that (R4) has something going for it.

[37] Cf. Miller 2021.

Mind-independence

Mind-independence may be a good enough criterion of realism in domains such as physical reality. To think that physical objects consist of, for example, nothing but sense data would be expressive of a form of idealism, and clearly opposed to the spirit of realism. But much, perhaps all, of social reality is mind-dependent in one way or another. Thus, either there is no room for realism in social ontology or mind-independence does not provide a useful criterion to identify realism in this domain (see e.g. Khalidi 2016). As expounded above, we treat mind-independence as only one of four criteria of realism on the map of realisms meant to apply to any domain.

Social constructions are clearly mind-dependent.[38] If there weren't any agents engaged in intricate forms of interaction and forming specific individual or shared attitudes, there wouldn't be social entities such as banks, football clubs, democratic elections, parking areas or labour unions. With the help of the basic map, we can see that such entities can nonetheless be considered real in the sense of (R1), (R2), and (R4). Anyone who holds that claims about social constructions have truth value, are sometimes true, and are about irreducible entities, is both a realist ((R1), (R2), and (R4)) and a social constructionist (and thus non-objectivist). According to the basic map, to be a social constructionist is only incompatible with being an objectivist realist (R3). But in all other senses, social constructionists can be realists. This is worth emphasising, as many social constructionists self-identify as realists (e.g. Haslanger 2012) whereas others (such as Ásta) are sometimes classified as anti-realist (see Ásta 2015, Barnes 2017, Mason 2020 for discussion).

That said, it still seems that some kinds of mind-dependence conform more with the realist spirit than do other kinds of mind-dependence. The intuition that mind-dependence amounts to anti-realism can perhaps partially be saved on a more sophisticated map. Social constructionism need not be equally seriously opposed to the spirit of realism as, say, textbook sense-data idealism about physical objects is (cf. Ásta 2015). A more sophisticated map would show that the nature of mind-independence is to be studied more carefully in view of this. We here outline two ways in which the basic map might be refined (see Page 2006 for a third way[39]).

(1) First, if you think of the social realm as directly dependent on the experiences and thoughts of *current* observers, your stance is analogous to idealism about material objects. But even if social reality is causally or constitutively dependent on past actions, it may be independent of current observers. Further, if some aspects of social reality are dependent on *intersubjective or collective* acceptance, these aspects may nonetheless be relatively independent of any individual. This marks a

[38] Our discussion in this section is similar in spirit to Thomasson's (2003, 584f.) treatment of varieties of mind-dependence. We will discuss further varieties of mind-dependence in a separate paper.

[39] Page 2006 distinguishes between ontological, causal and structural independence from 'individuative' independence.

crucial difference between subjectivist and intersubjectivist accounts.[40] What is thus needed is a more precise account of both the relata and the nature of the dependence relation in play. In other words, further work is needed for inventories and analyses (a) of what social reality is dependent on - be it beliefs, intentionality, science, declarations, collective acceptance, official decisions by institutions, informal communal recognition etc. -, (b) of the nature of the respective dependence - be it causal, constitutive, normative, ontological, etc. -, and especially (c) of arguments about what kind of dependence is the crucial kind of mind-dependence for the question of realism.[41] One could be, to adapt a term, a 'quasi-objectivist' and say that whereas social kinds, social injustices, oppressive practices, and everyday institutional facts are not strictly speaking objective, they are very close to being objective in that they do not display the kinds of mind-dependence at odds with the spirit of realism. They do not go away, whatever an individual thinks.

(2) Second, social reality contains heterogeneous elements, and some elements may be multiply mind-dependent, whereas others are relatively mind-independent (even though mind-dependent in some general sense). In this vein, Khalidi (2015) distinguishes between, on the one hand, social kinds of which a token can be a member of the kind only if the token is collectively regarded as such - for example, one can only be the president of the U.S in this way -, and on the other hand, kinds within which the general type is mind-dependent, but once the kind exists, tokens can become members without any thoughts targeted at the tokens - for example, there could be a dollar bill that was produced in the usual way but then lost so that no-one has ever had any awareness of it. The former social kinds are mind-dependent in two ways (as types and tokens), the latter only in one way. Similarly, although all action is mind-dependent, patterns of interaction may emerge without anyone intending or even noticing this. Emergent patterns of interaction (not necessarily noticed by anyone) are more mind-independent than conferred statuses, which exist only when conferred, and thus are more thoroughly mind-dependent. Consequently, whereas patterns of interaction and conferrals of statuses both depend on human action and mental attitudes in some general sense, conferrals of statuses are

[40] This distinction warrants more attention that we can give it in this paper. Our focus is on which strength of the claim concerning mind-dependence non-objectivist views would be advised to adopt. Although our explanations and most accounts mentioned in this section take a broadly intersubjectivist line - i.e. in terms of social practices or collective beliefs -, this alone may not dissuade those attracted to subjectivism. However, a proper defence of intersubjectivism and engagement with subjectivism will need to be provided on another occasion.

[41] See e.g. Vinueza 2001.

> thus dependent in a further way. The recipe for examining degrees of mind-dependence would then be to study which (if any) phenomena are mind-dependent in multiple ways and which are such only in some very general sense.

We cannot pursue this here, but our suggestion is that a more sophisticated map of realisms would capture positions as 'more objectivist' and 'less objectivist' by analysing the kinds of dependence at issue. One could then suggest a definition of 'quasi-objectivism' for a certain family of positions closest, or at any rate relatively close to objectivism.

The more sophisticated distinctions may be needed in accounting for the sense in which some social constructionist accounts are more realist (more objectivist) and some more anti-realist (less objectivist), even though they all subscribe to a mind-dependence view, i.e. (AR3) on our basic map.

For Success theory, against mere cognitivism

By contrast, minimal realism (R2) is a fully recommendable definition of realism in social ontology. This is the sense of realism in which one can be a realist about the less-than-fundamental, and the not-fully-mind-independent. And as commented above, this may be especially appealing when an entity deemed real has causal powers or normative roles.

The corresponding form of anti-realism is what Kit Fine (2001) calls the eliminativist or sceptical form of anti-realism. This view of anti-realism holds that if numbers are not real (in the sense of R2), then there are no prime numbers between 3 and 6. And as there *are* prime numbers between 3 and 6, we have good reason to be realists *in this sense*. There are also good reasons to be eliminativists about various entities from witches to phlogiston (namely, the ones that do not exist).

But why not adopt an even more minimal definition of realism in social ontology? Why doesn't, in other words, an affirmative answer to our (Q1) suffice in demarcating realist from anti-realist views? Here we hold that cognitivism alone is a too modest view, as it is compatible with error theory. Here, the key argument is simply that existence seems just too central for something to be real - the mere availability of views on which realism about *Xs* entails that *Xs* exist places a heavy burden of proof on views on which realism is compatible with non-existence. Such a massive majority of usages of 'real' in all contexts from everyday life and philosophy to social research connote being real and existing, that R1 cannot but feel inadequate. To avoid this inadequacy, realism in a domain *D* better entail a success theory about the central claims or sentences of domain D. The basic view we labelled 'most minimal realism' (R1) is, in short, too minimal to be an acceptable account of realism, when other options on which existence is a defining part of realism are available. Yet, here too there is conceptual space for more sophisticated views.

Defenders of *quasi-realism* typically argue that their theory can be success-theories while capturing the strengths of a non-cognitivist, expressivist analysis of the relevant discourse.[42] It is worth pointing out that in social ontology expressivism hasn't been popular at all (despite Wilfrid Sellars (1956 and 1968) being both an expressivist and an important early defender of we-intentions), and for example Mills (1998, 49) puts it aside as an irrelevant option. One reason for that may be that expressivism has an immediate appeal in metaethics that it lacks in social ontology: a salient feature of ethical or normative talk and thought is its practicality and connection to motivations. By contrast, the immediate salient feature in metaphysics of the social and institutional reality is its dependence on human constructions and conceptions. Social constructionism has a lot going for it, as a theory in social ontology. In metaethics, social constructionism has been an equally marginal position as expressivism is in social ontology; and expressivism typically does not go very well with social constructionism.[43] They both differ from robust objectivist mind-independence views (R3), but expressivism parts company already in Q1, whereas social constructionism only in Q3. If one is to defend social constructionism, one should opt for R1 and R2; expressivism doesn't (by definition, it rejects R1). Therefore, while a quasi-realist can argue that apparently realist phenomena can be reinterpreted on an expressivist basis, it may lack the initial motivation: why not be, say, a realist social constructionist instead?

Conclusion

We have suggested that there are at least four ways of defining realism in social ontology, labelled (R1) through (R4). Against non-cognitivism, (R1) holds that statements such as 'the bank closes at 4 pm' or 'the Kaizer Chiefs scored a goal,' abbreviated as statements about *Xs*, have truth-value. Against error theory, (R2) holds that some of the statements are true, and so that there are *Xs* in virtue of which those statements are true. Against non-objectivist (subjectivist or intersubjectivist) views granting the mind-dependence of social constructions, (R3) holds that for *Xs* to be real they have to be objective, mind-independent. And against reductionism (of reducible, non-fundamental, grounded entities to something more fundamental), (R4) holds that for *X* to be real, it has to be irreducible, ungrounded, fundamental, *sui generis*. This basic map helps to see how social constructionists can be realists in the sense of (R1), (R2), and (R4). A more sophisticated map would zoom in on the third of these demarcating issues and distinguish between kinds and multiple degrees of mind-dependence. In analogy to these forms of realism, there are four types of anti-realism (AR1)-(AR4).

[42] Cf. Blackburn (1993), van Roojen (2015).

[43] Cf. Essays in Lenman and Shemmer, eds. (2012), for arguments for and against the compatibility of constructivism and non-cognitivism.

The basic map is primarily guided by the clarificatory aim of making sense of debates about 'realism' in social ontology. Above, we assessed reasons for and against the four suggested definitions of realism. We argued that (R1) seems necessary, but insufficient: it is counter-intuitive to take error-theorists (eliminativists) to be realists. (R3), in its basic version, seems ill-suited for social ontology. Social reality just isn't fully mind-independent, but that does not seem to justify throwing social entities in the dustbin with phlogiston or witches. And (R4) has problems of its own: to distinguish between eliminativism and reductionism it may need to appeal to the distinction between 'really existing' fundamental entities and 'existing, but less real' non-fundamental entities. It may well be that it is best to merely talk about reducibility, fundamentality or grounding without the assumption that the status as *real* hangs on those investigations. Further, fundamentality as a definition of real is ill-suited for social ontology. A more fruitful understanding of 'non-redundancy' is that of causal or normative relevance: certainly at least non-redundant entities are real (even if mind-dependent, or even if non-fundamental). However, this may be a reason to believe in their reality rather than a suggested analysis of what it means that they are real: perhaps causally inert entities exist, or are real, as well.

For these reasons, realism as success-theory (R2) looks to be the most viable definition of realism in social ontology. The second best is then non-reductionism (R4) on the 'less permissive' reading we have detailed, which permits causally and normatively significant properties and entities to count as real, but not others. In some contexts, adopting this view on the 'irreducibility'-reading is fine, but substantively the same points can be made without the problematic distinction of 'existing' and 'really existing'.

Given the central appeal of mind-independence, a more sophisticated plotting of views in between subjectivist and objectivist realism is called for, and more detailed accounts of what is dependent on what, and what kind of dependence is at stake are called for, before it can be an acceptable definition of realism for social ontology. We have taken first steps in this direction, more may need to follow.

One might still wonder whether social ontologists need the sort of metametaphysical clarifications towards which we worked in this paper. That is, why not drop the moniker 'realism' and simply examine the debates of cognitivism vs. expressivism, error theory vs. success theory, mind-dependence vs. mind-independence, reducibility vs. non-reducibility, or causal or normative relevance and redundancy?

There are several reasons to examine which usages of 'realism' are most fitting in some contexts, such as the social and institutional world. The term 'real' - with all its more or less confusing usages - is so deeply embedded in different discourses that it is less futile to argue for reasoned usages than to hope that the term would simply disappear. These discourses include those of lay people in their everyday life, social scientists which take themselves to be studying something real and those philosophers interested in locating social entities in broader metaphysical understandings of the universe. In practical and political contexts, it is rather obviously important to be able to say that, for example, some forms of oppression are real. As realism-talk is likely not going to go away in any of these guises, it is better to promote disciplined usages than simply give up.[44]

[44] We owe special thanks to everyone who has commented our talks and previous drafts on this topic. You know who you are, and our gratitude to you is real.

References

Appiah, K. Anthony (1996): 'Race, Culture, Identity: Misunderstood Connections', in K. A. Appiah and A. Gutmann, *Color Conscious: The Political Morality of Race*, Princeton: Princeton University Press, 30–105.

Ásta [Sveinsdóttir] (2015): 'Social Construction,' *Philosophy Compass* 10(12): 1–9. URL = https://doi.org/10.1111/phc3.12265.

Ásta (2018): *Categories We Live by: The Construction of Sex, Gender, Race, and Other Social Categories*, Oxford: Oxford University Press.

Baker, Lynne (2007): *The Metaphysics of Everyday Life: An Essay in Practical Realism*, Cambridge: Cambridge University Press.

Barnes, Elizabeth (2014): 'Going Beyond the Fundamental: Feminism in Contemporary Metaphysics', *Proceedings of the Aristotelian Society* 114(3): 335-351. URL = https://www.jstor.org/stable/44122576.

Barnes, Elizabeth (2017): 'Realism and Social Structure,' *Philosophical Studies* 174(10): 2417–2433. URL = https://doi.org/10.1007/s11098-016-0743-y.

Blackburn, Simon (1993): *Essays in Quasi-Realism,* Oxford: Oxford University Press.

Cohen, Shlomit Wygoda (2022): 'Mind Independence versus Mind Nongroundedness: Two Kinds of Objectivism,' *Ethics* 132(1): 180–203. URL = https://doi.org/10.1086/715278.

Dennett, Daniel (1987): *The Intentional Stance,* Cambridge, MA: MIT Press.

Devitt, Michael (1984): *Realism and Truth,* Princeton: Princeton UP.

Devitt, Michael (2010): *Putting Metaphysics First*, Oxford: Oxford UP.

Díaz-León, Esa (2013): 'What Is Social Construction?', *European Journal of Philosophy* 23(4): 1137-1152. URL = https://doi.org/10.1111/ejop.12033.

Dreier, James (2004): 'Meta-Ethics and the Problem of Creeping Minimalism', *Philosophical Perspectives* 18(1): 23–44. URL = https://doi.org/10.1111/j.1520-8583.2004.00019.x.

Dummett, Michael (1978): *Truth and Other Enigmas*, Boston: Harvard University Press.

Enoch, David (2007): 'An Outline of an Argument for Robust Metanormative Realism', in R. Shafer-Landau (ed.), *Oxford Studies in Metaethics, 2,* Oxford: Oxford University Press, 21–50.

Epstein, Brian (2009): 'Ontological Individualism Reconsidered,' *Synthese* 166 (1): 187-213. URL = https://doi.org/10.1007/s11229-007-9272-8.

Epstein, Brian (2018): 'Social Ontology', *The Stanford Encyclopedia of Philosophy* (Summer 2018 Edition), Edward N. Zalta (ed.), URL = https://plato.stanford.edu/archives/sum2018/entries/social-ontology/.

Fine, Kit (2001): 'The Question of Realism,' *Philosopher's Imprint*, 1(2): 1–30. URL = http://hdl.handle.net/2027/spo.3521354.0001.002.

Gibbard, Allan (1990): *Wise Choices, Apt Feelings*, Cambridge: Harvard University Press.

Glanzberg, Michael, (2021): 'Truth', *The Stanford Encyclopedia of Philosophy* (Summer 2021 Edition), Edward N. Zalta (ed.), URL = https://plato.stanford.edu/archives/sum2021/entries/truth/.

Glasgow, Joshua (2008): *A Theory of Race*, New York: Routledge.

Guala, Francesco (2016): *Understanding Institutions. The Science and Philosophy of Living Together,* Princeton: Princeton UP.

Haslanger, Sally (2012): *Resisting Reality: Social Construction and Social Critique*, New York: Oxford UP.

Haukioja, Jussi (2021): 'Metaphysical Realism and Anti-Realism,' in R. Bliss and J.T.M. Miller (eds.), *The Routledge Handbook of Metametaphysics*, Routledge, 61–70.

Himmelreich, Johannes (2019): 'Existence, really? Tacit disagreements about 'existence' in disputes about group minds and corporate agents,' *Synthese* 198: 4939–4953. URL = https://doi.org/10.100 7/s11229-019-02379-3

Hindriks, Frank (2006): 'Acceptance-Dependence: A Social Kind of Response-Dependence', *Pacific Philosophical Quarterly* 87(4): 481–498. URL = https://doi.org/10.1111/j.1468-0114.2006.00272.x.

Hindriks, Frank (2017): 'Group Agents and Social Institutions: Beyond Tuomela's *Social Ontology',* in G. Preyer & G. Peter (eds.), *Social Ontology and Collective Intentionality: Critical Essays on the Philosophy of Raimo Tuomela with His Responses*, Cham: Springer, 197–210.

Hindriks, Frank (2020): 'How Social Objects (Fail to) Function,' *Journal of Social Philosophy* 51(3): 483–499. URL = https://doi.org/10.1111/josp.12334.

Joyce, Richard (2016): 'Moral Anti-Realism,' *The Stanford Encyclopedia of Philosophy* (Winter 2016 Edition), Edward N. Zalta (ed.), URL = https://plato.stanford.edu/archives/win2016/entries/moral-anti-realism/.

Khalidi, Mohammad Ali (2015): 'Three Kinds of Social Kinds,' *Philosophy and Phenomenological Research* 90(1): 96–112. URL = https://doi.org/10.1111/phpr.12020.

Khalidi, Mohammad Ali (2016): 'Mind-Dependent Kinds,' *Journal of Social Ontology* 2(2): 223–246.

Khlentzos, Drew (2021): 'Challenges to Metaphysical Realism', *The Stanford Encyclopedia of Philosophy* (Spring 2021 Edition), Edward N. Zalta (ed.), URL = https://plato.stanford.edu/archives/spr2021/entries/realism-sem-challenge/.

Lenman, James and Yonatan Shemmer, eds. (2012): *Constructivism in Practical Philosophy*, Oxford: Oxford University Press.

List, Christian and Pettit, Philip (2011): *Group Agency: The Possibility, Design, and Status of Corporate Agents*, Oxford: Oxford University Press.

Mäki, Uskali (2005): 'Reglobalizing realism by going local, or (how) should our formulations of scientific realism be informed about the sciences?', *Erkenntnis* 63(2): 231–251. URL = https://doi.org/10.1007/s10670-005-3227-6.

Mäki, Uskali (2008): 'Scientific Realism and Ontology,' in: St.N. Durlauf and L.E. Blume (eds.), *The New Palgrave Dictionary of Economics,* 2nd ed., Palgrave, 334–341.

Mallon, Ron (2016): *The Construction of Human Kinds*, New York: Oxford University Press.

Mason, Rebecca (2020): 'Against Social Kind Anti-Realism', *Metaphysics*, 3(1): 55–67. URL = https://doi.org/10.5334/met.30.

Mikkola, Mari (2017): 'On the Apparent Antagonism between Feminist and Mainstream Metaphysics', *Philosophical Studies* 174(10): 2435–2448. URL = https://doi.org/10.1007/s11098-016-0732-1.

Miller, Alexander (2021): 'Realism', *The Stanford Encyclopedia of Philosophy* (Winter 2021 Edition), Edward N. Zalta (ed.), URL = https://plato.stanford.edu/archives/win2021/entries/realism/.

Miller, J. T. M. (2022): *Metaphysical Realism and Anti-Realism*, Cambridge UP: Cambridge.

Mills, Charles (1998): *Blackness Visible: Essays on Philosophy and Race*, Ithaca, NY: Cornell University Press.

Page, Sam (2006): 'Mind-Independence Disambiguated: Separating the Meat from the Straw in the Realism/Anti-Realism Debate', *Ratio* 19(3): 321–335. URL = https://doi.org/10.1111/j.1467-9329.2006.00330.x.

Pettit, Philip (2009): 'The Reality of Group Agents,' in C. Mantzavinos, *Philosophy of the Social Sciences: Philosophical Theory and Scientific Practice*, Cambridge, 67–91.

Psillos, Stathis (2011): 'Living with the abstract: realism and models,' *Synthese* 180: 3–17. URL = https://doi.org/10.1007/s11229-009-9563-3.

Putnam, Hilary (1979): *Mathematics, Matter and Method - Philosophical Papers, Volume 1*, 2nd ed., Cambridge: Cambridge University Press.

Raatikainen, Panu (2014): 'Realism: Metaphysical, Scientific, and Semantic', in Kenneth R. Westphal (ed.), *Realism, Science, and Pragmatism*, Routledge, 139–158.

van Riel, Raphael and Robert Van Gulick (2019): 'Scientific Reduction', *The Stanford Encyclopedia of Philosophy* (Spring 2019 Edition), Edward N. Zalta (ed.), URL = https://plato.stanford.edu/archives/spr2019/entries/scientific-reduction/.

van Roojen, Mark (2015): *Metaethics: A Contemporary Introduction*, New York: Routledge.

Sayre-McCord, Geoffrey (1986): 'The Many Moral Realisms', *The Southern Journal of Philosophy* 24(S1): 1–22. URL = https://doi.org/10.1111/j.2041-6962.1986.tb01593.x.

Schaffer, Jonathan (2009): 'On What Grounds What.', in D. J. Chalmers, D. Manley and R. Wasserman (eds.), *Metametaphysics: New Essays on the Foundations of Ontology.* Oxford: Oxford University Press, 347–383.

Schaffer, Jonathan (2017): 'Social Construction as Grounding; Or: Fundamentality for Feminists, a Reply to Barnes and Mikkola,' *Philosophical Studies* 174(10): 2449-65. URL = https://doi.org/10.1007/s11098-016-0738-8.

Schweikard, David P. and Hans Bernhard Schmid (2013): 'Collective Intentionality', *The Stanford Encyclopedia of Philosophy* (Summer 2013 Edition), Edward N. Zalta (ed.), URL = https://plato.stanford.edu/archives/sum2013/entries/collective-intentionality/.

Sellars, Wilfrid (1956): 'Imperatives, Intentions, and the Logic of 'Ought',' *Methodos* 8: 227–268. URL = https://doi.org/10.1093/oso/9780192866820.003.0015.

Sellars, Wilfrid (1968): *Science and Metaphysics: Variations on Kantian Themes*, Dordrecht: D. Reidel.

Sider, Theodore (2011): *Writing the Book of the World*, Oxford: Oxford University Press.

Tahko, Tuomas (2015): *An Introduction to Metametaphysics,* Cambridge: Cambridge University Press.

Taylor, Charles (1989): *Sources of The Self,* Harvard University Press.

Thomasson, Amie L. (2003): 'Realism and Human Kinds,' *Philosophy and Phenomenological Research* 67(3): 580-609. URL = https://doi.org/10.1111/j.1933-1592.2003.tb00309.x.

Thomasson, Amie L. (2019): ‘The Ontology of Social Groups,' *Synthese* 196(1): 4829–4845. URL = https://doi.org/10.1007/s11229-016-1185-y.

Tollefsen, Deborah Perron (2015): *Groups as Agents*, Cambridge UK: Polity Press.

Tuomela, Raimo (2013): *Social Ontology: Collective Intentionality and Group Agents*, Oxford: Oxford University Press.

Tuomela, Raimo (2017): 'Response to Frank Hindriks', in G. Preyer & G. Peter (eds.), *Social Ontology and Collective Intentionality: Critical Essays on the Philosophy of Raimo Tuomela with His Responses*, Cham: Springer, 211–217.

Vinueza, Adam (2001): ‘Realism and Mind Independence’, *Pacific Philosophical Quarterly* 82(1): 51-70. URL = https://doi.org/10.1111/1468-0114.00118.

Ylikoski, Petri (2017): 'Methodological Individualism', in L. McIntyre & A. Rosenberg (eds.), *Routledge Companion to Philosophy of Social Science,* New York: Routledge, 135–146.

Zack, Naomi (2002): *Philosophy of Science and Race*, New York: Routledge.

Zahle, Julie and Finn Collin, eds. (2014): *Rethinking the Individualism-Holism Debate: Essays in the Philosophy of Social Science*, Dordrecht: Springer.

18
The nominalist theory of natural kinds and kind essences

Markku Keinänen

Introduction

The world is divided into entities that belong to *natural kinds*. It seems that we both use natural kinds to individuate entities at the different levels of constitution of reality and make use of natural kinds in scientific explanations. Natural kinds appear to collect sets of features relevant to objects' acting in a certain way in certain circumstances. Moreover, natural kind divisions seem to be independent of us and our classificatory faculties. For instance, it is not up to us that there are different kinds of atoms, molecules and chemical stuff.

It is one of the central questions of metaphysics and ontological category theory to specify the ontological status of natural kinds. Are there natural kinds? If they exist as constituents of the world, are they sui generis entities or, say, complexes of property universals? Moreover, natural kinds are often considered to have essences that collect the properties necessary to the members of the kind. Therefore, we may ask whether there are kind essences and how they must be characterized.

In this chapter, I defend a nominalist conception of natural kinds, which denies the existence of natural kinds as separate entities. Nevertheless, there are divisions of entities into natural kinds and truths about entities belonging to a natural kind. Therefore, I accept the general view Bird & Tobin (2022) call "naturalism about natural kinds" and reject the strong error theoretic version of eliminativism about

Keinänen, M. (2025). The nominalist theory of natural kinds and kind essences. In A. Laitinen, M. Keinänen, J. Reinikainen & A. Honkasalo (eds.). *Language, Truth, and Reality.* Tampere University Press. (pp. 227–236). https://doi.org/10.61201/h5fbdk64

natural kinds (cf. Ludwig 2018). My nominalist conception denies the existence of natural kinds and thus rejects any attempt to *reduce* natural kinds to any other entity such as a set of tropes. Therefore, I label my view "the *eliminativist nominalism* about natural kinds".

The eliminativist nominalism seems to create a problem about kind essences: if there are essential properties of certain natural kind K or kind essences but no natural kinds, which entity does have these properties or essences? The problem is still more pressing if we assume (as I do in this paper) that natural kind divisions are independent of us and our classifications. What is the basis of these divisions if there are no natural kinds or similar entities possessing the kind essences? My aim here is to argue that the eliminativist nominalism can solve this problem by means of the following strategy: although there are no natural kinds, there is general talk about entities belonging to natural kinds. This talk is made true by different kinds of entities and structures of entities, which are taken as instances of different natural kinds. Moreover, having certain features or a certain kind of structure are necessary to object's belonging to natural kind K if and only if their possession constitutes both sufficient *and* necessary condition for the application of the corresponding natural kind term. As an advocate of the trope theory SNT (the Strong Nuclear Theory) (Keinänen 2011; Keinänen & Hakkarainen 2024), I take the world to be ultimately constituted by tropes, that is, thin particular natures (such as certain determinate masses, charges and lengths) in some specific locations. However, the proposed view of natural kinds is not tied to trope theory and other nominalists (e.g., substance-mode-theorists) might adopt it.[1]

In what follows, I first specify what natural kinds are and why we need a metaphysical view of natural kinds. I then present the eliminativist nominalist view of natural kinds in more detail. The article ends with a brief concluding section.

Natural kinds

Prima facie, entities and concrete individual objects in particular share natural properties like the mass of 1kg and are therefore said to belong to the same *natural class* (class of 1kg objects). As there are divisions of objects into natural classes based on their mind-independent similarities, there are analogous divisions into natural kinds. However, objects belonging to a natural kind are typically required to share several distinct features and there is some exhaustive division of objects into natural kinds. Because of bringing effective classification to reality, natural kinds are important to scientific explanations and inductive generalizations.

We can point to prima facie examples of natural kinds at different levels of complexity. Fundamental microparticles divide into natural kinds such as electron,

[1] Of course, different nominalist ontological category theories have different resources for this task. For instance, John Heil (2012) has suggested that substances and their modes are sufficient truthmakers of attributions of natural kinds to substances.

down-quark or tau-neutrino. Similarly, there are natural kinds of atoms, molecules (e.g., water molecule) and chemical stuffs (e.g., water). Also living organisms seem to divide into natural kinds such as polar bear or oak tree. In addition to natural kinds of objects, there are natural kinds of processes such as different kinds of chemical reactions.

Very different kinds of beings are members of natural kinds, but we seem to give to natural kinds similar functions. Therefore, it is fruitful to present the different types of functions given to natural kinds:

1. Natural kinds permit inductive generalizations (Bird & Tobin 2022) and have a central role in scientific explanation (Boyd 1999, 2010; Hawley & Bird 2011).
2. The members of a natural kind possess certain basic dispositional properties, and some fundamental laws of nature concern the behaviour of every member of some kind K (Ellis 2001; Lowe 2009, 2015).
3. Natural kinds determine the identity conditions of their members (Loux 1978; Lowe 1998, 2009).
4. Natural kinds are referents of natural kind terms (Kripke 1980; Lowe 2009).

Function 1 has motivated the discussion of natural kinds in philosophy of science. Function 2 is present in metaphysics of science and in some theories of dispositional properties and laws of nature.[2] Function 2 is most notably advocated by Neo-Aristotelians (like Ellis and Lowe), who identify natural kinds with substantial kind universals. If natural kinds perform function 2, the corresponding classification of objects into natural kinds permits inductive generalizations and has a central role in scientific explanation, that is, these natural kinds also perform function 1. By contrast, the converse need not hold. We can well consider natural kinds having a central role in scientific explanation, but these natural kinds need not have any role in fundamental laws or as bearers of fundamental properties (Bird & Tobin 2022, sec.1).

The identificatory function 3 of natural kinds is put forth by Neo-Aristotelian metaphysicians. It is associated with the parallel function of natural kind terms or sortal concepts referring to natural kinds to provide us with the identity criteria of the objects belonging to the kind (Lowe 2009, 2015). Finally, both Neo-Aristotelian metaphysicians and certain advocates of the externalist theory of reference put forth function 4. Nevertheless, function 4 is not tied to the identification of kinds with kind universals. For instance, metaphysicians reducing natural kinds to complex property universals (Hawley & Bird 2011) or sets of particulars could assign it to natural kinds as well. Besides, natural kinds functioning as referents of natural kind terms

[2] Alexander Bird (2007) claims that we need not postulate natural kinds to perform this function, but fundamental dispositional properties are possessed by objects (e.g., basic microparticles).

need not be considered to fulfil stringent criteria: one might accept relationally or conventionally identified kinds to function referents of natural kind terms (Beebee & Sabbarton-Leary 2010, 4).

This brings another important dimension of comparison between the different accounts of natural kinds. We can ask how one can answer the *naturalness question*: what should be required of a kind to be considered *natural kind* instead of being an artificial or conventionally identified kind? Brian Ellis (2001, 19–23) sets six constraints on natural kinds as contrasted with conventionally defined or accidental kinds. In addition to demanding that natural kind divisions must be mind-independent, he requires that they are sharp and not gradual. Third, every natural kind must be determined by a set of features or a structure essential to the members of the kind. Fourth, these features must be intrinsic features of the kind members, or the structure must be intrinsic to each kind member. Finally, according to Ellis, every permanent difference in intrinsic features must lead to a division of natural kinds and natural kinds must constitute a hierarchy.[3]

As Ellis himself admits, these criteria are demanding and rule out biological species as natural kinds. The main problem with Ellis' constraints is that there seem to be mind-independent kind-like divisions among entities that do not fulfil constraints two, three or four. Here biological species might be a case in point. According to some accounts, the membership of an individual in a species (considered a biological kind) is determined (at least) by its lineage and (possibly) some other extrinsic features. Since species are in constant flux, they do not have sharp boundaries or even a clearly specifiable set of kind-determining features (Ereshefsky 2010; Bird & Tobin 2022, sec. 2.1). There might be equally legitimate alternative ways to divide living organisms into distinct species (Kitcher 1984). Still, our kind terms might well track some mind-independent divisions among biological organisms.

At another end of the spectrum, natural kinds of physical microparticles, atoms and molecules have a clear set of features or a structure intrinsic to every member of the kind. They seem to fulfil all, or almost all criteria Ellis sets to natural kinds.[4] Things get more complicated if we go to the kinds of chemical stuff and chemical compounds, in particular. Take, for example, water. Metaphysicians of science do not agree on whether water has a micro-structural essence (i.e., micro-structuralism about water). While Needham (2000) argues against micro-structuralism, Hendry (2006, 2023) puts forth and defends a qualified form of micro-structuralism about water, which might be extended to some other chemical compounds.

Being a Neo-Aristotelian realist, Ellis has a clear motivation to set restrictions on the different substantial kind universals (i.e., natural kinds) he postulates. By contrast, the eliminativist nominalist about natural kinds may adopt a more

[3] Here, I have changed the order of presentation of requirements and joined Ellis' speciation requirement and hierarchy requirement into one.

[4] Since water molecules, for instance, are in constant flux as parts some portion of water (cf. Hendry (2006, 869 ff.), it is, however, contestable whether we can draw sharp boundaries for the set of the instances of the kind water molecule.

relaxed view, according to which naturalness of kinds comes in degrees. The kinds of elementary particles, atoms and molecules might be considered *perfectly natural kinds*: the members of a kind have a set of intrinsic features individually necessary and jointly sufficient for being a member of the kind. By contrast, if the membership in a natural kind is partly determined by the relations the kind members bear to other individuals, we have a less than perfectly natural kind. Another possible case of a less than perfectly natural kind is that there are sets of alternative features necessary to an entity to be a member of natural kind. If biological species are natural kinds – which is also contestable - they might have sets of alternative kind-determining features.[5] Since there are no natural kinds as constituents of reality, the eliminativist nominalist view can remain tolerant to all these cases – it suffices that there is a natural division that can be specified by means of comparatively simple criteria.

The nominalist view of natural kinds

As indicated above, the eliminativist nominalist view of natural kinds denies the existence of kind universals (Ellis 2001; Lowe 1998, 2006, 2009) and the identification of natural kinds with complex property universals (Hawley & Bird 2011). Moreover, this view rejects the nominalist attempts to identify natural kinds with sets of objects (Quine 1969) or abstractions from natural kind terms (Keinänen 2015). The main motivation here is categorial ontological economy: we can take care of most of the above functions set to natural kinds without postulating them as separate constituents of reality.

In order to get my nominalist view off the ground, I tentatively adopt an application theory of natural kind terms. In other words, natural kind terms are predicates applying to objects rather than singular terms referring to natural kinds.[6] We might later replace the application theory with some better account, which provides us with a more precise conception of the special identificatory function of natural kind terms. Thus, the advocate of the nominalist view can agree with Neo-Aristotelians on the epistemic role of natural kind terms in the identification of objects. However, they deny that there are natural kinds having the corresponding function to (contributing to) determine the identity conditions of objects.[7]

The eliminativist nominalist is obliged to specify in more explicit terms what is for a structure of entities to function as a *truthmaker* of the attribution of natural kind to an object. Here it suffices to lay down three principles of truthmaking, which are

[5] Another alternative put forth in the discussion about species is to assume that instead of being natural kinds at all, species are complex individuals.

[6] This view of natural kind terms as predicates applying to objects is associated with Putnam (1975) and later advocated in different forms, for instance, by Devitt (2005) and Haukioja (2012).

[7] Following Lowe (2003), I distinguish between individuation in the epistemic sense (identification) and individuation in the metaphysical sense (individuation). Correspondingly, natural kind terms can have a central role in the identification of objects.

rather widely accepted.[8] First, truthmakers are entities or pluralities of entities. We need not assume that every plurality of entities constitutes a complex entity. Second, items made true are interpreted sentences, which I call statements. Instead of statements, one might consider items made true propositions. I consider propositions problematic postulations, but much does not depend on this matter here. Thirdly, the existence of the truthmaker necessitates the truth of the statement made true. Thus, the intuition that the statement is true because its truthmaker exists is (at least partially) cashed out by means of necessitation.

The eliminativist nominalism denies that there are any such entities as natural kinds that would be referents of natural kind terms (function 4). It depends on the preferred nominalist ontological category theory what would be taken as truthmakers of the attributions of natural kinds to objects. For example, according to our trope theory SNT (the Strong Nuclear Theory) (Keinänen 2011; Keinänen & Hakkarainen 2014), only fundamental objects are trope bundles. Necessarily, if they exist, they have certain nuclear tropes as their proper parts. These nuclear tropes are also truthmakers of the claim that the corresponding object belongs to a certain natural kind. For instance, certain determinate mass, electric charge and spin tropes make true the claim that a certain micro-particle is an electron.

We can present a similar explanation for why complex objects that have certain kinds of objects as their parts necessary to their existence belong to a natural kind. Roughly, the parts related in a certain way are sufficient truthmakers for the complex object belonging to a natural kind. Here we have a recourse to the relations between proper parts, which unify the parts into a complex object.[9] This kind of account of natural kind membership is microstructural. It might perhaps be applied to the natural kinds of microparticles, atoms and molecules. I have presented this account of kind membership as a view that is committed to de re necessities; according to it, the kind determining features and natural kinds would be necessary to the members of the natural kind. The eliminativist nominalism is not committed to this claim. Depending on the nominalist ontological category theory assumed, one might take some or even all kind-determining features as contingent to the members of a natural kind. In this case, we would have only de dicto necessities involved in our account: because of having certain kind determining features, an object belongs to a certain natural kind. Moreover, having these kind-determining features constitutes the necessary condition for application of the kind term to the object. Therefore, the kind-determining features are necessary for every member of the kind.[10]

[8] See, for instance, Rodriguez-Pereyra (2002), Maurin (2002) and Simons (2000). Some theorists (e.g., Rodriguez-Pereyra) assume that truthbearers and items made true are propositions, but I consider it as an additional commitment. Similarly, I will not consider here the later attempts to elucidate truthmaking by means of metaphysical grounding.

[9] See Mckenzie & Muller (2017) and Hendry (2023) for empirically motived answers to the special composition question: in which conditions objects related in a certain way constitute a complex object.

[10] The view that natural kinds are necessary to their instances has remained contestable. Brian Ellis (2001, sec.7.5), for instance, who advocates Neo-Aristotelian realism, denies the need to consider natural kinds necessary to their instances.

According to the eliminativist nominalism, there are no natural kinds as constituents of reality. Therefore, the above functions assigned to natural kinds are either left unoccupied or given to the entities the different nominalist ontological category theories might postulate. According to the SNT, the identity conditions of fundamental objects are determined by their nuclear tropes. Additionally, one might introduce rigid existential dependencies between complex objects and their proper parts and/or relations between the proper parts. These parts and/or relations between parts would be necessary to the complex objects and contribute to their individuation. My purpose here is only to illustrate how *certain* eliminativist nominalists can take care of function 3 assigned to natural kinds without postulating substantial kind universals, but not to argue for this proposal.

Similarly, according to eliminativist nominalists, we need not introduce kind universals to function as entities collecting fundamental properties, which are truthmakers of law statements (function 2). The eliminativist nominalist can maintain, for instance, that tropes or modes already play this role, and we need not postulate kind universals. However, there are difficult issues left to the nominalist in the metaphysics of laws, to which I cannot go to in the limits of this chapter. Finally, since not being *error-theoretic eliminativists* about kind-talk, eliminativist nominalists can accept the use of natural kind terms in scientific explanations and inductive generalizations (function 1).

In addition to natural kind terms having their application conditions fixed by the theoretical context in which they occur (like “hydrogen atom” or “water molecule”), there are natural kind terms such as “gold” and “water”, whose application conditions appear to be fixed by some manifest criteria. Again, eliminativist nominalists must provide us with an account of such natural kind terms that is consistent with rejection of natural kinds as separate entities. Moreover, it is preferable to adopt a semantic theory that avoids strong metaphysical implications and leaves the possible commitment to de re necessities to the preferred ontological category theory.

Jussi Haukioja’s (2012, sec. 3) theory of natural kind terms as *actuality dependent expressions* is the most promising available account fulfilling these constraints.[11] According to it, every general term has an *applicability role*, which specifies the criteria that an object or stuff must satisfy for a general term applying to the object. The applicability roles of such paradigmatic natural kind terms as “water” and “gold” are specified by the descriptions of the manifest features of an object or stuff. In the case of paradigmatic natural kind terms, the applicability role is realized by empirically. discovered features or a structure distinct from the manifest features. The latter determine the membership of an object or stuff in a natural kind and its manifest features (at least in the actual world). The paradigmatic natural kind terms are *actuality dependent* expressions: the same features/structure that actually realize their applicability role realize the applicability role in every possible world.

[11] Other theories of “rigidity” of general terms make more problematic ontological assumptions. For instance, Devitt (2005) commits himself to the view that natural kinds are necessary to their instances.

For instance, if a structure constituted by H_2O molecules and some other molecules and ions actually realizes the applicability role of the kind term “water”, the same structure realizes the role in every possible world.

Haukioja’s theory allows for alternative structures or groups of features realizing a single applicability role. Moreover, the realizers of an applicability role can contain extrinsic features of objects. The theory fits with the eliminativist nominalism about naturel kinds: instead of kind universals, there are groups of features or structures of entities which are truthmakers of attributions of natural kinds to objects and stuffs. If objects having certain features or a certain kind of structure are the only truthmakers of attribution of natural kind K to the object, these features or structure are necessary to the members of K. We need not introduce any kind essences or de re necessities to explain this, but all work is done actuality dependence; if an object having certain features is the realizer of the applicability role, this type of object realizes the role in every possible world.[12]

Conclusion

According to the eliminativist nominalist view of natural kinds, there are no natural kinds. Since there are no natural kinds, there are no natural kind essences or de re necessary properties of natural kinds. There is true general talk about the members of natural kinds and classifications of objects with the help of natural kind terms, which track mind-independent divisions. The nominalist theory is not committed to the necessity of natural kinds to their members; the nominalist ontological category theories into which the theory is integrated might be so committed, at least in some cases. Nevertheless, in many cases, the source of the necessity of certain features to the members of a natural kind might be only the fact that the kind term tracks objects having certain features or a certain kind of structure in every possible world.

The nominalist theory stresses the epistemic and explanatory functions of natural kinds and natural kind classifications (function 1). By contrast, the metaphysically heavy functions of collecting the necessary properties of the members of the kind (function 2) and determining the identity conditions of objects (function 3) are taken care of the nominalist basic ontologies. Because of its flexibility, this nominalist view of natural kinds interlocks well with the new theory of reference Panu Raatikainen (2020, 2021) defends.

[12] Haukioja (2015) adopts a conferralist view of the necessity of certain kind-determining features to the members of the kind. Roughly, because of our dispositions to use kind term, we also confer the necessity of certain features to the members of a natural kind. From the eliminativist nominalist perspective, this conferring is not required because there are no natural kinds.

References

Beebee, Helen & Sabbarton-Leary, Nigel (2010): 'Introduction', in *The Semantics and Metaphysics of Natural Kinds*, H. Beebee and N. Sabbarton-Leary (eds.), London: Routledge, 1–24.

Bird, Alexander (2007): *Nature's Metaphysics,* Oxford: Oxford University Press.

Bird, Alexander & Tobin, Emma (2022): 'Natural Kinds', *The Stanford Encyclopedia of Philosophy* (Spring 2022 Edition), Edward N. Zalta (ed.), URL = http://plato.stanford.edu/archives/spring2022/entries/natural-kinds/.

Boyd, Richard (1999): 'Homeostasis, Species, and Higher Taxa', in *Species: New Interdisciplinary Essays*, R. Wilson (ed.), Cambridge, MA: MIT Press, 141–186.

Boyd, Richard (2010): 'Realism, Natural Kinds, and Philosophical Methods', in *The Semantics and Metaphysics of Natural Kinds*, H. Beebee & N. Sabbarton-Leary (eds.), London: Routledge.

Devitt, Michael (2005): 'Rigid Application', *Philosophical Studies* 125(2): 139–165. URL =https://www.jstor.org/stable/4321625.

Ellis, Brian (2001): *Scientific Essentialism,* Cambridge: Cambridge University Press.

Ereshefsky, Marc (2010): 'What's Wrong with the New Biological Essentialism', *Philosophy of Science* 77(5): 674–685. URL = https://doi.org/10.1086/656545.

Haukioja, Jussi (2012): 'Rigidity and Actuality-Dependence', *Philosophical Studies* 157(3): 399–410. URL = https://doi.org/10.1007/s11098-010-9660-7.

Haukioja, Jussi (2015): 'On Deriving Essentialism from the Theory of Reference', *Philosophical Studies* 172(8): 2141–2151. URL = https://doi.org/10.1007/s11098-014-0402-0.

Hawley, Katherine & Bird, Alexander (2011): 'What are Natural Kinds?', *Philosophical Perspectives* 25(1): 205–221. URL = https://doi.org/10.1111/j.1520-8583.2011.00212.x.

Heil, John (2012): *The Universe as We Find It*, Oxford: Oxford University Press.

Hendry, Robin (2006): 'Elements, Compounds, and Other Chemical Kinds', *Philosophy of Science* 73(5): 864–875. URL = https://doi.org/10.1086/518745.

Hendry, Robin (2023): 'Structure, Essence and Existence in Chemistry', *Ratio* 36(4): 274–288. URL = https://doi.org/10.1111/rati.12387.

Keinänen, Markku (2011): 'Tropes – the Basic Constituents of Powerful Particulars?', *Dialectica* 65(3): 419–450. URL = https://doi.org/10.1111/j.1746-8361.2011.01276.x.

Keinänen, Markku (2015): 'A Trope Nominalist Theory of Natural Kinds', in G, Guigon, & G. Rodriguez-Pereyra (eds.), *Nominalism about Properties*, London: Routledge, 156–174.

Keinänen, Markku & Hakkarainen, Jani (2014): 'The Problem of Trope Individuation: A Reply to Lowe', *Erkenntnis* 79(1): 65–79. URL = https://doi.org/10.1007/s10670-013-9459-y.

Keinänen, Markku & Hakkarainen, Jani (2024): 'Trope Bundle Theories of Substance', in A.R.J. Fisher and A-S. Maurin (eds.), *The Routledge Handbook of Properties*, London: Routledge, 239–249.

Kitcher, Philip (1984): 'Species', *Philosophy of Science* 51(2): 308–333. URL = https://doi.org/10.1086/289182.

Kripke, Saul (1980): *Naming and Necessity,* Cambridge: Cambridge University Press.

Loux, Michael (1978): *Substance and Attribute*, Dordrecht: D. Reidel.

Lowe, Edward Jonathan (1998): *The Possibility of Metaphysics*, Oxford: Clarendon Press.

Lowe, Edward Jonathan (2003): 'Individuation', in M. Loux & D. Zimmerman (eds.), *The Oxford Handbook of Metaphysics*, Oxford: Oxford University Press, 75–95.

Lowe, Edward Jonathan (2006): *The Four-Category Ontology*, Oxford: Clarendon Press.

Lowe, Edward Jonathan (2009): *More Kinds of Being – A Further Study of Individuation, Identity and the Logic of Sortal Terms,* Oxford: Wiley-Blackwell.

Lowe, Edward Jonathan (2015): 'In Defence of Substantial Universals', in G. Galluzzo & M.J. Loux (eds.), *The Problem of Universals in Contemporary Philosophy*, Cambridge: Cambridge University Press, 65–84.

Ludwig, David (2018): 'Letting Go of "Natural Kind'. Towards a Multidimensional Framework of Non-Arbitrary Classification', *Philosophy of Science* 85(1): 31–52. URL = https://doi.org/10.1086/694835.

Maurin, Anna-Sofia (2002): *If Tropes*, Dordrecht: Kluwer.

McKenzie, Kerry & Muller, F. A. (2017): 'Bound states and the special composition question', in M. Massimi, J. W. Romeijn, and G. Schurz (eds.), *EPSA15 selected papers: The 5th conference of the European Philosophy of Science Association in Düsseldorf*, Dordrecht: Springer, 233–242.

Needham, Paul (2000): 'What is water?', *Analysis* 60(1): 13–21. URL = https://doi.org/10.1093/analys/60.1.13.

Putnam, Hilary (1975): 'The Meaning of "Meaning"', in K. Gunderson (ed.), *Language,Mind and Knowledge*, Minneapolis: University of Minnesota Press, 131–193.

Quine, Willard van Orman (1969): 'Natural Kinds', in *Ontological Relativity and Other Essays*, New York: Columbia University Press, 114–118. URL = https://doi.org/10.7312/quin92204-006.

Raatikainen, Panu (2020): 'Theories of Reference: What was the Question?', in A. Bianchi (ed.), *Language and Reality from a Naturalistic Perspective*, Cham: Springer, 69–103.

Raatikainen, Panu (2021): 'Natural Kind Terms Again', *European Journal for Philosophy of Science* 11(1): 1–17. URL = https://doi.org/10.1007/s13194-020-00344-3.

Rodriguez-Pereyra, Gonzalo (2002): *Resemblance Nominalism: Solution to the Problem Universals*, Oxford: Oxford University Press. URL = https://doi.org/10.1093/mind/fzi457.

Simons, Peter (2000): 'Truth-maker Optimalism', *Logique et Analyse*, 43(169–170): 17–41.

19
On the irrelevance of freedom to the causal relevance of will

Renne Pesonen

Introduction

Assume that there is no free will. Does it follow that everyone can do as they please, since no one can be held accountable for their actions? Many find the affirmative answer intuitive. Some of them embrace moral nihilism, but I believe most philosophers who consider the affirmative answer correct argue that we must assume the existence of free will because we indeed are morally accountable for our actions.

There is also an analogous—and maybe deeper—question concerning the ontological status of actions themselves: If there is no free will, does it follow that no intentional action is possible, since our behavior is predetermined, perhaps by the laws of nature, regardless of our beliefs, desires, and reasons? Or, to rephrase: if intentional action is indeed possible, does it follow that there is free will? In this essay, I examine two views that accept this inference but arrive at opposing stances on the question of free will.

The first view denies the existence of free will based on reductive physicalism. Since everything is physical and every event is determined by the laws of physics, there can be no true causal principles or powers that transcend the causal closure of the physical. Any putative higher-level causation would violate the fundamental

Pesonen, R. (2025). On the irrelevance of freedom to the causal relevance of will. In A. Laitinen, M. Keinänen, J. Reinikainen & A. Honkasalo (eds.). *Language, Truth, and Reality*. Tampere University Press. (pp. 237–247). https://doi.org/10.61201/h5fbdk64

laws of nature. In particular, free will is impossible because mental causation cannot exist. I call this stance *metaphysicalism*. Its adherents may claim that their view is based on natural science, but it is a theory of metaphysics, rather than physics, that licenses their inferences. The gist of their view is the causal closure of the physical and the reducibility of every thing, property, and event to physics. However, it does not ultimately matter which theory of physics is correct and what the actual laws are—for example, whether they are deterministic or not.

I call the second view *intentional realism*, which refers to realism about the ontology of intentional ascriptions. Intentional realists rely on science, but it is the behavioral, cognitive, and social sciences rather than physics. According to them, intentional ascriptions may not constitute proper scientific laws, but they are nevertheless indispensable for explaining and predicting human behavior. Moreover, if our best theories of psychology, economics, sociology, and so on, require that people make and enact choices among alternative possibilities, we are licensed to presume the existence of free will (or its psychological cognates). Thus, intentional realists are scientific realists who may or may not subscribe to the thesis of causal closure of the physical. What they insist on is that mental or intentional states are not superfluous but have explanatory and predictive relevance.

Below, I formulate a defense of intentional realism based on standard arguments for anti-reductive physicalism and the autonomy of the special sciences. While the argument salvages the causal efficacy of will from the ultimately absurd and militantly anti-scientific attack of metaphysicalism, it implies nothing about the question of freedom in terms of the metaphysical possibility of genuine alternatives. I argue that the metaphysics of freedom should be disentangled from the scientific questions concerning the role of will (or related mental states) in intentional explanations. Whether or not the will is free in any metaphysically or morally relevant sense is ultimately irrelevant to explanations and predictions that invoke mental causation.

Reasons for anti-reductionism

Once we look at the actual theories and practices of science, it immediately becomes evident that metaphysicalism is an anti-scientific ideology. Even if everything is material and governed by the laws of nature, the theories and methods of physics cover only a small portion of legitimate scientific questions. If you want to know whether rising interest rates mitigate inflation or whether schizophrenia is heritable, you need other methods and theories to answer these questions. Denying the reality of these phenomena, or the legitimacy of any means beyond physics to investigate them, clearly represents a doctrine of militant anti-science. The next section discusses psychological explanation and metaphysicalism. However, let’s first take a glance at scientific reductionism, which is a less militant and more general program in comparison to metaphysicalism.

Back in the day, positivists argued that it should be possible to reduce the entities and laws of "higher-level" sciences, such as biology, to lower and more fundamental levels, such as chemistry, and ultimately all the way down to fundamental physics (e.g., Oppenheim & Putnam, 1958, Nagel, 1961). This branch of reductionism did not necessarily deny the existence of higher-level entities or laws. Rather, it aimed to show that, in principle, they could be derived from, or explained by, the laws of physics. However, in recent decades, few philosophers of science have subscribed to this program. Most of us are *non-reductive* physicalists now.

Non-reductive physicalists believe in metaphysical materialism and the causal closure of the physical, but they also maintain that there are higher-level (such as emergent or supervenient) properties and causal regularities that cannot be defined or derived from theories of physics. For example, a 50-euro bill is obviously a physical thing, but its value or status as legal currency does not stem from its physical properties. Every financial transaction is a physical event, but there are unlimited ways to physically implement these transactions—whether through gold coins, fiat bills, exchanges of information in computer networks, or whatever the future may bring. Moreover, aggregates of those transactions form higher-level social and economic patterns that sometimes can be predicted and manipulated. The fact that prices tend to go up as demand increases is not something that can be derived from fundamental physics. Furthermore, we do not need physics to investigate whether this pricing trend can be reversed by increasing production. Discoveries concerning the laws of fundamental physics will almost certainly have no impact on theories in fields such as macroeconomics, population genetics, or psychology.

In a nutshell, the above is the standard argument put forward, for example, by Fodor (1974, 1997), in support of non-reductive physicalism and the autonomy of the "special" sciences. While the argument is traditionally formulated in terms of theories and laws, this is not necessary. Instead, it is common to conceive of scientific explanations in terms of variables and dependencies between their values (Woodward, 2000): If you change the value of variable X, the value of Y tends to change. If this relationship is an established invariance, you can predict and potentially explain the values of Y based on changes in X. For example, if you increase the intake of vitamin C in a malnourished population (or the production of a good in a market), the incidence of scurvy (or the price of the good) decreases. Using such regularities for prediction and scientific explanation does not require them to be laws in any strict sense. Non-reductive physicalists further argue that these kinds of causal generalizations cannot be reduced to lower-level sciences if the higher-level mechanisms and properties that make them work can be physically realized in multiple ways.

Functional explanations in psychology

Since the advent of functionalism in the philosophy of mind (Fodor, 1968; Putnam, 1967; Block & Fodor, 1972), it has been commonplace to hold that psychological states are prime examples of multiply realizable phenomena. Our beliefs and desires are somehow realized in our brains, but they also have characteristic consequences for both mental and overt behavior, which can be identified and investigated without knowledge of their physical realization. According to functionalists, it is these characteristic consequences, identified at the intentional rather than the physical level of causation, that make them instances of psychological states such as beliefs, desires, and so on.

Following Dennett's (1971) classic analysis, consider, for example, a computer running a chess program. If the program's behavior is rational enough, we can explain its moves simply by referring to the strategies and rationales of the game. The same qualitative behavior can be implemented by infinitely many algorithms, and we do not need to know the exact algorithm to explain or predict the moves. Furthermore, the same algorithm can be executed on physically vastly different machines, and a complete description of the machine would not usually help in predicting the program's behavior, simply due to its excessive complexity. Likewise, with humans, we do not need to know much about the exact mental machinery, and even less about the brain, to predict and explain people's everyday rational behavior. Intentional explanation is justified purely by explanatory and pragmatic necessity. Moreover, there is no extra mystery in the relationship between the mind and the brain compared to that between a program and the machine.

With these teachings in mind, I discuss an example borrowed from Raatikainen (2010): Suppose John desires a bottle of beer and believes that there is some in the fridge. Without any knowledge of his brain, we can safely bet that, soon enough, he will head to the fridge. However, before John gets a chance to get there, we tell him that we already drunk all the beer and the fridge is empty. Hence, John's belief changes from "there is some beer in the fridge" to "there is no beer in the fridge." This intervention changes his behavior from "go to the fridge" to "go to the store to get more beer." While changes in John's beliefs and behavior certainly implicate changes in his brain states, we can explain the change in his behavior solely in terms of changes in his beliefs. We have absolutely no information on exactly how his brain was affected, and we do not need that information in order to make inferences about his behavior.

Raatikainen (2010) argued that since psychological states are multiply realizable, we could, in principle, manipulate John's brain states without manipulating his beliefs concerning the beers and the fridge. Such an intervention would not affect John's decisions and behavior, while an intervention on his relevant beliefs would. What follows is that, surprisingly, it is the changes in intentional states that explain change in his behavior, not changes in the brain!

I think that it is largely correct. However, Raatikainen uses the thesis of the multiple realizability of the mental in a slightly non-standard and potentially problematic way. The standard view holds that mental states and processes could, in principle, have vastly different realizations *beyond* the (typical) human brain. This is because, if mental states are identified functionally—based on their causal properties concerning perception, action, and other mental states—similar causal networks could, in principle, be realized in vastly different brains or even without brain tissue at all, for example, in silicon-based life forms or robots. This is not the same as claiming that changes in an individual brain could occur without changes in behavior. Therefore, one could accept the antecedent of the argument without accepting the consequent, because not just any intervention on the brain should count as relevant, regardless of whether mental states are multiply realizable.

We could, for example, manipulate the firing rates of some random neurons at the periphery of John's visual cortex, and no one expects this would change his behavior. Therefore, we need a specification of which interventions count as relevant. Raatikainen (2010, 359) argues that, in the given example, the only route through which we can change John's behavior is by altering his beliefs. I think this is almost correct. However, if we allow direct interventions in John's brain, it should be possible to stimulate his motor areas to make him head for the grocery store instead of the fridge without altering his beliefs or desires. Hence, behavior may not always align with beliefs and desires, and it could be argued that the brain is causally explanatory for John's behavior after all.

However, functionalists identify psychological states with *patterns* of behaviors and inferences. If we only hijack John's motor cortex, he presumably still believes there is beer in the fridge and remains disposed to make inferences based on that belief. He may still want to go to the fridge, but his behavior is now under external control, and he is acting in a way that he cannot rationally explain. Our intervention has interfered with the normal functioning of John's brain and mind. When this happens, we can no longer explain his behavior in terms of intentional states because their normal causal functions no longer exist. Such an intervention would not count as an argument against the causal efficacy of mental states, as the relevant causal mechanisms are severed.

What if we hijack a part of John's brain so that we shift the entire *pattern* of behavior and inferences stemming from his beliefs about the beer and the fridge? That sort of intervention surely counts as relevant. However, according to functionalists, such an intervention would amount to altering John's beliefs, because beliefs are identified precisely by such patterns. That intervention would not rob John of his capacity to act freely or at least rationally. We routinely create such interventions by simply telling people that there is no more beer in the fridge and so on. This slight alteration of Raatikainen's (2010) argument, based on the functionalist theory of mental states rather than on mere multiple realizability, retains its original conclusion.

However, not everyone believes in functionalism, and perhaps we could dispense with belief/desire explanations altogether. Maybe we could scan John's brain and use theoretical calculations (based on some future neuroscience) to make even better predictions about his behavior. Well, perhaps we could, yet we don't. We know how intentional explanations work, but we have absolutely no idea how to predict complex behavior from brain activity, let alone from fundamental physics. Commonsense belief/desire ascriptions are surely too blunt an instrument for serious scientific psychology, but we routinely rely on them for quotidian and social scientific explanations. Swaths of cognitive psychology, behavioral economics, and related fields operate on the intentional level of explanation, even when they attack our folk psychological platitudes. I am not here defending functionalism as such but scientific realism in the realm of intentional explanation.

I close this section by outlining what the metaphysicalist alternative would be. What would John do if he believed that there was beer in the fridge that he desires? That depends on the laws of nature and the elementary particles that comprise his body. The particles do not care about John's desires, so we would need to know the complete physical description of his body and perform extremely complex calculations to predict what is going to happen. I am not sure if such predictions are possible even in principle, but that is the only option available for metaphysicalists. Furthermore, once we tell John that the fridge is empty, what would follow? The imagined situation leaves John's physical state and the impact of the intervention completely unspecified. Therefore, according to metaphysicalists, absolutely no predictions follow, and nothing in the given description could explain John's behavior.

Consistent metaphysicalists should not even care about arguing that it is the brain that drives behavior. Brain processes are surely physical, but for metaphysicalists, there are no higher-level properties or causal regularities to be identified beyond fundamental physics. In the end, this tenet does not render only special sciences impossible but science itself, including physics. If metaphysicalism is true, no one would conduct experiments for epistemological reasons or accept hypotheses for rational reasons. In fact, metaphysicalists themselves would not hold their beliefs because it is the rational thing to do, but simply because things turned out that way.

So much for the metaphysicalism. This section conveys three main points: (1) "The brain made me do it" may, in a sense, always be a correct answer when you need to explain your actions. However, it almost never is the only correct option or the most informative one. (2) The debates between intentional versus other scientific explanations of behavior (such as brain centered explanations) are not the same as the debate between metaphysicalism and intentional realism. This is further discussed below. (3) Nothing said thus far bears on the question of freedom in terms of metaphysically possible alternatives. Intentional states can be causally efficacious or relevant even if there is no freedom in that sense. With these conclusions in mind, I next discuss potential misconceptions about determinism, consciousness, and external influences in debates concerning the freedom of will.

Some implications of the argument

Some enemies of free will mount their attack from neuroscience or biology. It is your brain or genes calling the shots, so your thoughts and will are mere illusions, or at least their presumed significance for your behavior. These arguments may look much like metaphysicalism, but instead of radical materialism, they rely on behavioral sciences to argue that our actions are determined by external or non-conscious factors beyond our control. While I cannot discuss these debates in detail here, I will briefly address some confusions they harbor concerning these factors in relation to intentional explanations. I believe that at least some misunderstandings can be straightened out simply by disentangling questions concerning freedom from an entirely different question about the role of will or mental states in psychological explanations.

The determinism/non-determinism dimension

Incompatibilists believe that free will requires the universe to be indeterministic, for if every event is determined by the laws of nature, there is simply no room for the will to operate. This topic veers back into metaphysical debates, on which I have nothing more to say. However, if we accept the autonomy of intentional explanations, we can have indeterminism without metaphysics, simply because intentional ascriptions are not fundamental laws but probabilistic generalizations.

For example, if we tell John that there is no beer in the fridge, nothing in the intentional description determines what he will do next. Maybe he still goes to the fridge to check. Maybe he skips the trip to the grocery because he doesn't bother. Typically, we use intentional ascriptions for explanation rather than prediction, and often they serve merely as *post hoc* rationalizations (see Cushman, 2019). However, this does not mean that they cannot factor into legitimate scientific explanations.

The aims of science are diverse, and the aims of explanation and understanding are not always aligned with the aim of prediction (Potochnik, 2015). That is why many explanatory models in science abstract and idealize. What intentional ascriptions capture are not laws; they provide only an approximate model for explaining human behavior, which is highly patterned but still not deterministic at the intentional level of description. Reasons, instincts, norms, and so on may be in conflict and they rather motivate than determine decisions. Perhaps what we experience as freedom of action is simply the result of a host of variables that render any particular action practically unpredictable. At any rate, indeterminism in the domain of psychology does not imply any metaphysical commitments, but neither does it imply randomness or a lack of control.

The internal/external dimension

Perhaps it is our genes or the environment that truly determines our actions. Indeed, instincts, habits, and social norms may often explain our behavior better than our conscious volitions (Cushman, 2019), and factors such as mental illness or coercion can also rob us of the possibility to act freely. An unquestionably important fact is that many factors external to our conscious intentions guide our behavior, and we are not always as free as we believe or want to be.

However, these considerations do not imply anything fundamental about human freedom, except that the extent to which our choices are free is partly an empirical and partly a moral consideration. Many of the factors mentioned simply modulate our intentions, and they are comparable to beliefs and desires in that they affect our behavior without strictly determining it on each occasion. At gunpoint, you may still be free to make the rational decision to give up your wallet before your life. Your chess moves or your next steps when the fridge proves empty may be strongly habitual, but they are not strictly determined by your genes or your past any more than each move of a chess machine is hard-coded into its program.

I believe the future of behavioral sciences will push the boundaries of what we can predict and explain, and this will have consequences for people's intuitions about which behaviors are free and which are determined beyond our control. The point I want to make here is that the debate about the relative importance of external versus internal factors in the explanation of behavior is not the same as the debate between metaphysicalists and intentional realists. Metaphysicalists cannot even differentiate between internal and external causes because, for them, there are no identifiable higher-level determinants of behavior. For intentional realists, the question about the relative importance of external and internal factors is not a fundamental or moral question about human freedom but an empirical one concerning the causal variables involved in intentional explanations.

The consciousness/non-consciousness dimension

Finally, some regard consciousness as crucial for free will (e.g., Hodgson, 2012). I argue that it actually isn't, at least insofar as free will pertains to decision-making and cognitive control.

The most famous example of the dominance of unconscious over conscious brain processes is the Libet experiment (Libet et al., 1983). It demonstrated that a brain signal can be reliably detected before subjects experience a conscious volition to move their arm. Thus, in this simple task, the brain makes the decision first, and apparently the conscious intention follows. However, this is not very alarming.

Behavioral and decision scientists have held for decades that human decision-making is largely intuitive and often eludes conscious control. However, this does not mean we lack control or freedom over our decisions. For example, an influential dual-process model by Kahneman (2003) posits that unconscious processes make automatic decisions that we rely on during routine activities. The function of conscious processes is to monitor these decisions and ongoing behavior, intervening when there is a reason to do so. Therefore, simple and trivial decisions that do not require planning or control, such as those investigated by Libet et al., are expected to stem from non-conscious processes. The conscious mind simply monitors what is happening and exercises regulatory control when necessary. Importantly, dual-process and related theories do not conceptualize the difference between non-conscious and conscious processes in terms of brain versus mental processes. Non-conscious processes may be automatic but still intentional. This is particularly clear when automatization results from habituation during social or skill learning.

But can an entirely unconscious mechanical system, such as artificial intelligence, make decisions? I don't see why not. The entire field of reinforcement learning investigates decision-making and learning in artificial agents, with many methods inspired by psychological and biological learning theories (Sutton & Barto, 2018). As far as I can tell, it should be possible to model human decision-making as accurately as we wish. It might be natural to view these non-conscious systems as mere complex automata, incapable of doing anything beyond what their programs dictate. However, if the human mind surpasses the capabilities of mere machines, consciousness is probably not among the reasons. The notion of consciousness used by dual-process and related decision theorists refers to access consciousness (see Block, 1995) rather than subjective experience. Access consciousness is a functional concept involved in metacognition. As such, it should be amenable to causal description like any other mental function.

For example, Hodgson (2012) believes that consciousness does not conform to any rules or laws, but it may contribute to free decision-making, for example, by resolving inconclusive reasons. If this means a capacity to form an arbitrary choice, it seems to me that the same function can be implemented simply by drawing decisions randomly from some appropriate distribution.

In conclusion, the fact that some action-guiding processes are not conscious does not mean that they are mere mindless brain processes. Moreover, consciousness is not essential for the ability to form and exercise one's will, insofar as it refers to the ability to rationally choose one's goals and actions.

Conclusion

I have argued that the problem of free will involves two separate questions. On the one hand, it involves the mind–body problem, which, in this context, boils down to the question of the causal relevance of mental states. On the other hand, there is the question of whether the will can be truly free. There are scientifically respectable answers to the first problem, and metaphysicalism is not among them. To me, the second question appears to be primarily moral or metaphysical. I do not wish to imply that the question is therefore meaningless, but rather that it should be disentangled from the question of the causal efficacy of the will. The remaining scientific question is not whether the will is free, but whether it plays a role in scientifically respectable explanations of behavior.

But does it? You rarely encounter the term "will" in the literature of cognitive or behavioral sciences. For the purposes of this essay, I have considered it to be a folk-psychological abstraction that captures aspects of executive function, such as decision-making, goal selection, and cognitive control. As with other folk-psychological notions, such as beliefs and desires, it serves only as an approximate but heuristically useful description of the determinants of intentional behavior. If "will" cannot be given any psychologically meaningful functional interpretation, it is futile to debate whether it is free or real, as it would lack any explanatory relevance anyway.

I suspect that arguments from genes and the brain against the existence of free will appeal to some because they reveal the human mind to be a physical system, after all. However, functionalists do not deny this. Their point is simply that human behavior can also be described in terms of intentional agency, where causal variables are identified based on their causal properties in cognition and behavior rather than their intrinsic physical properties. It is all about levels of description that justifies the use of intentional explanations. Dennett's (1971) argument for intentional interpretations can be criticized on the grounds that it justifies anthropomorphizing machines when it is convenient to think of them as intentional agents. But surely we cannot be guilty of anthropomorphizing humans!

Hence, it is futile to argue against intentional realism on the basis that science shows the mind to be a complex causal mechanism. At least functionalists already believe this. For them, science is there to uncover the nature and exact mechanisms of mental functions, often correcting our preconceived folk-psychological ideas about them. As research progresses, it may very well chip away at our intuitive belief in human freedom. However, no harm is necessarily done. The facts uncovered thus far do not justify metaphysical nihilism or biological determinism; rather, they only place restraints on excessively libertarian conceptions of human freedom and reason.

References

Block, Ned (1995): ‘On a confusion about a function of consciousness’, *Behavioral and Brain Sciences,* 18(2): 227–247. URL = https://doi.org/10.1017/s0140525x00038188.

Block, Ned & Fodor, Jerry (1972): ‘What Psychological States are Not’, *The Philosophical Review,* 81(2): 159–181. URL = https://doi.org/10.2307/2183991.

Cushman, Fiery (2019): ‘Rationalization is rational’, *Behavioral and Brain Sciences,* 43*(*28). URL = https://doi.org/10.1017/S0140525X19001730

Dennett, Daniel (1971): ‘Intentional Systems’, *The Journal of Philosophy,* 68(4): 87–106. URL = https://doi.org/10.2307/2025382.

Fodor, Jerry (1968): *Psychological Explanation: An Introduction to the Philosophy of Psychology,* New York: Random House.

Fodor, Jerry (1974): ‘Special sciences (Or: The Disunity of Science as a Working Hypothesis)’, *Synthese,* 28(2): 97–115. URL = https://doi.org/10.1007/bf00485230.

Fodor, Jerry (1997): ‘Special Sciences: Still Autonomous After All These Years’, *Noûs,* 31(S11), 149–163. URL = https://doi.org/10.1111/0029-4624.31.s11.7.

Hodgson, David (2012): *Rationality + Consciousness = Free Will,* Oxford: Oxford University Press.

Kahneman, Daniel (2003): ‘A Perspective on Judgment and Choice: Mapping Bounded Rationality’, *American Psychologist,* 58(9), 697–720. URL = https://doi.org/10.1037/0003-066X.58.9.697.

Libet, Benjamin, Gleason, Curtis, Wright, Elwood & Pearl, Dennis (1983): ‘Time of conscious intention to act in relation to onset of cerebral activity (readiness-potential): The unconscious initiation of a freely voluntary act’, *Brain,* 106(3): 623–642. URL = https://doi.org/10.1093/brain/106.3.623.

Nagel, Ernest (1961): *The Structure of Science: Problems in the Logic of Scientific Explanation,* New York: Harcourt, Brace & World.

Oppenheim, Paul & Putnam, Hilary (1958): ‘Unity of Science as a Working Hypothesis’, *Minnesota Studies in the Philosophy of Science* 2, 3–36.

Potochnik, Angela (2015): ‘The diverse aims of science’, *Studies in History and Philosophy of Science,* 53: 71–80. URL = https://doi.org/10.1016/j.shpsa.2015.05.008.

Putnam, Hilary (1967): ‘Psychological Predicates’, in William H. Capitan & Daniel Davy Merrill (eds.), *Art, Mind, and Religion*, Pittsburgh: University of Pittsburgh Press, 37–48.

Raatikainen, Panu (2010): ‘Causation, Exclusion, and the Special Sciences’, *Erkenntnis,* 73(3): 349–363. URL = https://doi.org/10.1007/s10670-010-9236-0.

Sutton, Richard & Barto, Andrew (2018): *Reinforcement Learning: An Introduction,* 2ndedition, Cambridge: The MIT Press.

Woodward, James (2000): ‘Explanation and Invariance in the Special Sciences’, *British Journal for the Philosophy of Science,* 51(2): 197–254. URL = https://doi.org/10.1093/bjps/51.2.197.

20
Mental and normative causation

Teemu Toppinen & Vilma Venesmaa

Like cats, we often do what we do simply because we want to. Sometimes we might even choose a certain course of action because we think that it is the right thing to do. Our joy tends to be contagious and may evoke smiles and laughter; on occasion we may wake up, in the middle of the night, due to a terrifying nightmare. The impact of the mental on the physical is a pervasive and a very familiar phenomenon. Yet it is also, of course, a very familiar fact that this phenomenon gives rise to a number of philosophical puzzles, some of which we often lump together under the heading of *mental causation*. In his work on the topic, Panu Raatikainen (2006, 2007, 2010, 2013, 2018) offers an account of mental causation, drawing on an interventionist approach to causation – developed, especially, in the context of philosophy of science – and on the idea that causal claims would carry an (often) implicit reference to contrast classes.

Raatikainen notes that while he focuses on the case of mental causation, his conclusions "are applicable across the board" in relation to the special sciences. This, it seems, would be a very nice feature of the account. However, we argue that if the kind of account that Raatikainen proposes is correct, then this has implications not just for the special sciences, but also for the case of the *normative*. The Raatikainen-style account makes it relatively easy for normative properties such as rightness, wrongness, being supported by the balance of reasons, having aesthetic merit, or being morally depraved, to have causal power. In particular, we suggest that the following conditional is true:

Toppinen, T. & Venesmaa, V. (2025). Mental and normative causation. In A. Laitinen, M. Keinänen, J. Reinikainen & A. Honkasalo (eds.). *Language, Truth, and Reality*. Tampere University Press. (pp. 248–260). https://doi.org/10.61201/h5fbdk64

> Conditional. If the interventionist account of mental causation of the kind that Raatikainen proposes is correct, then normative properties have causal power, even given a non-naturalist or a quasi-realist understanding of such properties.

We suggest, moreover, that normative properties should not turn out to have causal power given a non-naturalist or a quasi-realist construal of such properties. And so, the truth of Conditional turns out to be problematic for the style of account of causation that Raatikainen favors. We start by presenting Raatikainen's proposal with regard to how to make sense of mental causation (§§1–3). We then explain why the account generalizes to the case of the normative properties (§§4–5) and briefly conclude (§6).

§1. Following Raatikainen (2010, sec. 2; Raatikainen, on his part, follows Bennett 2007 here), we may approach the problem of mental causation in the context of what is known as *the exclusion problem*. Very briefly, the problem is that the following claims all seem very plausible, yet incompatible:

(1) Distinctness. Mental properties (and perhaps events) are distinct from physical properties (or events) (i.e., the type-identity theory is false).
(2) Completeness. Every physical occurrence has a sufficient physical cause.
(3) Efficacy. Mental events sometimes cause physical events and sometimes do so in virtue of their mental properties.
(4) No overdetermination. The effects of mental causes are not systematically overdetermined.
(5) Exclusion. No effect has (at a particular time *t*) more than one sufficient cause unless it is overdetermined.

Different solutions have been proposed to the problem (see Bennett 2007), but our focus is solely on the one offered by Raatikainen (2006, 2007, 2010, 2013, 2018).[1] Raatikainen's solution takes issue with the way the entire problem is set up. On his account, the exclusion problem is based on a confusion, really, or on a failure to adequately understand the notion of a cause. As a result, Raatikainen rejects – or at least refuses to accept – all of (2), (4), and (5), as these theses are formulated above. *Prima facie*, this of course doesn't look very promising. As noted, the whole problem arises because all of (1)–(5) are all very attractive theses. But Raatikainen's idea is to explain why we can let go of the relevant theses, once we properly understand the concept of causation. We next turn to Raatikainen's account of this concept.

[1] The references to Raatikainen's work, below, are all to the 2010 paper.

§2. Raatikainen's proposal is an application of an *interventionist* approach to causation.[2] The fundamental – and quite sensible – thought behind the interventionist approach is to start by asking what the point of our having the concept of causation is in the first place. What role or function does this concept perform? What are we up to when we make judgments about what causes what? The answer, very roughly, is that

> […] knowledge of genuine causal relationships is, *sometimes*, practical and applicable: by manipulating the cause we can influence the effect. If there is a real causal relationship between *A* and *B*, manipulating *A* is a way to change *B*; Mere correlation between *C* and *D*, on the other hand, just disappears if one attempts to affect *D* by manipulating *C* (p. 353).

The interventionist view "connects causal claims with counterfactual claims concerning what would happen to an effect under interventions on its putative cause" (p. 353):

> Roughly, *C* causes *E* if and only if an intervention on *C* would bring about a change in *E*. Slightly more exactly, causal claims relate, in this approach, variables, say *X* and *Y*, that can take at least two values. These may often be some magnitudes (such as temperature, electric charge or pressure), but in simple cases, they may also be just discrete alternative events or states of affairs. The idea now is that were there an intervention on the value of *X*, this would also result [in] a change in the value of *Y* (p. 353).

The interventionist view, then, provides truth-conditions of roughly the following kind for claims about what causes what:

> A change in *X* causes a change in *Y* if and only if, if *X* were to be changed by an intervention to such and such a value, the value of *Y* would change.

Raatikainen subscribes to the popular idea that "causal claims do not in fact describe a simple binary relation between two events, but rather involve (even if often only implicitly) a contrastive class for both cause and effect" (p. 354). The contrast classes are contextually determined, but given "default contrast classes," where the alternatives for *X* or *Y* having a certain value simply is their not having that value, (see pp. 354–355), we then get:

> *X*'s being x_1 (rather than not being x_1) causes *Y*'s being y_1 (rather than not being y_1) if and only if, if *X*'s being x_1 were to be changed by an intervention to *X*'s not being x_1, then *Y* would change from being y_1 to not being y_1.

[2] Raatikainen provides helpful references to the literature in which the interventionist theory has been developed, giving special credit to the work of James Woodward (see, e.g., Woodward 2003).

§3. According to Raatikainen, the interventionist account implies that a mental state can be a cause of behavior, and, given certain assumptions about the contrast classes, the physical state that underlies the relevant mental state may fail to be such a cause. Raatikainen (pp. 355–356) argues for this by appealing to the case of John who desperately wants a beer.[3] Happily, John remembers having earlier bought a six-pack and having placed it in the refrigerator. He then forms a belief that there is some beer in the refrigerator and consequently walks to the refrigerator. What causes John's walking to the refrigerator? Is it his belief that there is some beer in the fridge, or perhaps his brain state, *B*, at the moment – the one underlying his belief?

Given the default contrasts, it turns out that John's belief causes his behavior, but his brain state doesn't. Consider, first:

(i) If John's belief that there is beer in the refrigerator were to be changed by an intervention to not having the belief, he would not have gone to the refrigerator.

According to the standard possible world analysis of counterfactual conditionals, 'P → Q' is true if and only if either there is no P-world, or some P & Q -world is more similar to the actual world than any P & not-Q -world. In the case of (i), it is not the case that there are no P-worlds. But there are P & Q -worlds that are closer than any P & not-Q -worlds. Intuitively, the idea is that if we just change John's belief state, so that he does not believe that there is beer in the refrigerator, then he does not go to the refrigerator (but, say, goes to the grocery instead). If we would want to make it true that John's belief state changes, but that John nevertheless goes to the refrigerator (and not to the grocery, for example), then more changes would be needed to the way things actually are, and we thus move to possible scenarios that are further from the actual world. (i), then, plausibly comes out as true. Consider next:

(ii) If John's brain state *B* were to be changed by an intervention to his not having that state, he would not have gone to the refrigerator.

In the case of (ii), it is not true that all of the *P* and not-*Q* -worlds are further away from the actual world than some *P* and *Q* -world. Given that belief states are multiply realizable by different brain states, there are possible worlds in which John's brain state B is changed to a different brain state, B', but in which John nevertheless goes to the refrigerator. Also, these worlds would seem to be closer to the actual world than the worlds in which John's brain state is manipulated in the relevant way, but he goes to the grocery, say, instead of heading to the refrigerator. And so, (ii) plausibly is not true.

According to Raatikainen's interventionist proposal, again:

[3] Raatikainen notes that his argument has been inspired by Tim Crane's (2001) "similar argument with respect to a more traditional counterfactual approach to causation" (p. 358, n. 15).

> *X*'s being x_1 (rather than not being x_1) causes *Y*'s being y_1 (rather than not being y_1) if and only if, if *X*'s being x_1 were to be changed by an intervention to *X*'s not being x_1, then *Y* would change from being y_1 to not being y_1.

In the light of the above, if we now replace X's being x_1 with John's believing that there is beer in the refrigerator, X's not being x_1 with John's not believing that there is beer in the refrigerator, Y's being y1 with John's going to the refrigerator, and Y's not being y_1 with John's not going to the refrigerator (but going to the grocery instead), we get the result that John's belief causes his going to the refrigerator. By the interventionist analysis, the causal power of John's belief is vindicated.

By contrast, if we replace X's being x_1 with John's being in a certain brain state and X's not being x_1 with John's not being in this brain state (but in some other brain state instead), we get the perhaps somewhat surprising result that John's brain state does not cause his action. This is not to say, Raatikainen emphasizes, that John's being in the relevant state is not a *sufficient condition* for his performing the action in question. However, by Raatikainen's analysis, it is not a *cause* of his action: "Being sufficient condition for the occurrence of something, and being its difference-making cause, must thus be clearly distinguished" (p. 358).

Raatikainen notes that the argument that he gives "certainly deserves, and requires, further elaboration" (p. 358). For instance, in order for John's belief to cause his action, it must be the case that the alleged potential interventions are *genuine interventions*. They must not directly cause the change in John's behavior, for example. So, the mere truth of conditionals such as (i) and (ii) does not, strictly speaking, suffice for establishing the causal power of John's belief. The account is, in fact, more complicated. But Raatikainen suggests that the extra complexities will not cause any trouble for his argument (for a brief discussion of this, see pp. 358–359). We see no reason to suspect that this would not be so, and grant this assumption here. That is, we grant that by the standards of the interventionist theory, John's belief gets to cause his going to the refrigerator.

What is essential, for our purposes here, is that Raatikainen's account vindicates the causal efficacy of the mental. But as Raatikainen's account has been presented in the context of the exclusion problem, we may briefly note what Raatikainen takes to be the implications of his proposal in relation to the problem. Let us consider, then, in the light of Raatikainen's view, the exclusion problem again. In particular, consider (2) and (5):

> (2) Completeness. Every physical occurrence has a sufficient physical cause.
>
> (5) Exclusion. No effect has (at a particular time *t*) more than one sufficient cause unless it is overdetermined.

According to Raatikainen (p. 360), both of these assumptions involve confusing causes with sufficient conditions:

> There are causes, which are difference-makers; and there are sufficient conditions, which are wholly different issues and not causes of any sort; there are no such things as *sufficient causes*. Hence, I do not think that these two assumptions are so much false (or true) as mongrels based on a conceptual confusion which fail to make clear sense.

If we try to reformulate these theses in terms of difference-making causes, the resulting theses turn out to be false. The revised version of (2), for instance, would state that every physical occurrence has a physical difference-making cause, but Raatikainen suggests that his example of John establishes that this is mistaken (p. 361). (Raatikainen (p. 360) suggests that (4), or the thesis that the effects of mental causes are not systematically overdetermined, also involves confusion, but we shall not delve into this issue here).

However, again, what is really essential, for our purposes, is the way in which Raatikainen's account allows for the possibility of the mental getting some real causal work done. We next turn to metaethics and address the way in which Raatikainen's ideas generalize to the realm of the normative.

§4. Moral properties plausibly play a role in causal *explanations*. A helpful pair of examples may be lifted from a classic paper by Nicholas Sturgeon. First, we'll play the Nazi card: it is plausible that Hitler initiated a world war and ordered the "final solution" at least in part because he was morally depraved (Sturgeon 1984, p. 249). Second:

> An interesting historical question is why vigorous and reasonably widespread moral opposition to slavery arose for the first time in the eighteenth and nineteenth centuries, even though slavery was a very old institution; and why this opposition arose primarily in Britain, France, and in French- and English-speaking North America, even though slavery existed throughout the New World. There is a standard answer to this question. It is that chattel slavery in British and French America, and then in the United States, was much *worse* than previous forms of slavery, and much worse than slavery in Latin America (Sturgeon 1984, p. 245).

The second explanation, at least, is undoubtedly controversial. But this is irrelevant. What these examples are meant to illustrate is simply the fact that in giving causal explanations of various events, it is sometimes quite natural to appeal to certain things having certain normative (e.g., moral) properties. This would seem to be so because it is plausible that counterfactuals such as the following are true:

> HITLER. Had Hitler not been morally depraved, he would not have initiated a world war and ordered the "final solution."

> SLAVERY. Had the slavery in British and French America, and then in the United States, not been worse than previous forms of slavery, and slavery in Latin America, vigorous and reasonably widespread moral opposition to slavery would not have arisen in the way it did.

Just consider the possibility that Hitler would not have been morally depraved, but would rather have been "humane and fair-minded, free of nationalistic pride and racial hatred" (Sturgeon 1988, p. 250). In this case, he would not have done what he did. Or consider the possibility that he would have been, not morally depraved, but just a bit of a jerk. In this case, too, he plausibly would not have done what he did. Similar considerations apply to SLAVERY.

We may also give these counterfactuals an interventionist twist:

> HITLER*. Had Hitler's character been changed by an intervention from being morally depraved to not being morally depraved, he would not have initiated a world war and ordered the "final solution."

> SLAVERY*. Had the slavery in British and French America, and then in the United States, been changed by an intervention to not having been worse than previous forms of slavery, and slavery in Latin America, vigorous and reasonably widespread moral opposition to slavery would not have arisen in the way it did.

HITLER* and SLAVERY* also seem to be true. The mere addition of the relevant changes being brought about by an intervention makes no difference.

We make the simplifying assumption, here, that given the interventionist view, establishing that normative properties support counterfactuals such as HITLER, HITLER*, SLAVERY, and SLAVERY* suffices to establish the causal efficaciousness of the normative. We were willing to grant, above, that whatever further conditions must apply (e.g., regarding the alleged interventions really being genuine interventions), do apply in the case of the mental. We now assume that this plausibly is the case also in the case of the normative. That is, we see no reason to think that there would be any significant difference between the cases of mental and normative causation, with respect to the complications of the relevant sort. Given the assumption that the cases are analogous, in this respect, the truth of HITLER* and SLAVERY* establishes, by the lights of Raatikainen's account of causation, that Hitler's moral depravity and the level of badness of the slavery in British and French America, and then in the United States, are properties with causal power.

§5. According to Raatikainen's view, then, it is quite easy for normative properties to be causally efficacious. They just need to support the truth of counterfactuals such as HITLER, HITLER*, SLAVERY, and SLAVERY* (and pass whatever further conditions the interventionist theory involves, which we are now supposing they do pass).

Interestingly, it seems that their supporting such counterfactuals can be made sense of on a wide variety of different views in metaethics.

Consider, first, *non-naturalist* views. These views are *representationalist* in that according to these views, normative judgments represent, in some substantive and theoretically interesting sense, the ways the world is or, as we could also put it, the ways in which normative properties are instantiated. Moreover, according to non-naturalism, these properties are *sui generis*, irreducibly normative properties. On this view, normative properties such as the properties of being morally depraved or being very bad are something very different from the *natural* properties, which are properties such that can figure in empirical regularities and are amenable to study by empirical science. (For examples of non-naturalist views, see, e.g., Enoch 2011; Bengson, Cuneo & Shafer-Landau 2024.)

However, even though normative properties are, according to the non-naturalist, very different from the natural properties, the normative doesn't, even on the non-naturalist account, float free of the natural – to deploy a much-used phrase from Simon Blackburn (1984, p. 221). Rather, despite normative properties being so very different from natural properties, the normative ways the world is nevertheless *supervene* on the natural ways the world is, in something like the following way:

> Super. Metaphysically necessarily, for all *x* and all properties *F* in the set of normative properties *N*, if *x* has *F*, then there is some *G* in the set of natural properties *D* such that *x* has *G*, and metaphysically necessarily, for all *y*, if *y* has *G*, it has *F*.

It is metaphysically impossible, then, that there would be a normative difference between two possible things without there being some difference in their natural properties; it is metaphysically necessary that the natural ways the world is fix – with metaphysical necessity – the normative ways the world is.[4]

A similar assumption – implicit, in the discussion, above – about the supervenience of the mental on the physical, is crucial to the interventionist account of mental causation. In motivating the problem of mental causation, it was assumed that mental properties are not simply identical with physical ones. But, in order for it to be possible to manipulate the physical via interventions on the mental, it must nevertheless be assumed that the mental supervenes on the physical.

Now, it is an excellent question how non-naturalists can explain the supervenience of the normative on the natural (for an overview of the issues that this raises for the non-naturalist, see Väyrynen 2018). But what is relevant here is that non-naturalists, in any case, *accept* that the normative supervenes on the natural. We then have a guarantee, given this kind of non-naturalist view, that had Hitler not been morally

[4] Some non-naturalists reject Super because they deny that it is metaphysically necessary that the natural ways the world is fix, with *metaphysical necessity*, the normative ways the world is. These non-naturalists believe that the natural ways the world is only determine the normative ways the world is with a weaker *normative necessity*. For this kind of view, see, e.g., Rosen 2020. We set these views aside here.

depraved, his natural properties, too, would have been different. Given a suitable story about the kinds of natural properties that moral depravity supervenes on, we then have a guarantee that had Hitler not been morally depraved, he would have been free of the kind of combination of nationalistic pride and racial hatred that he sadly manifested. This allows normative properties to support, even given a non-naturalist understanding of such properties, the sorts of counterfactuals that would make it true, on the interventionist account of causation, that Hitler's moral depravity caused his initiating a world war and ordering the "final solution."

It seems, then, that according to the kind of account of causation favored by Raatikainen, the irreducibly normative *sui generis* properties postulated by the non-naturalist view have causal power. This seems like an interesting consequence for a theory of causation to have. For standardly, perhaps, non-natural properties are understood to be properties such that do *not* have causal power (see, e.g., Enoch 2011). Indeed, not having causal power would be one candidate for a feature that makes these properties non-natural. (For an account that grants that non-natural normative properties have causal powers, see Oddie 2005; for a recent attempt at navigating the complexities involved in giving a good account of what it is for a property to be non-natural, see Leary 2021).

Another type of view that we wish to highlight here is *quasi-realism*. Quasi-realism is a *non-representationalist*, *expressivist*, view, according to which normative language does not represent, in any substantive or theoretically interesting sense, normative properties and facts. Instead, on the expressivist view, normative language has a "dynamic," broadly practical, meaning. Its function is to guide attitude-formation and action. Very roughly, according to expressivist views, judging that Hitler was morally depraved is not a matter of representing Hitler as having had a certain kind of specifically normative property. Or at least this is not a very illuminating way of understanding the nature of this kind of judgment. Instead, judging that Hitler was morally depraved is more helpfully, and more fundamentally, understood in terms of being committed to acting or feeling in certain ways – in terms of being committed to disapproving of Hitler's character, say. (For this kind of way of understanding expressivism, see Dreier 2004, 2015; for expressivist views, see, e.g., Blackburn 1998, Gibbard 2003, Ridge 2014.)

Quasi-realism is what we get when we combine an expressivist view about the meaning of normative language and the nature of normative thought with the idea that there are normative properties, truths, and facts. According to the quasi-realist, some actions really have the property of being right, while others have the property of being wrong; it really is true – a fact – that Hitler was morally depraved; and so on. (Moreover, according to quasi-realism, such properties, truths, and facts are objective, in a certain interesting sense, but we can set this issue to one side here).

Now, it is an interesting question how expressivists can earn the right – to use another familiar phrase of Blackburn's – to accept the existence of normative properties, truths, and facts. A part of the reply is often taken to be that properties, truths, and facts don't carry a steep metaphysical price. For Hitler to have the property of being morally depraved, and for it to be a truth, or a fact, that Hitler was morally depraved, just is for it to be the case that Hitler was morally depraved. That Hitler was morally depraved is a normative claim that seems entirely compatible with quasi-realism. And so, it also seems entirely compatible with quasi-realism that there are normative properties (e.g., moral depravity, which Hitler instantiated), truths, and facts (e.g., the truth and the fact that Hitler was morally depraved). In any case, what is relevant here is just that quasi-realists accept the existence of normative properties, truths, and facts. (Almost all expressivists are quasi-realists, these days; see, again, e.g., Blackburn 1998, Gibbard 2003, Ridge 2014.)

Quasi-realists also accept that the normative properties of things supervene on their natural properties. Again, there are interesting questions that may be asked about whether they can really make good sense of the truth of something like SUPER. (They can.) But what is relevant here is that quasi-realists, in any case, accept the supervenience claim.

Given that quasi-realists accept the existence of normative properties, and that these properties supervene on the natural, quasi-realists, too, are in a position to accept claims such as HITLER* and SLAVERY*. Consequently, it seems that on the assumption of the Raatikainen-style account of causation, quasi-realists, too, get to have normative properties with causal power. As was the case with non-naturalism, here, too, the standard view presumably is that normative properties, as understood by the quasi-realist, are *not* the kind of thing that would have causal power. Quasi-realists have tried to explain how they, too, can make sense of *causal explanations* with normative explanantia (Blackburn 1991, Gibbard 2003, Sinclair 2012). But the idea has been to just explain the causal *relevance* of normative properties, or the relevance of normative properties to causal explanation. One possibility, here, is to appeal to the idea that when something has a normative property, this non-causally secures its also having some natural properties, which, in turn, can then do the *causing* (for this kind of account of macro-level causal explanation, more generally, see Jackson & Pettit 1990). But it is often thought that for a quasi-realist, normative properties are merely ‘shadows of predicates’ – that for a quasi-realist, being morally depraved just amounts to being a thing such that the predicate ‘morally depraved’ may truthfully be applied to it. And it would be natural to think that things are not imbued with causal power in virtue of having such shadowy properties, even if their having these shadow properties may be helpfully appealed to in giving causal explanations.[5]

[5] It is not clear that quasi-realists should treat normative properties as mere shadows of predicates. Some quasi-realists accept a reductive view, according to which normative properties are natural properties. It’s just that claims about which natural properties normative properties reduce to are given an expressivist interpretation (for this kind of view, see, e.g., Bex-Priestley 2024). Given a reductive quasi-realist account, the idea of a normative property as being causally efficacious does not sound so puzzling. But even on this kind of expressivist view, some claims about what causes what, namely the claims about normative causation, turn out to have a non-representational meaning, which some might find surprising. In any case, the Raatikainen-style account of causation seems to entail that normative properties need not be causally inert even if they are mere shadows of normative predicates that play a non-representational role.

§6. CONDITIONAL says, again, the following:

> CONDITIONAL. If the interventionist account of mental causation of the kind that Raatikainen proposes is correct, then normative properties have causal power, even given a non-naturalist or a quasi-realist understanding of such properties.

We have suggested that CONDITIONAL is plausibly true. We have also suggested, in the previous section, that normative properties do *not* have causal power, given a non-naturalist or a certain kind of quasi-realist understanding of such properties. If this is true, then this entails, together with CONDITIONAL, that the kind of interventionist account of causation defended by Raatikainen is not correct.

Is there a way for a defender of the interventionist account to resist this line of argument? One way to do so would be to turn our *modus tollens* into a *modus ponens* and suggest that given the plausibility of the interventionist account of causation, non-natural normative properties or normative properties as shadows of predicates should be construed as causally efficacious. This strikes us as an unpromising strategy. Another possible line of response would be to reject our assumption that there are no significant asymmetries between the cases of the mental and the normative, when it comes to the applicability of the interventionist account (e.g., the conditions that something must satisfy in order to be a genuine intervention). While we are not optimistic about the prospects of finding asymmetries of a relevant kind, a defense of the assumption that none exist remains outside the scope of this paper.

References

Bengson, John, Cuneo, Terence & Shafer-Landau, Russ (2024): *The Moral Universe*. Oxford: Oxford University Press.

Bennett, Karen (2007): 'Mental Causation', *Philosophy Compass* 2: 316–337. URL = https://doi.org/10.1111/j.1747-9991.2007.00063.x.

Bex-Priestley, Graham (2024): 'Expressivists Should be Reductive Naturalists,' in R. Shafer-Landau (ed.), *Oxford Studies in Metaethics, Vol. 19*: Oxford: Oxford University Press.

Blackburn, Simon (1984): *Spreading the Word*, Oxford: Oxford University Press.

Blackburn, Simon (1991): 'Just Causes', *Philosophical Studies* 61(½): 3–17.URL = https://www.jstor.org/stable/4320165.

Blackburn, Simon (1998): *Ruling Passions*, Oxford: Oxford University Press.

Crane, Tim (2001): *Elements of Mind*, Oxford: Oxford University Press.

Dreier, James (2004): 'Metaethics and the Problem of Creeping Minimalism', *Philosophical Perspectives* 18: 23–44. URL = https://www.jstor.org/stable/3840926.

Dreier, James (2015): 'Explaining the Quasi-Real', in R. Shafer-Landau (ed.), *Oxford Studies in Metaethics, Vol. 10*: Oxford: Oxford University Press.

Enoch, David (2011): *Taking Morality Seriously*, Oxford: Oxford University Press.

Gibbard, Allan (2003): *Thinking How to Live*, Harvard, Mass.: Harvard University Press.

Jackson, Frank & Pettit, Philip (1988): 'Program Explanation: A General Perspective', *Analysis* 50(2): 107–117. URL = https://doi.org/10.1093/analys/50.2.107.

Leary, Stephanie (2021): 'What is Non-Naturalism?', *Ergo* 8: 787–814. URL = https://doi.org/10.3998/ergo.2253.

Oddie, Graham (2005): *Value, Reality, and Desire*, New York: Oxford University Press.

Raatikainen, Panu (2006): 'Mental Causation, Interventions, and Contrasts,' unpublished manuscript. URL = https://philpapers.org/rec/RAAMCI.

Raatikainen, Panu (2007): 'Mentaalinen kausaatio,' in H. Gylling, I. Niiniluoto & R. Vilkko (eds.): *Syy*, Helsinki: Gaudeamus, 227–237.

Raatikainen, Panu (2010): 'Causation, Exclusion, and the Special Sciences,' *Erkenntnis* 73: 349–363. URL = https://doi.org/10.1007/s10670-010-9236-0.

Raatikainen, Panu (2013): 'Can the Mental Be Causally Efficacious?,' in K. Talmont-Kaminski & M. Milkowski (eds.): *Regarding the Mind, Naturally: Naturalist Approaches to the Sciences of the Mental*. Newcastle: Cambridge Scholars Press.

Raatikainen, Panu (2018): 'Kim on Causation and Mental Causation', *E-LOGOS - Electronic Journal for Philosophy* 25: 22–47. URL = https://doi.org/10.18267/j.e-logos.458.

Ridge, Michael (2014): *Impassioned Belief*, Oxford: Oxford University Press.

Rosen, Gideon (2021): "The Modal Status of Moral Principles," in R. Shafer-Landau (ed.), *Oxford Studies in Metaethics, Vol. 16*: Oxford: Oxford University Press.

Sinclair, Neil (2012): 'Expressivist Explanations', *Journal of Moral Philosophy* 9: 147–177. URL = https://doi.org/10.1163/174552412X625745.

Sturgeon, Nicholas (1984): 'Moral Explanations', reprinted in G. Sayre-McCord (ed.), *Essays on Moral Realism*, Ithaca: Cornell University Press, 1988.

Väyrynen, Pekka (2018): 'The Supervenience Challenge to Non-Naturalism', in T. McPherson & D. Plunkett (eds.), *The Routledge Handbook of Metaethics*, New York: Routledge.

Woodward, James (2003): *Making Things Happen*, New York: Oxford University Press.

21
Mental causation, folk psychology, and rational action explanation

Tomi Kokkonen

When I give a reason for someone's action, do I identify a cause for the person's external behaviour? This is one of the issues within the multitude of philosophical problems that are bound together as "the problem of mental causation". This particular issue stems from the fact that in our folk-psychological practices, we seem to both rationalize the action and to give it a causal explanation at the same time – but, given the causal exclusion of physical reality and the non-reducibility of rationalizations to physics, this cannot be the case. Therefore, something has got to give. The now-standard solution to the problem, pioneered by Panu Raatikainen (2007 & 2010; see also Menzies 2007 & 2008; Shapiro & Sober 2007; Woodward 2008), uses the interventionist idea of causation and the contrastive theory of causal explanation (Woodward 2003) to argue that the reasons for action indeed identify the cause, while identifying the underlying physical processes do not answer to the relevant causal explanatory question. While I sympathize with this solution, I will argue that since our folk-psychological explanations are inherently ambiguous, there is no solution to *the* problem of mental causation, but a clarification of the issue leads into a more multi-layered explication of mental causation events. The standard solution gives an account to one but only one of the issues.

Kokkonen, T. (2025). Mental causation, folk psychology, and rational action explanation. In A. Laitinen, M. Keinänen, J. Reinikainen & A. Honkasalo (eds.). *Language, Truth, and Reality.* Tampere University Press. (pp. 261–279). https://doi.org/10.61201/h5fbdk64

The ambiguity of folk psychology

Folk-psychological descriptions such as "I opened the can because I wanted to drink what is inside the can" are ambiguous on whether they refer to *propositional attitudes ascribable to a person as a whole* or *representational states participating in cognitive processes within the person's cognition*. They seem to do both at the same time, which is one source for the problems we have with mental causation. I suggest, however, that these are two different functions that the folk-psychological descriptions have, but our everyday practices need not to distinguish between the two functions, so they do not do so. As a preliminary, let us take a look at what these practices are supposed to be about.

There are several theories of propositional attitudes and the nature of folk psychology and several possible ways to classify them. For current purposes, I will group philosophers of mind into *cognitivists*, who think that propositional attitudes are causally effective psychological states (for example, Fodor 1981; Kim 2005; Shea 2018), and *ascriptionists*, who think that propositional attitudes are the states that we attribute to agents. The ascriptionists include interpretationists (for example, Davidson 1963; Anscombe 1967; Dennett 1987) and dispositionalists (for example, Ryle 1949; Marcus 1990). The theories put forward by these philosophers are attempts to give the semantics for folk-psychology and account for how it relates to the causal structure of the world. However, if we approach folk psychology as a natural human practice (see Bogdan 1997; Gopnik & Meltzoff 1997; Wellman & Liu 2004; Hutto 2007 & 2008; Call & Tomasello 2008; Duval et al 2011; Henry *et al* 2013; Zawidzki 2013; Andrews 2012; 2015a & 2015b; Apperly 2020), it may be that no philosophical theory is a correct explication of what folk psychology.

Consider the following example. A person, call him Aaron, is drowning. Another person, call her Bea, sees him in trouble. Bea jumps to the water and saves him. She does this because she wants to, and she has no further external aims or reasons for doing so, such as glory or gratitude. This seems to be an altruistic act. However, when we take a closer look at Bea's motivation, we might get more confused. There is a debate in psychology whether the perceived distress in others (Aaron in this case) can be a directly motivational factor (*the empathy-altruism hypothesis*) or does the motivation always go through some self-regarding process, such as the distress that Bea undoubtedly feels in the situation, caused by the perception of Aaron in distress (see Batson 2011). The latter view holds that all helping is always motivationally selfish. This is not a conceptual issue: if the latter is true, motivation to help can be blocked by blocking the agent's distress, but if the empathy-altruism hypothesis is correct, Bea would help Aaron anyway. But does it really make sense Bea's action would be selfish even if the egoistic theory of human motivation is correct? Let us assume that Bea is Aaron's mother. Seeing him drowning creates distress, sure, but she might try to help him even to the point of self-sacrifice. How would this be selfish? We can make a distinction between the opposition between egoism and altruism on the "deeper" psychological level of description (that has to do with the psychological motivation mechanisms) and the opposition between egoism and altruism on the agentive level of description (that has to do with the aims of the action; see Kokkonen 2021, chapter 6 for a throughout discussion). The need to make this distinction is clear in this example where the folk-psychological understanding of egoism and altruism becomes inherently contradictory, but the distinction itself is between two general levels of description.

Folk-psychological concepts may refer, then, to both holistic states of an agent (call them *agentive level* states) and entities within the cognition (*psychological level* proper), even at the same time, since the practice itself does not recognize the difference between the two. (The debate on the correct level of description for the reference of folk-psychological concepts is also a debate on how to understand intentionality, so I prefer to use the term "agentive" instead of "intentional.") Furthermore, there is no *need* to make the distinction between the two levels in folk-psychological practice. Here I will follow the Pluralistic Folk Psychology idea of Kristin Andrews (2012, 2015a & 2015b), the view that folk psychology is inherently pluralistic as a theory and as a practice, whether in the psychological processes involved in the practices, what its function in social life is, or what the references of its core concepts are. I have discussed both philosophical and empirical reasons for thinking this is the case elsewhere (Kokkonen 2021, chapter 5), but I will discuss some philosophical issues as a prelude to my argument for the nature of mental causation now. My aim is to discuss how the two levels of description interact in folk psychology, not to present an alternative interpretation for the correct level of description of folk psychology. Later, I will discuss how this issue fundamentally changes the issue of mental causation.

The different intuitions about the nature of folk psychology in philosophy – and these are not only theoretical disagreements about folk psychology but also normative disagreements about what the philosophical theories using its conceptual framework should be about – may be symptomatic of its pluralistic nature. On one hand, folk psychology seems to have theory-like characteristics: we explain, predict and manipulate others' future behaviour by manipulating their mental states. Mental states, whatever they are, seem to work as if they are causal factors, according to the Woodwardian understanding of causality. At the same time, propositional attitudes are intentional and rational, and this seems to be foundational for the semantics of folk psychology. These are different aspects of human agency and psychological phenomena related to it that are unified in folk psychology for pragmatic reasons. Daniel Dennett's distinction between intentional, design, and physical stances (Dennett 1987) is one way to make sense of this and to make ascriptionism compatible with a causal interpretation of psychology proper. In this view, the ascribed states are abstracted properties of the system, rather than parts of the system, and thinking of them as parts with causal role would be a category mistake. According to Dennett's (1991) metaphor, beliefs and desires are more like the centres of gravity than the concrete states of a mechanism.

All this seems somewhat vague, however, and there have been more recent attempts to analyse how intentionality arises from brain processes in a more detailed way from the representationalist perspective. One idea is that intentional states can be understood as robust outcome functions that have been stabilized by evolution and learning (for example, Godfrey-Smith 2006; Sterelny 2015; Shea 2018). These processes are controlled by sub-personal sub-systems, and their functional operations have representational content. Person-level attributions of beliefs and desires, however, are robust states of the individual (or the whole "system") that describe their cognitive relations with the world, descriptive and directive, and these relations are constituted by the parts of the representational system. This is plausible and I will not challenge the idea as such. However, it would still be a category mistake to reduce the states of the system to the parts of the system. Beliefs and desires are dependent on the system of representations, not parts of it. If the robust outcome function approach works, it explains the constitution of systemic states, but this does not build a conceptual link between the levels of description. What I suggest, instead, is that both agentive and psychological level are sensible levels of analysis, even if we also understand intentionality and representations at the psychological level. The latter is a separate question asking what explains, on the cognitive/psychological level, agentive-level intentionality – if anything does.

There are also interesting alternatives to the representational theory of mind within the naturalistic context that are still compatible with describing humans as agents. The most extreme is radical enactivism (Hutto & Myin 2013 & 2017), which takes the biological processes outside the central processing system more seriously as part of cognition. It proposes that much of cognition lacks any representational content at all and has more to do with how the sensory-motor system functions as a whole. The so-called "4e movement" (enactive, extended, embodied and embedded; see Newen, De Bruin & Gallagher 2018) approaches to mind and cognition in general challenge the classical representational theory of mind. But even if this approach provided the correct account, this would not make intentional action descriptions inadequate on the agent level. The debate between representationalists and their critics is not about attributing mental states to agents but about how the cognition works. The latter includes the issue of what explains the applicability of folk-psychological attributions to human agents, but this is a different issue, which highlights the need for the distinction.

Furthermore, and even more importantly, the precise relation between psychological and agentive levels is more difficult to understand with directive mental states than with descriptive ones. Psychological-level descriptive representations and agentive-level beliefs can be thought of as being in a complex constitutive relation. But how the drives and motivational salience relate to agentive-level pro-attitudes is trickier. Motivational salience is a crucial explanatory component in the emergence of pro-attitudes, but it is difficult to see how it alone could have the right kind of propositional content. It is simply a causal factor, incentivising or aversive, that instigates behaviour. Motivational salience explains preferences in part but is not itself a preference with content. Furthermore, pro-attitudes are about particular goals, not behavioural tendencies towards or away from a type of behaviour in a type of context, which motivational salience entails. The goals implied by a pro-attitude may be quite general and abstract, of course, such as world peace, being famous, or whatever goals moral values entail even prior to knowing these entailments. Folk psychology also accommodates moods and personality traits as more general and robust dispositional states. However, these are not the same as a tendency to be motivated by a certain type of things in certain contexts. This is also evident in how folk psychology is inadequate in capturing mental episodes such as depression. Depression has effects on individuals that make it difficult to rationalize their behaviour. The origin histories of depression cannot be fully understood in terms of folk psychology, either – that is, we cannot always give a rationalizing reason for being depressed, and it may be dangerously misleading when we try. Depression simply is not a reason-like state nor a collection of reasons or desires, and neither explaining depression nor explaining with depression is a rationalizing explanation (see Goldie 2007).

To sum up this part, there seem to be several philosophical and scientific reasons to distinguish the different frameworks conceptually, whatever their relation turns out to be. Without this revision, it looks like rationalizations of behaviour causally explain it, which seems to be both true (we *do* explain people's behaviour using reasons as if they were causes) and false (agentive descriptions do not refer to causal processes). If we make the distinction, the nature of the problem changes. The (rationalizing) agentive and (causally explanatory) psychological levels of description are also connected in various ways (for example, there is causally efficient rational deliberation that uses folk-psychological categories in reflection, and individual rationalizations have causal presuppositions) but folk psychology as a practice does not make the distinction or provide the tools to discuss how exactly the levels are connected. Consequently, folk-psychological explanations cannot be used directly in more sophisticated action explanation – philosophical, psychological, or evolutionary. I will take a closer look at this proposition now.

Rationality and rationalization

An essential source for philosophical difficulties (as well as the connectedness between psychological and agentive levels) is rationality. The agentive description attributes reasons to agents. The relationship between reasons to each other and the action is rational. However, the rationality of action seems to presuppose some sort of rationality in the causal processes that produce behaviour if agentive descriptions are given a causal explanatory role. Rationality itself cannot be a causal factor, but the causal processes must have systematicity in their functioning that exhibits behaviour that we perceive as rational. Furthermore, rational deliberation about the goals and means to achieve them is a part of human psychology, not just a property of action attributions. Attributing rationality to human psychology seems to be unavoidable.

There are, however, different concepts of rationality that may be applied to action and should not be conflated, especially if we are interested in their connection to causal explanation of behaviour. I will call the notion of rationality used in the philosophical theory of action *agentive rationality*. It is the idea that there is a reason for action. The action is rational when it is in accordance with the goals and beliefs of the agent in a way that can be expressed as giving the action a reason. There are both descriptive and normative elements in this: the action can be described as intentional by giving it a rationalizing conceptualization, but rationality is also evaluative in the sense that we consider action itself appropriate or not, given the reasons it was taken (see McGeer 2007; O'Brien 2019). Rationality may come in degrees in the sense that the action may be more or less appropriate, but *it is a qualitative property of an action that it can be given a reason*. It is about the intelligibility of behaviour as the action of an agent. The proposition that humans are rational agents is a categorical proposition about rationalization, both in its descriptive and normative dimensions. If humans are rational in this sense, rationalization is an adequate way to conceptualize humans and human behaviour. The normativity of rationality in this sense is what makes human action rational or *irrational*, while some other animals, for example, are not rational or irrational but *arational* (see Hurley & Nudds 2006). In contrast, the notion of rationality used in cognitive science, call it *cognitive rationality*, is a quantitative measure of cognitive capacity – but it is, likewise, also a normative notion. Rationality is measured against the *chosen optimality model,* which specifies what counts as rational, either in the *epistemic* (belief-formation) or *instrumental* (decisions about which course of action to take given the context) sense, and the *degree* of rationality and irrationality in human action and thinking is evaluated by comparing the performance to the model (see Stanovich 2011 & 2012).

The two senses of rationality, and the notions of normativity accompanying them, are different. The philosophical analysis of folk psychology uses the agentive notion. It is supposed to capture something that is *constitutive* of agency. Its normativity is about the adequacy of action given its reasons, and the failure to be rational is a failure to be an agent and for the behaviour to be intelligible as human action (see O'Brien 2019). Cognitive rationality and its normativity are instrumental: there are models that we *choose* to represent optimal decision in a context, *given the aims of the agent*, and we compare the behaviour to this. Moreover, these models (and the concept of rationality) could be applied to non-intentional systems, too, such as those animals that we consider not to be intentional, and to Artificial Intelligence systems. Agentive rationality does not imply any specific model of cognitive rationality. Moreover, it cannot be assumed that agentive rationality implies any specific *degree* of cognitive rationality that would enable *agentive rationality itself* to be an explanatory factor for behaviour, for example (see also Henderson 1993; Ylikoski & Kuorikoski 2016).

The two notions of rationality are also related. The models of cognitive rationality are *meant* to be about what a rational agent would ultimately choose, given their goals. This implies a third notion of rationality, *normative rationality*: how one *should* reason and choose action, given the goals. This is a stronger notion of rationality than the one used in rationalization of action – the assumption (rational) agency is not an assumption complete rationality. However, although this is a more demanding normative notion of rationality than the other two concepts in their normative component, the normativity of normative rationality is *instrumental:* it depends on chosen goals and acknowledged constraints on achieving these goals. This is the notion of rationality for fields such as Decision Theory and does not concern us here. However, the ability to be a rational agent in the agentive sense requires some cognitive capacities that explain it. Cognitive rationality is a measurement of how well some of the cognitive capacities function in certain tasks, and having these capacities is a partial explanation for why humans are agentively rational. These capacities are what we should be interested in when causally explaining human behaviour and how it fits with the causal understanding of human behaviour that humans are also (agentively) rational. I will refer to the agentive notion as "rationality" from now on, unless otherwise specified.

An essential feature of folk-psychological explanations is that they rationalize the behaviour into actions that have reasons behind them and goals to look forward to. Reasons (and their constituents, beliefs and desires, the "two directions of fit," as Elizabeth Anscombe (1967) put it, descriptive and directive) are connected to each other and to the action in rational relations: the propositional contents entail other propositional contents and are attributed to agents as holistic sets. At the same time, the attributions identify what action the behaviour is, and the identification of the behaviour as doing *x* is a part of the interpretation of which beliefs and pro-attitudes of the agent constitute their reason for action in the situation. That is, the action descriptions are a part of the same holistic net of semantic connections as the mental states that make behaviour intelligible. Some philosophers take this to mean that rationalizations cannot be causal explanations, since semantic entailments are not causal relations (Anscombe 1967; von Wright 1971; Sehon 1997 & 2005). Others think that this merely makes the ontology of action somewhat anomalous (Davidson 1970). It seems that folk-psychological practices require the attributions to have at least some causal counterfactual power: the point of persuasion and reasoning with a person, for example, is to change their underlying structure of desires and beliefs to affect their future behaviour. This is a causal intervention, not a matter of interpretation after the fact. Mental attributions should not be causal attributions, under some conceptual and metaphysical considerations, but they seem to function as if they were. Hence the attempts to reduce the rationalizing elements into something that also has psychological reality. (See Henderson 1993; Crane 1995; Mele 2000; Kokkonen 2011.)

Furthermore, folk-psychological practices seem to presuppose that of all the reasons we can attribute to the agent, give the agent's mental states, there is a *primary* reason among them that is why the agent actually did what they did. It determines what the action was about – it is not just an alternative description for the behaviour. How should we understand this? For a causalist like Davidson, the primary reason is the one that caused the action. Under the ascription view, this is a problem known as *Davidson's Challenge*: a mere ascription is only about pattern fitting, it does not explain action (see Davidson 1963; Mele 2000; O'Brien 2019). For the causalists the problem is how the reasons can be causes. This problem can be broken into two parts. First, how can mental states in general (that is, agentive states under the description of folk-psychological conceptualization) be causes of physical behaviour? I call this the *Core Problem of Mental Causation*. Second, how could rationalization of action reliably capture states that are causally efficient for behaviour? In other words, how can *reasons*, identified by their modal and logical properties, be causal? I call this the *Hard Problem of Action Explanation*.

The hard problem of action explanation

The recently popular solution to the core problem of mental causation has been to use the contrastive theory of causal explanation and the manipulationist theory of causation as a framework to identify causal factors (Raatikainen 2007 & 2010; Menzies 2007 & 2008; Shapiro & Sober 2007; Woodward 2008). Folk-psychological descriptions make robust but imprecise claims about causal processes and behavioural dispositions of the agents on which the behaviour depends, with relevant counterfactual contrasts. These robust states are the states of the agents. There may be psychological states that implement these more or less directly and have causal relations with other psychological states and the behaviour, and similarly with neural states – but these are further issues. What matters is that the mental descriptions identify states that have intelligible contrast classes, and the difference between the explanatory state and its contrast class is a difference-maker between the explained behaviour and its contrast class. The *explananda* and *explanantia* need not be described on the same level, for as long as the framework identifies the correct dependence relation. In other words, reasons can be causes when having a reason is the adequate identification of a causal disposition.

Furthermore, the contrast classes of explanation may be different when referring to the causal process on different levels. In fact, given that we attempt to explain behaviour that is specified with a goal, a folk-psychological description (including reasons and intentions) may be a *more* adequate way of identifying the contrast class than an alternative explanation on a different level (see Raatikainen 2010). This seems to solve the causal explanatory part of the problem regardless of what the relationship between agentive states and the underlying causal processes may be. Moreover, it grants autonomy to causal explanations on different levels and fits the general pluralistic approach adopted here. Some issues remain untouched with this solution, however. These include problems such as the ontological relation between the objects of the different descriptions. More importantly, this solution does not touch the issue about the role of rationality and rationalization itself (the Hard Problem): how can we discover causes of behaviour by rationalizing action (or rather: can we do so, and how can we justify this practice)?

The problem has two components. First, how is it possible that humans are natural beings whose behaviour is a part of the causal structure of the world but follow the dictates of rationality at the same time? (*The role of rationality in naturalism.*) Second, is rationalization a form of (causal) explanation? There are only two possible solutions to the first part: some sort of anomalous monism (Davidson 1970), or that humans are not actually as rational as rationalization practices presuppose. There are good empirical reasons to think that humans are not fully rational when it comes to *cognitive* rationality (see Kahneman, Slovic & Tversky 1982; Kahneman 2011; Gigerenzer 2007; Stanovich 2011 & 2012). As discussed above, the agentive notion of rationality is a different notion, but there is a substantial connection between the two notions in *explaining* rationality. If rationality itself does not have causal powers (and it follows from the naturalistic premises that it does not) there must be something in the psychology that explains this. Rational deliberation is a part of how mind works, and it has causal consequences, but the empirical research seems to imply that it plays a limited role in cognition. This also implies limitations on the extent of agentive rationality in humans. (See also Henderson 1993; Ylikoski & Kuorikoski 2016.) Partial rationality (whether it is because of deliberation or something else) may be enough to justify the interpretative practices, however, and it is not an unsolvable problem for a naturalistic view of humans. Humans have complex cognitive systems adapted to survive flexibly in complex, changing environments. A part of this process has been the decoupling of representations from what is immediate, and this has also created a need to represent states of affairs as related to each other and make inferences between them (Godfrey-Smith 1996; Sterelny 1999 & 2003). In other words, humans have evolved psychological processes that are causal (and implemented by neural processes) but deal with representational states in a way that is partially rational, since the psychological mechanisms have been selected for having rational outcomes. But this rationality is relative to selected tasks and their proper contexts, and even there it is limited by how reliably rational the outcomes the underlying biological structure can produce. There is no selection for universal rationality. Even if there was, an organ such as the brain could not produce universal rationality through causal operations. Then again, we are not universally rational. This solution also makes the rationality of human behaviour and psychology (to the extent that it is rational) an *explanandum* itself – *rationality* (the entailments between propositional attitudes) does not explain rationality of behaviour. *Rationality is a part of descriptions of the behaviour to be explained.*

There are explanations for partial rationality. Folk psychology is a crucial part of our social practices and the cognitive skills related to them, and it evolved to be functional for the many different needs of our many kinds of social interaction (Byrne & Whiten 1988; Bogdan 1997; Corbalis & Lea 1999; Tomasello 2009; Emery 2012; Devaine *et al* 2014). The need for effective mindreading for various social activities to be possible, entails selection pressures on our behavioural tendencies too, as well as the "control structures" in our cognition, to be more in accordance with the kind of rationality that we use as a guide in folk psychology (Sterelny 2015). Furthermore, the folk-psychological practices, the language related to them (see Gopnik & Meltzoff 1997; Zawidzki 2013), and the agent-based narrative structure we learn in childhood (see Hutto 2008) affect our thinking. They do have not only mindreading but also *mindshaping* functions – they are an extra-genetic form of inheritance to shape our behaviour and its underlying psychology to be in line with folk-psychological assumptions, as suggested by Matteo Mameli (2001) (see also Zawidzki 2013; Sterelny 2015). Moreover, folk psychology has regulative and justificatory functions in social interaction (Andrews 2015a & 2015b; see also McGeer 2007; Zawidzki 2013). All this makes rationality understandable from a naturalistic point of view as far as it is limited, but rationality as such does not play an explanatory role in why we think and act rationally.

This still leaves us with the second problem: How could rationalizing with a reason itself be an explanation? One possible solution would be to revise the non-causalist stance on attribution of mental states by proposing that psychological states (in the narrow sense) are references to causal states, but rationalizations are about agentive states. I have already alluded to something like this as the first approximation. But making this distinction alone would cut the connection between rationalizations and causal explanations, and rationalizing attributions seem to work as attributions of causal factors, as discussed earlier. We could go even further: to reason to rationalize action in the first place is only because it captures something causal that is useful to us. Moreover, it would leave us with Davidson's Challenge: the notion of primary reasons, the intended reasons for action, require some further explanation if intention is not causally effective (see Mele 2000; O'Brien 2019).

A causal presoppositionalist account of rational action explanation

Consider the following option. It is not the *reasons* the agent has that are manipulated in an interaction, but something that *having the reasons depends on* (that is, something causal that can be described on the psychological and/or neurophysiological level). If the connection between the reasons and their underlying conditions is sufficiently robust, reasons identify causal relations, albeit under an imprecise description. Reasons are attributed by rationalization, and they depend on psychological processes. This would be a form of anomalous monism that is not anomalous, given there are explanations available for why the two are correlated. However, this is not a sufficient solution. Describing intentional action involves ascribing an *intention* to the agent, not just rationalizing reasons for action that can be interpreted for the agent: some reasons express what the agent intends to do, and these intention references are clearly meant to capture something causal (Davidson 1978; Bratman 1987; Mele 1992 & 2009). And as Elizabeth Anscombe (1967) (albeit a non-causalist herself) pointed out already, we also seem to have direct knowledge of our own intentions. Our knowledge of all the factors that play roles in why we do what we do may be fallible, but the experience of intending to do something specific is direct, not a process of interpretation. Within the causal interpretation, we identify some of our reasons as causes. Furthermore, we do not just act and interpret the action; we reason about our goals and the means to achieve them, and this reasoning seems to make some causal contribution to producing behaviour. Hence, there seems to be a connection between agentive rationalizations and causal psychological processes.

It is, however, one thing to say that we have more intuitive understanding of ourselves as agents than a mere interpretation and another thing completely to say that this understanding involves direct observations of the causal processes that guide our behaviour. We are only conscious of a part of our cognitive processes and motivations for action. Cognitive and social psychologists distinguish two kinds of processes in mind (the so-called *dual process* and *dual system* theories of cognition): Type I (or System 1) and Type II (or System 2). Type I processes are *automatic*; they are fast, reactive, non-conscious, associative, heuristic, and effortless. Type II processes are *analytic*; they are slow and effortful but controlled and deliberative. (See Kahneman, Slovic & Tversky 1982; Evans and Over 1996; Bargh & Chartland 1999; Stanovich 1999 & 2011; Kahneman 2011; Gigerenzer 2007; Frankish & Evans 2009; Evans & Stanovich 2013.) These processes (or systems) are jointly activated, and they give a rise to more complex cognitive operations, but only some processes are conscious, and we are (indirectly) aware of only some of the non-conscious processes. We have no access to all the processes that influence our thinking, even our conscious thinking. When people are asked about the reasons for their actions, they do not identify an *effective motivation* behind them, but describe a *state with a goal*, and this may be just as much a rationalization after the fact as if they were explaining another person's action, even in highly deliberative contexts such as making a moral judgment (Haidt 2001; see also Nisbett & Wilson 1977; Nisbett & Ross 1980; Bargh & Chartrand 1999).

As mentioned earlier, the notion of rationality used in cognitive science is different from the one used in the analysis of folk-psychological conceptualizations, although there are substantial connections. There are two properties of the two-level cognitive system that are consequential for the issue at hand. First, the analytic processes that we are conscious of and constitute our deliberation are the ones we identify as our thinking and decision-making in our cognitive phenomenology. We experience other states too, such as emotions, and we are usually aware that we have other psychological motivating factors, but reasoning is what we consider to be our "actual" thinking and we have an impression that it is responsible for our decision-making. We can disregard the normative, gradual notions of cognitive rationality for a while and concentrate on some of the qualitative aspects of the analytic processes. First, they process propositional contents: this part of cognition is closest to what folk-psychological rationalization presumes human thinking to be like. Second, our thinking and decision-making involves the non-conscious processes as well, even while we deliberate, and they have inputs into the deliberation. When we deliberate, we become aware of the products of non-conscious processes as our own thoughts (even if we do not have access to the processes producing them), and they become a part of further deliberation. Third, the agentive rationalizing attributions to agents (as whole persons) do not distinguish between these two kinds of processes.

If folk psychology is pluralistic both in its mechanisms but also in its reference, this extends to self-reflection. When we reflect our motives and decisions, we attribute rationalizing intentional states (desires, beliefs, reasons) to ourselves according to folk psychology. However, the sources for these states include both the deliberative process and the other processes that participate in guiding our thinking and behaviour. If this is the case, the object of reflection on our own mental states is a combination of deliberative conscious states, products of non-conscious processes that we are aware of and that we interpret in folk-psychological categories, and quasi-theoretical assumptions about ourselves that are folk-psychological postulates. Our self-understanding is fallible regarding these differences. Even if our self-attributions of mental states are correct in terms of folk psychology in the moment of action, and even if they are based on epistemically reliable self-observations, our *justificatory* self-rationalization does not necessarily identify the *causal* processes of how we came to the decision correctly. Furthermore, we are not necessarily correct in our self-attributions either, and our self-observations are not always reliable.

However, reflection is not mere rationalization. Sometimes we explain our own behaviour with non-rational causes, such as anger, sorrow, or intoxication. But the point here is that sometimes we also misidentify having non-rationalizable psychological processes as having reasons. Conscious reasoning (as a part of cognition) and interpretation using the theory of mind (on the agentive level) are confused in the simplified image of rational agency, and they should be distinguished. Moreover, although we experience intending and identify it correctly as the motivational state that triggers action, the content (the reason, or a plan) we accompany it with may sometimes nevertheless be an interpretation within a folk-psychological conceptual framework, not an experience of a deliberated state. There are also problems with prediction of one's own behaviour: people are notoriously bad at predicting their future actions based on their current self-perceived states of minds – although it is not clear whether this is because of misinterpretation of one's own motives or underestimating the situational factors that are not present in the context of prediction (Poon, Koehler & Buehler 2014).

Having an intention (in the sense of intending) does, however, presuppose that there is at least one causal factor that is identified in experiencing intending. Psychologically speaking, we experience motivational forces, aversive and incentivising saliences that guide our behaviour. A successful agentive explanation does not need to specify these processes precisely to be a form of causal explanation. But a successful agentive description must include a reference to the *existence* of such factors. The identification of an intention in the context of action involves attributing a reason that adequately describes the agent's relation to the world in a robust way in the context, given both her epistemic states and active motivational forces. Moreover, even if we think the agent *knows* what they are doing (or what their intention is), this does not require them to know all the psychological processes involved. On the other hand, when I reflect my own motives, or an external observer of the situation wonders what it is that I am doing, the observation or speculation (depending on which one is doing it) targets the *psychological states*, but this may still use the same *goal-directive semantics*.

Rational deliberation is a *part* of our cognitive capacities. It is a part of the causal makeup of mind, not just a passive reflection of cognition. But reason, in this sense, is not a determinative factor. At the same time, the object of rationalization is the action, not a partial factor of it: we use folk psychology to represent our own holistic agentive states, such as beliefs and desires, in metacognition (whether in conscious deliberation or in automatized processing, which also has metacognitive functions). We do not represent just the reasoning part of our cognition, although this is the part we mostly identify our thinking with, and we tend to conflate the two – the contents of reasoning and the contents of the holistic states. Folk-psychological categorizations affect how we deliberately plan our actions, but once again, this is causal influence of folk psychology on our cognition – it does not make agentive and psychological states the same. The same applies the other way around; not all behaviour needs to be produced by deliberation alone to be rationalizable in the sense of agentive rationality. Much of the unconscious, automatized processing has a positive function in reaching the chosen goal of action (see Bargh & Chartrand 1999; Gigerenzer 2007; Marewski, Gaissmaier & Gigerenzer 2010).

Re-thinking mental causation (again)

We can summarize the outcome of the discussion in this paper so far in the following propositions about rationality and action: (1) Reasoning (as rational deliberation) is a cognitive process that participates in the causal production of behaviour. (2) People are conscious of this part of their own cognitive processes, while the other processes manifest only in the products of these processes. (3) The non-conscious parts of cognition are often instrumental to the chosen goals, and therefore they participate in producing the action that we rationalize *without being a part of the rational guidance* of the action on the cognitive level. (4) People rationalize both their own and others' actions within the folk-psychological framework and this rationalization has more to do with justification and evaluation than causal explanation, but it functions both ways. (5) People conflate the *rationalization* they apply to their action and the *experienced intention* that triggers this action *whether it is the outcome of rational deliberation or some other process*. It can be either. To the degree that non-rational processes are instrumental to chosen goals, this does not make a difference in understanding the action as guided by reasons. But giving only reasons in the causal explanation misidentifies the causes. (6) People are aware of non-rationalizable causes such as emotions and use them in the folk-psychological explanations as well, and these explanations are not rationalizations. Emotions, personality characteristics, reasons and other factors are not separated as being on different levels; but there is only one folk-psychological "level". (7) Sometimes, people have no idea what motivated them, and their rationalization of their own action is simply incorrect.

If these conclusions are accepted, agentive, rationalizing descriptions do not refer directly to psychological processes with causal powers, but they *presuppose* that there are causal processes that are responsible for the action in order for the folk-psychological practices to work (see O'Brien 2019). In these practices, agentive attributions of reasons and attributions of psychological states proper that underlie the agentive states are mixed into a heterogeneous category, and the distinction between them would not make a difference. The connection between agentive ascriptions and the underlying psychology is strong enough to allow rational arguments, persuasion and other folk psychology-based practices to enter the cognitive system and influence behaviour. In philosophical and scientific scrutiny, however, different levels of description need to be acknowledged. Slices of the causal process that result in behaviour can be described on any level, although they do not make the same causal explanatory claims (since they have different contrast classes) and sometimes there is no rationalizing action explanation at all – that is, when the behaviour under scrutiny is irrational. For psychological and philosophical purposes, however, the two levels should be kept apart.

As I mentioned above, the problem of mental causation breaks into two sub-problems that I referred to as the Core Problem and the Hard Problem. The classic Davidsonian problem of mental causation was about the relationship between the physical (causal) domain of regularities that humans as natural entities follow and the rational level of action explanations that refer to reasons. This relationship has three steps in total. The first step is how biological design emerges from physical regularities. This is well understood, but it is worth noting that the causal basis in mental causation is not physics and regularities in its processes, but rather neurobiology which is constituted by physical processes but also has evolved functional structure. The second issue is how the psychological level (cognitive and conative) descriptions are related to neurobiology. We may have some understanding about this on empirical grounds, as I alluded above. The third step is how rationalizable states emerge from the psychological mechanisms and processes – and, again, we have some idea how this works, once we keep the two levels distinct and do not conflate them into one level of "mental descriptions". As for the problem of mental causation, I will make the following conjectures. First, the solution to the Hard Problem (that is, how could rationalizations reliably capture causal states) is that they do not. Not reliably – but often enough to make folk-psychological practices useful most of the time, for the reasons discussed above. Second, the solution to the Core Problem (that is, how can mental states be causes) is that when the folk-psychological statements refer to the psychological level proper, we can understand them as functional descriptions of brain processes and mechanisms that get their semantics partially from folk psychology, but when they do not refer to them (given the answer to the Hard Problem), they do *not* refer to causes but make a presupposition of an existence of a causal structure with the right kind of effects. Rather than identifying the cause, as in the first case (in the accordance with the standard solution), they identify an effect that, nevertheless, gives us information about what the action is about. Given the pluralistic nature of folk psychology, there is not *the* solution to the problem of mental causation.

If the account put forward here is correct, the standard solution to the problem of mental causation, put forward by Raatikainen and others, is not wrong, but it is imprecise and does not always capture the correct causal structure. However, it gives the correct ontology on what kind of events mental causation events are. It also captures the logic of causal explanations of action in the cases in which reason-based explanations are causal. It is also worth noting that introducing the standard solution was a groundbreaking shift in the discussion on mental causations and the account given here is also building upon it, by adding detail and incorporating research and discussions on folk psychology in other fields.

References

Andrews, Kristin (2012): *Do Apes Read Minds? Toward a New Folk Psychology*, Cambridge, Ma: MIT Press.

Andrews, Kristin (2015a): 'Pluralistic folk psychology and varieties of self-knowledge: an exploration', *Philosophical Explorations* 18(2): 282–296. URL = https://doi.org/10.1080/13869795.2015.1032116.

Andrews, Kristin (2015b): 'Folk psychological spiral: explanation, regulation, and language', *The Southern Journal of Philosophy* 53(S1), Spindel Supplement: 50–67. URL = https://doi.org/10.1111/sjp.12121.

Anscombe, Elizabeth (1967): *Intention*, 2nd edition (2000 reprint), Cambridge, MA: Harvard University Press.

Apperly, Ian (2010) *Mindreaders: The Cognitive Basis of Theory of Mind*, New York, NY: Psychology Press.

Bargh, John A. & Thanya L. Charthrand (1999): 'The unbearable automaticity of being', *American Psychologist* 54(7): 462–479. URL = https://doi.org/10.1037/0003-066X.54.7.462.

Batson, C. Daniel (2011): *Altruism in Humans*, Oxford: Oxford University Press.

Bogdan, Radu J. (1997): *Interpreting Minds*, Cambrige, Ma: MIT Press.

Bratman, Michael (1987): *Intention, Plans, and Practical Reason*, Cambridge, MA: Harvard University Press.

Byrne, Richard W. & Andrew Whiten (eds.) (1988): *Machiavellian Intelligence: Social Expertise and the Evolution of Intellect in Monkeys, Apes and Humans*, Oxford: Oxford University Press.

Call, Josep, & Michael Tomasello (2008): 'Does the chimpanzee have a theory of mind? 30 years later', *Trends in Cognitive Science* 12(5): 187–192. URL = https://doi.org/10.1016/j.tics.2008.02.010.

Corballis, Michael C., & Stephen E. G. Lea (eds.) (1999): *The Descent of Mind: Psychological Perspectives on Hominid Evolution*, Oxford: Oxford University Press.

Crane, Tim (1995): 'The mental causation debate', *Proceedings of the Aristotelian Society* (Aristotelian Society Supplementary), 69: 211–236. URL = https://doi.org/10.1093/aristoteliansupp/69.1.211.

Davidson, Donald (1963): 'Actions, Reasons and Causes', *Journal of Philosophy* 60(23): 685–700. URL = https://doi.org/10.2307/2023177.

Davidson, Donald (1970): 'Mental Events', in Lawrence Foster and J. W. Swanson (eds.), *Experience and Theory*, London: Duckworth.

Davidson, Donald (1978): 'Intending', in Yirmiaku Yovel (ed.), *Philosophy of History and Action*, D. Reidel and the Magnes Press, 41–60.

Dennett, Daniel C. (1987): *The Intentional Stance*, Cambridge, Ma: MIT Press.

Dennett, Daniel C. (1991): 'Two contrasts: folk craft versus folk science, and belief versus opinion', in John Greenwood (ed.): *The Future of Folk Psychology: Intentionality and Cognitive Science*, Cambridge: Cambridge University Press, 135–148.

Devaine Marie, Guillaume Hollard & Jean Daunizeau (2014): 'Theory of Mind: Did Evolution Fool Us?"' PLOS One 9(2): e87619. URL = https://doi.org/10.1371/journal.pone.0087619.

Duval, Céline, Pascale Piolino, Alexandre Bejanin, Francis Eustache & Béatrice Desgranges (2011): 'Age effects on different components of theory of mind', *Consciousness and Cognition* 20(3): 627–642. URL = https://doi.org/10.1016/j.concog.2010.10.025.

Emery, Nathan (2012): 'The evolution of social cognition', in Alexander Easton & Nathan Emery (eds.), *The Cognitive Neuroscience of Social Behaviour*, Hove and New York: Psychology Press, 115–156.

Evans, Jonathan St. B.T. & David E. Over (1996): *Rationality and Reasoning*, Psychology Press.

Evans, Jonathan S.B.T., & Keith E. Stanovich (2013): 'Dual-process theories of higher cognition: advancing the debate', *Perspectives on Psychological Science* 8(3): 223–241. URL = https://doi.org/10.1177/1745691612460685.

Fodor, Jerry (1981): *RePresentations: Philosophical Essays on the Foundations of Cognitive Science*, Cambridge, Ma: MIT Press.

Frankish, Keith & Jonathan Evans (eds.) (2009): *In Two Minds*, Oxford: Oxford University Press.

Gigerenzer, Gerd (2007): *Gut Feelings: The Intelligence of the Unconscious*, New York: Viking Penguin.

Godfrey-Smith, Peter (1996): *Complexity and the Function of Mind in Nature*, Cambridge: Cambridge University Press.

Godfrey-Smith, Peter (2006): 'Mental Representation, Naturalism and Teleosemantics', in David Papineau & Graham Macdonald (eds.), *New Essays on Teleosemantics*, Oxford: Oxford University Press, 42–68.

Goldie, Peter (2007): 'There are reasons and reasons', in Daniel D. Hutto & Matthew Ratcliffe (eds.), *Folk Psychology Reassessed*, Dordrecht: Springer, 103–114.

Gopnik, Alison & Andrew Meltzoff (1997): *Words, Thoughts and Theories*, Cambridge, Ma: The MIT Press.

Haidt, Jonathan (2001): 'The Emotional Dog and Its Rational Tail: A Social Intuitionist Approach to Moral Judgment', *Psychological Review* 108(4): 814–834. URL = https://doi.org/10.1037/0033-295x.108.4.814.

Henderson, David (1993): *Interpretation and Explanation in the Human Sciences*, Albany: State University of New York Press.

Henry Julie D., Louise H. Phillips, Ted Ruffman & Phoebe E. Bailey (2013): 'A meta-analytic review of age differences in theory of mind', *Psychology and Aging* 28(3): 826–839. URL = https://doi.org/10.1037/a0030677.

Hurley, Susan & Matthew Nudds (2006): 'The questions of animal rationality: Theory and evidence', in Susan Hurley & Matthew Nudds (eds.), *Rational animals?,* Oxford: Oxford University Press, 1–83.

Hutto, Daniel D. (2007): 'Folk Psychology without Theory or Simulation', in Daniel Hutto & Matthew Ratcliffe (eds.), *Folk Psychology Reassessed*, Dordrecht: Springer, 115–135.

Hutto, Daniel D. (2008): *Folk Psychological Narratives: The Sociocultural Basis of Understanding Reasons*, Cambridge, Ma: MIT press.

Hutto, Daniel D. & Erik Myin (2013): *Radicalizing Enactivism: Basic Minds Without Content*, Cambridge: MIT Press.

Hutto, Daniel D. & Erik Myin (2017): *Evolving Enactivism – Basic Minds Meet Content*, Cambridge: MIT Press.

Kahneman, Daniel (2011): *Thinking, Fast and Slow*, London: Macmillan.

Kahneman, Daniel, Paul Slovic & Amos Tversky (eds.) (1982): *Judgment Under Uncertainty: Heuristics and Biases*, Cambridge: Cambridge University Press.

Kim, Jaegwon (2005): *Physicalism, or Something Near Enough*. Princeton: Princeton University Press.

Kokkonen, Tomi (2011): ’Mielen teoria, selittäminen ja ymmärtäminen’, *Tiede & edistys 36*(4): 277–290. URL = https://doi.org/10.51809/te.105048.

Kokkonen, Tomi (2021): *Evolving in Groups: Individualism and Holism in Evolutionary Explanation of Human Social Behaviour*, Doctoral Thesis, Helsinki: University of Helsinki. URL = http://hdl.handle.net/10138/333344.

Marcus, Ruth B. (1990): ‘Some revisionary proposals about belief and believing’, *Philosophy and Phenomenological Research* 50: 132–153. URL = https://doi.org/10.2307/2108036.

Marewski, Julian N., Wolfgang Gaissmaier & Gerd Gigerenzer (2010): ‘Good judgments do not require complex cognition’, *Current Directions in Psychological Science* 11(2): 103–121. URL = https://doi.org/10.1007/s10339-009-0337-0.

McGeer, Victoria (2007): ‘The Regulative Dimension of Folk Psychology’, in Daniel D. Hutto & Matthew Ratcliffe (eds.): *Folk Psychology Re-Assessed*, Dordrecht: Springer, 137–156.

Mele, Alfred (1992): *The Springs of Action*, New York: Oxford University Press.

Mele, Alfred (2009): *Effective Intentions: The power of conscious will*, Oxford: Oxford University Press.

Menzies, Peter (2007): ‘Mental Causation on the Program Model’, in G. Brennan, R. Goodin, F. Jackson, & M. Smith (eds.): *Common Minds: Themes from the Philosophy of Philip Pettit*, Oxford: Oxford University Press, 28–54.

Menzies, Peter (2008): ‘Exclusion problem, the determination relation, and contrastive causation’, in J. Hohwy & J. Kallestrup (eds.): *Being Reduced—New Essays on Reduction, Explanation and Causation*, Oxford: Oxford University Press, 196–217.

Newen, Albert, Leon De Bruin & Shaun Gallagher (2018): *The Oxford Handbook of 4E Cognition*, Oxford: Oxford University Press.

Nisbett, Richard & Lee Ross (1980): *Human Inference: Strategies and Shortcomings of Social Judgement*, Englewood Cliffs: Prentice Hall.

Nisbett, Richard & Timothy Wilson (1977): ‘Telling more than we can know: Verbal reports on mental processes’, *Psychological Review* 84(3): 231–259. URL = https://doi.org/10.1037/0033-295X.84.3.231.

O'Brien, Lilian (2019): ‘Action Explanation and its Presuppositions’, *Canadian Journal of Philosophy* 49(1): 123–146. URL = https://doi.org/10.1080/00455091.2018.1518629.

Poon, Connie S. K., Derek J. Koehler & Roger Buehler (2014): ‘On the psychology of self-prediction: Consideration of situational barriers to intended actions’, *Judgment and Decision Making* 9(3): 207–225. URL = https://doi.org/10.1017/S1930297500005763.

Raatikainen, Panu (2007): ’Reduktionismi, alaspäinen kausaatio ja emergenssi’, *Tiede & Edistys* 32(4): 284–296. URL = https://doi.org/10.51809/te.104902.

Raatikainen, Panu (2010): ‘Causation, exclusion, and the special sciences’, *Erkenntnis* 73(3): 349–363. URL = https://doi.org/10.1007/s10670-010-9236-0.

Ryle, Gilbert (1949): *The Concept of Mind*, Chicago: University of Chicago Press.

Sehon, Scott (1997): ‘Deviant Causal Chains and the Irreducibility of Teleological Explanation’, *Pacific Philosophical Quarterly* 78(2): 195–213. URL = https://doi.org/10.1111/1468-0114.00035.

Sehon, Scott (2005): *Teleological Realism: Mind, Agency, and Explanation*, Cambridge, Ma: MIT Press.

Shea, Nicholas (2018): *Representation in cognitive science*, Oxford: Oxford University Press.

Stanovich, Keith E. (1999): *Who is Rational? Studies of Individual Differences in Reasoning*, Lawrence Erlbaum.

Stanovich, Keith E. (2011): *Rationality and the Reflective Mind*, Oxford: Oxford University Press.

Stanovich, Keith E. (2012): ‘On the Distinction between Rationality and Intelligence: Implications for Understanding Individual Differences in Reasoning’, in Keith J. Holyoak & Robert G. Morrison (eds.), *The Oxford Handbook of Thinking and Reasoning*, Oxford: Oxford University Press, 343–365.

Sterelny, Kim (1999): ‘Situated Agency and the Descent of Desire’, in Valerie Gray Hardcastle (ed.), *Biology Meets Psychology: Constraints, Conjectures, Connections*, Cambridge: MITPress.

Sterelny, Kim (2003): *Thought in a Hostile World: The Evolution of Human Cognition*, Oxford: Blackwell Publishing.

Sterelny, Kim (2015): ‘Content, Control and Display: The Natural Origins of Content’, *Philosophia* 43(3): 549–564. URL = https://doi.org/10.1007/s11406-015-9628-0.

Tomasello, Michael (2009): *Why We Cooperate*, Cambridge, Ma.: MIT Press.

von Wright, Georg Henrik (1971): *Explanation and Understanding*, Cornell University Press.

Wellman, Henry M. & David Liu (2004): ‘Scaling of Theory-of-Mind Tasks’, *Child Developing* 75(2): 523–541. URL = https://www.jstor.org/stable/3696656.

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

Woodward, James (2008): ‘Mental Causation and Neural Mechanisms’, in Jakob Hohwy & Jesper Kallestrup (eds.), B*eing Reduced: New Essays on Reduction, Explanation, and Causation*, Oxford: Oxford University Press.

Ylikoski, Petri & Jaakko Kuorikoski (2016): ‘Self-interest, norms, and explanation’, in Mark Risjord (ed.): *Normativity and Naturalism in the Philosophy of the Social Sciences*, New York: Routledge, 212–229.

Zawidzki, Tad (2013): *Mindshaping: A New Framework for Understanding Human Social Cognition*, Cambridge, Ma: MIT Press.

22
Could Raatikainen have written otherwise?

Valtteri Arstila

I had the privilege of having Panu as the external examiner of my licentiate thesis in 2002. The following year, we met occasionally while both abroad, and it was during this time that I became more closely acquainted with his research. As anyone who knows him will attest, Panu's intellectual interests are both wide-ranging and deeply considered. One of the themes I have especially enjoyed reading about in his work concerns the notion of free will. With a short commentary on his work on this topic, I would like to extend my warm congratulations to Panu on the occasion of his 60th birthday.

Introduction

The question of free will is one of the most classic and debated topics in philosophy, and the related problem of mental causation has played a central role in the philosophy of mind since Descartes' interactionism (Raatikainen 2018). This question can be summarized as follows: can we act freely—that is, do we genuinely possess free will—or are our actions ultimately determined by natural laws or other deterministic factors?

The question sets free will against determinism. According to the former, we can genuinely choose between different courses of action, and our actions are truly our own. Hence, it is, for example, reasonable to assume that at least some of Panu Raatikainen's claims in his extensive writings on the subject (2007; 2010; 2013; 2015; 2017; 2018) were made due to his own choices. Given that the capacity for free will

Arstila, V. (2025). Could Raatikainen have written otherwise? In A. Laitinen, M. Keinänen, J. Reinikainen & A. Honkasalo (eds.). *Language, Truth, and Reality*. Tampere University Press. (pp. 280–295).
https://doi.org/10.61201/h5fbdk64

is often tied to moral responsibility for decisions and actions, Raatikainen would deserve praise (or blame) for his writings.

Determinism, traditionally seen as challenging free will, claims that prior occurrences and deterministic natural laws determine all events. According to this view, given certain initial conditions, the universe can evolve in only one way under such natural laws. Consequently, all human actions would be predetermined, including Raatikainen's writing process and the texts he produced. Thus, he could not have written otherwise.

Several solutions have been proposed to the question of free will. Considering the robustness of our intuitions regarding free will, the most straightforward solution might be to reject determinism. Supporting this view, current quantum physics suggests that natural laws may be inherently probabilistic rather than strictly deterministic. If, for instance, some version of quantum consciousness theories proves correct, we might consciously make choices that are not predetermined. However, most consciousness researchers regard such views of consciousness with considerable skepticism.

Another solution is to reconcile free will and determinism, as compatibilists have attempted to do. This approach hinges on how free will is defined. The term is used differently in various discussions, and some definitions are compatible with determinism. Many contemporary compatibilists argue, for instance, that even if brain activity determines our decisions, our actions can be free in the sense that they arise from our desires and values, rather than from external coercion or chance. (Lavazza 2019; Raatikainen 2017.)

In recent decades, discussions concerning free will have revolved around two themes. The first pertains to the traditional argument against free will, namely the claim that all our decisions are causally determined and that this is incompatible with free will. A more recent development on this theme centers on the (causal) *exclusion argument*, introduced and developed by Jaegwon Kim (1998; 2005) in particular. This argument is directed against substance dualism and non-reductive physicalism. Consistent with Raatikainen's position and argumentation, I shall confine my discussion to the latter, which asserts that mental states supervene on physical states but are not identical with them. The exclusion argument holds that all physical events can be explained by physical processes. As a result, non-reductive physicalism leads to *epiphenomenalism*—the view that mental states and events are mere byproducts of physical processes without being causally efficacious within the physical domain. For example, neural activity in the brain might cause physical actions (such as extending a finger) and mental states (such as deciding to extend a finger), but the mental state would not affect the brain processes. In short, the causal efficacy of mental states in the physical domain is an illusion.

The second recent theme concerns *scientific epiphenomenalism*. Unlike metaphysical epiphenomenalism, this view is grounded in neuroscience findings rather than metaphysical principles. It is inspired by the experiments conducted by Benjamin Libet and colleagues since the late 1970s (e.g., Libet 1985; Libet et al. 1979;

1993). These experiments showed that brain activity related to actions occurs before subjects consciously report deciding to act (e.g., extend a finger). The results were interpreted as evidence that the brain initiates action preparation before conscious decision-making, suggesting that actions result from unconscious processes rather than conscious will. That is, the decision to extend a finger is not the cause of the finger's movement. Nonetheless, this causal inefficacy of our conscious decisions does not (necessarily) arise from determinism but from the observation that the brain "decides" before our conscious decision.

Both themes are featured in Raatikainen's extensive work on free will. He has focused mainly on analyzing the exclusion argument, critiquing it in light of various theories of causation, and arguing that the argument is ultimately unsuccessful. However, he has also criticized interpretations of Libet's experiments, which often underpin scientific epiphenomenalism. In this paper, I examine Raatikainen's critiques and offer critical observations regarding his arguments.

The causal exclusion argument

The argument against free will

The causal exclusion argument is one of the most central modern arguments challenging the causal efficacy of mental states. In recent decades, it has also been debated in connection with the problem of free will. The argument can be presented through four premises:

1. *Non-reductive physicalism*: Mental states (e.g., beliefs, desires, intentions) supervene on physical states but are not identical with physical states. Kim (1998; 2005) presented this argument against non-reductive physicalism (as well as dualism), which the critics of the exclusion argument typically support. According to this theory, mental states depend on brain activity and supervene on physical states, but they are not reducible to physical states. Hence, mental states are distinct from their physical bases. Non-reductive physicalism is often defended on the grounds of the multiple realizability of mental states (Putnam 1967), and Raatikainen (2010; 2013; 2015) is no exception.
2. *Causal closure of the physical*: Every physical event has a sufficient physical cause. This assumption asserts that all physical events have a sufficient physical cause. For example, the act of extending a finger is fully accounted for by neural impulses in the brain, without the need for anything else, such as a mental state, to account for the occurrence.
3. *Causal exclusion*: If a physical event has a sufficient physical cause, it does not have another distinct cause unless it is a case of overdetermination. The third premise holds that if a physical cause can account for a physical event, it cannot have another independent cause unless there is causal

overdetermination. Overdetermination means that the same event results simultaneously from two separate causes, for instance, mental and physical causes. However, the fourth assumption of the exclusion argument rules out this possibility.

4. *No systematic overdetermination*: The causes of physical events are not systematically and continuously overdetermined. The fourth premise denies the possibility that the causes of physical events are systematically overdetermined. Thus, causal overdetermination, if it occurs, would be a rare and exceptional phenomenon and cannot be applied to ordinary physical events.

From the last three premises, we can conclude the following: Since all events in the physical world have sufficient physical causes, and no other causes exist for them, mental states cannot be causes of physical events unless they are identical with physical states. However, the first premise, which concerns non-reductive physicalism, denies the identity of mental and physical states. This leads to the conclusion of the exclusion argument: mental states are epiphenomenal, meaning they have no causal role in the physical world.

Raatikainen's critique of the exclusion argument

Raatikainen adopts a critical stance toward the exclusion argument and aims to show that it does not force one to accept reductive physicalism or the epiphenomenality of mental states. His critique is grounded in a conceptual analysis of causation: he argues that the exclusion argument erroneously relies on a conception of causation that cannot be applied universally across all scientific disciplines and causal relationships. As a result, he maintains that the exclusion argument cannot establish that mental states are causally inefficacious. [1]

Raatikainen (2013; 2018) follows the distinction made by Ned Hall (2004; see also Lewis 1973) and divides theories of causation into two main categories: those based on *production* and those based on *dependence*. He argues Kim holds a contemporary version of a production-based view of causation, which emphasizes the role of physical processes that involve real connectedness in the relationship between cause and effect. Raatikainen refers to this theory as the "causation-as-transmission view" because it is grounded in the idea that a causal relationship is based on the transfer of energy or some other physical magnitude (e.g., charge or momentum) from cause to effect.

The causation-as-transmission view in the context of the exclusion argument is problematic in two ways, however (Raatikainen 2013; 2018). First, he notes that the production-based theory of causation is not viable in the special sciences (such as biology, psychology, or history), where causal relationships do not rely on energy

[1] Raatikainen's critique parallels similar criticisms put forward independently by other scholars around the same time (e.g., List and Menzies 2009; Menzies and List 2010; Sober et al. 2007; Woodward 2008). In this festschrift contribution, I will focus solely on his views and arguments.

transfer or other physical processes. In these sciences, causation is grounded in complex dependency relations among the variables and phenomena under investigation. This observation suggests that the causation-as-transmission view is, at most, suited to fundamental physics. Indeed, when commenting on Phil Dowe's theory of causation, which is the best-developed causation-as-transmission view, Raatikainen (2018, 40) concludes that "it is undeniable that Dowe's theory directly applies only in the domain of fundamental physics." Therefore, it is unwarranted to assume that the causation-as-transmission view is a universally valid theory of causation, particularly regarding its suitability for assessing the causal efficacy of mental states.

The second problem is that if causation is assumed to require the transfer of a physical magnitude, the theory effectively presupposes that mental states cannot be causally efficacious unless they are identical to physical states. This presupposition, according to Dowe himself, requires a commitment to reductionism. Consequently, the exclusion argument becomes circular and incapable of supporting reductive physicalism over dualism or non-reductive physicalism (Raatikainen 2018).

Raatikainen regards the causation-as-transmission view as an "outdated idea" of causation (2010, 351) and favors James Woodward's interventionist theory of causation instead. This theory is a modern version of the counterfactual theory of causation—and thus a version of the causation-as-dependency view—and has proven to provide a successful framework for understanding causation in many special sciences. It approaches causal relationships from the perspective of manipulation: a causal relationship exists when we can manipulate one factor (the cause) and observe a change in another factor (the effect). Thus, causation is not a matter of a necessary sequence of events (in which some physical magnitude is transferred) but is based on causal dependency that can be revealed through interventions. This provides a practical definition of causation: if changing the cause alters the effect, there is a causal relationship.

In the interventionist theory, mental states can be causally efficacious if they meet the theory's criteria for causal dependency (Raatikainen 2010; 2015; 2018). For instance, if a person's desire to drink water changes and this change leads to physical action — such as reaching for a glass — this shows that the desire has causal relevance. The mental state functions here as a "difference-maker" alongside physical events without transferring energy or any other physical magnitude. This perspective challenges the exclusion argument's claim that mental states are causally inefficacious or have no causal relevance. What is particularly noteworthy in this argumentation is the distinction between the causes of an event, which are determined by causal dependency relations, and the physically sufficient conditions for an event to occur. Indeed, Raatikainen criticizes the exclusion argument for conflating these two notions. In the interventionist theory, these two aspects are separate because an event can have simultaneously both sufficient reasons to exist at a physical level and a causally relevant "difference-making" cause at the mental level.

Based on the interventionist theory, Raatikainen presents three arguments for why the exclusion argument is unsuccessful. The *proportionality argument* (2013; 2015) claims that mental states are causally efficacious, even when the underlying physical state suffices to explain the event. This is because causes and effects should be proportional: physical states are often too precise or complex to explain behavior, whereas mental states are closer to the level of the effect and, therefore, can provide a better explanation for our actions. It is also often the case that even if a physical state changes, the mental state can remain a cause of a bodily behavior, providing a more accurate explanation of the effect. Thus, the proportionality argument suggests that mental states can be genuine causes—understood as causal dependencies—even when the physical state is a sufficient cause in a physical sense.

Raatikainen (2013; 2015; 2018) also criticizes the exclusion argument's implicit assumption that mental and physical causes for an effect together would bring about causal overdetermination. From the perspective of the interventionist theory, overdetermination occurs only if two causes can be manipulated independently of each other and a change can be observed in the effect in both cases. However, according to non-reductive physicalism as assumed by the exclusion argument, mental states supervene on physical states. Hence, the independent causal examination of mental states is impossible: we cannot change a mental state without simultaneously changing the physical state. Therefore, within the interventionist theory, mental and physical states cannot lead to overdetermination. As a result, the soundness of the exclusion argument is, to say the least, questionable (Raatikainen 2015, 188).

Raatikainen's (2015) third argument against the exclusion argument in the context of the interventionist theory targets the causal exclusion premise, according to which a physical event can have only one causally sufficient cause (if we exclude the cases of overdetermination). Raatikainen emphasizes that in the context of interventionist theory, this premise is mistaken. He illustrates the issue with J. L. Mackie's (1965) example of a fire in a house, in which it was concluded that a short circuit, oxygen in the air, and flammable material are causally relevant for the fire to occur. None of these causes excludes the others because manipulating any of these factors affects whether the fire occurs. Thus, in the interventionist theory, multiple causes can be causally efficacious without resulting in causal overdetermination. This conclusion challenges Kim's exclusion premise and opens the possibility that mental states can be causally efficacious alongside physical causes without one excluding the other.

Observations on Raatikainen's critique

Raatikainen's critical analysis of the exclusion argument seeks to show that the argument fails in the context of any theory of causation. From the perspective of the causation-as-transmission view, the argument becomes circular, as it assumes from the outset the exclusion of mental causation. In contrast, within the context of the interventionist theory that Raatikainen supports, the premises of the exclusion argument—such as the claim that an event can only have one cause—are incorrect.

The interventionist theory offers a solution to the problems posed by the exclusion argument, as it allows mental states to be causally relevant without requiring them to be identical to physical causes.

This critique has sparked counterarguments. For example, Bram Vaassen (2021) examines Raatikainen's claim that exclusion arguments misuse the concept of causal sufficiency. According to Raatikainen, genuine causes are difference-makers, whereas causally sufficient phenomena are not necessarily causes. For example, a short circuit alone does not cause a fire, just as a thrown stone does not break a window without additional factors (such as the stone's mass and the window's fragility). Vaassen (2021, 10346) proposes replacing the concept of causal sufficiency with physical sufficiency, by which he means "any possible world in which the same fundamental laws of physics as in our world hold." After replacing causal sufficiency with physical sufficiency, Vaassen reformulates the exclusion argument and argues that mental causation does not happen under these terms.

There are also reasons to question the soundness of some of Raatikainen's arguments within the context of interventionist theory. For example, his interventionist solution can be criticized because it appears to assume the causal efficacy of mental states from the outset (at least sometimes). In fact, avoiding the conclusion that mental states have causal powers is nearly impossible: Since mental states supervene on physical states in non-reductive physicalism, mental states gain causal efficacy in the physical world by virtue of the physical states they supervene on. Similarly, every time a mental state changes, by definition, the underlying physical brain state also changes. Hence, different mental states have different causal effects; there are no situations in which a mental state changes, but its causal effects would remain the same. These points raise the question of whether the interventionist theory sidesteps the problem of free will by defining causation in a way that anticipates the desired conclusion. Recall that Raatikainen criticizes Kim's view of causation because it makes the exclusion argument circular. Here, a similar critique appears to apply to his own solution to the exclusion argument based on the interventionist theory.

Raatikainen's critique, which targets the premise that causal overdetermination must be excluded, raises a different concern. In non-reductive physicalism, changes in mental states are impossible without changes in physical states. For this reason, Raatikainen argues that the entire question of overdetermination is meaningless in this context. This critique, however, can be turned on its head because it implies that the causal efficacy of mental states cannot be empirically tested within the framework of interventionist theory. The problem of empirical testability arises because mental states, supervening on physical states, cannot be manipulated without affecting their physical basis. This is a problem for interventionist theory since it requires variables to be independently manipulated (see, e.g., Baumgartner 2018).

The above counterarguments to the interventionist solution to the exclusion argument highlight a more fundamental issue: interventionist causation theory does not offer a satisfactory answer to the traditional problem of free will. This claim may seem surprising, given the theory's success in many special sciences, including research on the human mind. Since the most typical difference-making causes of our actions are often intentional and psychological — that is, our mental states are perhaps the most closely correlated with our actions — interventionist theory would appear to provide a particularly suitable framework for examining free will.

Upon closer inspection, however, the interventionist theory does not resolve the problem of determinism, which many regard as central to the issue of free will. This is because the interventionist solution accepts multiple levels of causal explanation (Woodward 2008). Mental and physical causation operate at different levels, and there is no direct competition between them in their role as causal explanations. Mental states are causally efficacious at the psychological level, and physical events are causally efficacious at a physical level, such as the neurophysiological level, without these levels excluding each other.

This theory and Raatikainen's arguments have the merit of providing an account of mental causation that allows us to understand mental states as causally efficacious. Unlike contemporary compatibilists, who seek to reconcile determinism and free will by offering a concept of free will that is compatible with determinism, Raatikainen's solution is based on reevaluating the concept of causation.

However, since this solution is conceptual, it does not alter the nature of events at the physical level. This observation aligns with Vaassen's critique: even if we agree with Raatikainen that the exclusion argument conflates the concepts of causal sufficiency and cause, his account does not change what occurs at the physical level. While Raatikainen does not emphasize the conceptual nature of the interventionist solution, he would likely accept this assessment. This speculation is supported, for example, by his acknowledgment that the causation-as-transmission view may function at the level of fundamental physics.[2] Furthermore, when presenting his proportionality argument, Raatikainen (2013, 152–153) notes that from an interventionist perspective, "at least in some ways of conceptualizing the situation," the physical state is not the cause of a behavior, thus emphasizing the conceptual nature of his solution.

In short, and in relation to the question posed by the title of this article, the preceding means that based on his arguments, Raatikainen could not have acted differently. For instance, he could not have written otherwise and defended the exclusion argument in his articles.

[2] Raatikainen (2010, 351–352) also points out that the concept of causation is problematic in fundamental physics. However, some of the arguments he discusses rely on the interventionist theory, which does not provide impartial support for the claim. This is because the laws of fundamental physics are temporally symmetric, implying that causal relationships consistent with them would also be temporally symmetric. In contrast, under the interventionist theory, this is not the case: for instance, a fire is not a relevant factor in the occurrence of a short circuit.

Scientific epiphenomenalism

Varieties of scientific epiphenomenalism

The second central theme in contemporary discussions on free will concerns scientific epiphenomenalism. It originates particularly from the experiments conducted by Benjamin Libet and his colleagues starting in the late 1970s. These experiments challenged the traditional view of the causal role of conscious will in action, since the results suggested that neural processes associated with an action begin before a person becomes aware of deciding to act. If true, conscious intention does not (always) initiate action; instead, actions begin before conscious decisions are made. Hence, conscious decisions and intentions would be epiphenomenal in relation to the corresponding actions.

Scientific epiphenomenalism has received further support from new experimental findings over the past few decades. Some studies continue in the footsteps of Libet's experiments. For example, in a study by Chun Siong Soon and colleagues (2008), participants were asked, similar to the Libet experiments, to move their hand, but this time to press a button with either the left or the right hand. The results showed that brain activity indicates which hand the participant will use to press the button several seconds before the participant consciously becomes aware of their decision. These experiments support the idea that conscious decision-making is not the causal source of action; decisions are formed unconsciously in the brain before consciousness comes into play.

The challenge to free will has also expanded to include other types of research. For instance, David Milner and Mel Goodale's (1995) studies on the ventral and dorsal streams of the visual system—two pathways in which visual information processing can proceed—suggest that actions based on visual information do not necessarily require conscious visual experience. Daniel Wegner (2017), in turn, argues that the experience of free will itself is a kind of cognitive illusion, where conscious will is not causally effective. Instead, actions are determined by the brain and social factors.

In summary, Libet's studies now represent just one piece of evidence utilized to contest free will. Because scientific epiphenomenalism relies on neuroscientific studies, its forms vary depending on the experiments and results examined. Nonetheless, despite the differences among the studies, they communicate a consistent message: the nature of action and decision-making is not how it appears to be. Most of us believe that our conscious mind guides our decisions and actions, but these studies suggest this belief is mistaken.

It is important to note that these studies do not directly address the issue of free will, but focus on mechanisms related to decision-making, intentional action, and the conscious experience of these processes. Thus, the philosophically intriguing challenge lies in how these findings about mechanisms connect to the concept of free will.

Raatikainen's critique of the interpretation of Libet's experiments

Raatikainen (2015) presents three criticisms of Libet's interpretation of his experimental findings. For this reason, it is necessary to examine more closely what Libet's experiments measured and how these results were interpreted.

Libet aimed to investigate the temporal relationship between two types of cortical neural processes, namely, those related to conscious intention and voluntary action. In his experiments, participants looked at a clock face, where the hand (or a light) made a full rotation in about two and a half seconds. They were instructed to follow the clock hand and freely bend a wrist or a finger whenever they felt the urge to do so. Additionally, they were asked to report the exact position of the clock hand when they decided to bend their wrist or finger. Simultaneously, the electrical activity in their brains was measured using an electroencephalogram (EEG), which recorded activity in the motor cortex. Libet was particularly interested in the so-called readiness potential (RP, a slow negative shift in EEG readings) because it was known to occur when we make voluntary movements.

The results showed that the readiness potential began, on average, 550 milliseconds before the hand or finger movement. However, participants reported deciding to move their hand or finger only 350 milliseconds after the readiness potential had started, that is, about 200 milliseconds before the movement itself. In other words, brain activity was already underway significantly before participants decided to move. Libet interpreted these results as evidence that decision-making is not governed by conscious (free) will but results from unconscious brain processes.

This interpretation is supported by the difference in the readiness potential observed in Libet's experiments between voluntary and forced actions. Specifically, the EEG readings of participants who acted upon hearing a signal—such as bending their finger automatically after hearing a tone—did not show a readiness potential. Since readiness potential was present when subjects bent their fingers without being prompted by an external signal, Libet concluded that it is specifically linked to decision-making rather than automatic, forced, or reflexive actions.

Raatikainen presents three critical comments about Libet's findings and what can be concluded from them. First, he points out that the actions performed in Libet's experiments—moving a wrist or finger—are fundamentally different from the actions typically associated with the problem of free will. The actions in the experiments do not correspond to the complex, deliberative decisions requiring conscious reflection that we make in everyday life. Raatikainen emphasizes that, whereas such decisions take time and involve thorough consideration, the decision-making in Libet's experiments occurs (seemingly) instantaneously. Alfred Mele (2014) agrees with this point and concludes similarly to Raatikainen that Libet's experiments do not provide sufficient grounds to question the existence of free will at a more general level.

Second, Raatikainen criticizes the interpretation of Libet's results from the perspective of theories of causation. The problem lies in the implicit assumption that because something occurs in the brain before a conscious decision, this earlier event somehow excludes the causal efficacy of the later conscious decision. According to

Raatikainen, such an interpretation would lead to absurd conclusions. For example, throwing a stone could not be considered the cause of a window breaking because the stone-throwing itself is always preceded by some event that causes it. He contends that, based on this rationale, only the Big Bang as the first event could be regarded as the cause of later events.

Finally, Raatikainen questions the causal relevance of the readiness potential for the decisions made. This critique concerns the assumption that an unconscious readiness potential would precede every decision we make.[3] Given the assumption, the point Raatikainen makes is the following: if every decision—such as the choice between coffee, taking a walk, or watching a movie—were preceded by the same or similar readiness potential, this potential would not explain which option we choose, as it would occur regardless of the decision made. Thus, according to Raatikainen, the readiness potential cannot be a causally relevant factor for our actions.

Observations on Raatikainen's critique

Raatikainen's critique is ultimately unconvincing because it fails to consider the background from which scientific epiphenomenalism arises. In other words, it does not consider the context in which cognitive and neuroscientific research is conducted, nor how this context affects the interpretations of research findings.

Starting with the critique of the generalizability of Libet's experiments, Raatikainen correctly highlights the difference between decision-making in everyday situations and those faced by Libet's participants. In everyday situations, the things we decide on are often complex and require considerable deliberation, which takes time, whereas this is not the case for the subjects of Libet's experiments. However, one should not confuse the deliberation processes and the act of decision-making. Even if the first one is complex and takes time, the decision that leads to action could happen almost instantly. Thus, although the generalizability of Libet's findings to everyday life is not straightforward, the temporal difference that Raatikainen emphasizes between the decision-making in Libet's experiments and that in everyday contexts is not necessarily significant. In both cases, a decision is made, and Libet's experiments specifically addressed the moment of decision-making, not the deliberation preceding it.

Moreover, Libet's findings are further extended by the research of Soon and colleagues (2008), in which participants were asked to make a simple choice between pressing a button with either their left or right hand. Participants could freely choose which hand to use and then press the button. Brain activity was monitored using functional magnetic resonance imaging (fMRI). The results showed that the activity in specific brain areas could predict with 60 percent accuracy which hand

[3] Although this assumption is rarely explicitly stated, it is typically accepted by those who endorse scientific epiphenomenalism. This is because if it is rejected, then some of our actions could result from conscious decisions rather than preceding brain activity. That is, scientific epiphenomenalism would hold for some actions—something acceptable to most people—but our free will would be "safe" because there are actions that we freely choose to do.

the participants would choose up to 7–10 seconds before they made a conscious decision. Although the choices made in this study were still simple, unlike in Libet's experiments, this task required participants to choose between options. Yet unconscious brain processes, which researchers regarded as being associated with the decision, were active long before the conscious decision. Furthermore, in more everyday contexts, deliberation processes are likely influenced by additional factors. From the perspective of scientific epiphenomenalism, Richard E. Nisbett and Timothy DeCamp Wilson's (1977) findings are particularly interesting, as they suggest that we are often unaware of the basis of our conscious decisions and instead construct explanations retrospectively when asked.

Raatikainen's second critique is correct on a general level: if we deny the potential causal efficacy of phenomena occurring between the readiness potential and the resulting action, we risk ending up in an untenable situation where no event after the Big Bang could be the cause of subsequent events. However, the issue can also be examined from the perspective of the interventionist theory, which Raatikainen prefers. Suppose Libet's findings are accurate and that the readiness potential precedes every consciously made decision and is absent in the contrasting case of "forced" action. In that case, it is reasonable to claim that manipulating the readiness potential (for example, by inhibiting it with a magnetic pulse) would prevent participants from moving their hands or fingers. Therefore, within the framework of the interventionist theory, readiness potential can be considered a causally relevant factor in explaining participants' actions. In contrast, the causal relevance of conscious decision-making has not yet been experimentally demonstrated, and if our earlier analysis stands, it cannot be demonstrated.

Raatikainen's final critique was that readiness potential is not a causally relevant factor in explaining our actions because it does not explain choices, such as deciding between coffee, taking a walk, or watching a movie. This critique, however, can be addressed in at least three ways, two of which have already been discussed above. First, it is reasonable to distinguish the deliberation processes from the decision-making processes. Libet's experiments targeted only the moment of decision-making, and it is therefore unclear why the readiness potential should explain our choice between coffee and walking, for example. Therefore, while Raatikainen's observation is valid, it does not invalidate the significance of Libet's findings. Second, Raatikainen's claim that readiness potential is not a causally relevant factor in explaining our actions is questionable within the interventionist framework. As noted in response to his earlier critique, if Libet's results hold, manipulating readiness potential would influence behavior. Third, Raatikainen's (2015, 193, my translation) argument was based on the claim that "a similar readiness potential precedes all choices (such as going for a walk or to the movies)." This claim about the similarity of readiness potentials is important to his argument since if it were true, the readiness potential would not be a causally significant factor in explaining our specific choices. However, it is unclear why the same readiness potential would precede all choices, nor does Raatikainen justify this claim.

When assessing the last, unjustified claim, it is crucial to recognize that readiness potential is measured using EEG, which records electrical brain activity through electrodes placed on the scalp. The activity of individual neurons is not strong enough to be detected by this method. Rather, detecting readiness potential requires the synchronized activity of thousands or millions of neurons, and the recorded readiness potential does not differentiate between the roles of individual neurons or neuron groups in decision-making or movement preparation. This means that readiness potentials could well reflect our choices — that is, they may differ for different choices and actions — but our current methods cannot differentiate between the neuron groups activated in decision-making. This interpretation is supported by the assumption, which both reductionist and non-reductionist physicalist frameworks accept, that different mental states (e.g., choices) manifest as differences in brain activity.

In conclusion, based on the interpretation of Libet's findings and the above considerations, decision-making appears to be determined by unconscious processes. Thus, Raatikainen could not have refrained from writing his articles. However, the content of those writings might have been different, as Libet's experiments did not address the deliberation preceding decisions. As a result, Raatikainen could have also defended the exclusion argument, provided that Soon and colleagues' findings are not generalizable to more complex situations than those in their experimental design. Furthermore, regarding the findings of Soon and colleagues, it is worth repeating that the prediction accuracy was at most 60% and other times even lower. Given that a random guess would be correct 50% of the time, the reported accuracy is relatively modest. Therefore, it cannot be justifiably concluded from that study that our future decisions are predetermined.

Moreover, this examination of Libet's experiments would be incomplete without mentioning that not just their interpretation but the findings themselves are problematic according to current knowledge (for a recent review, see Dominik et al. 2024). For instance, readiness potential does not precede all voluntarily made actions. Additionally, EEG data analysis is problematic, as voluntary actions occur more frequently during certain phases of brain activity lasting a few seconds. As a result, when EEG data are "summed," the readiness potential appears to begin at least 200 milliseconds earlier than it actually does (Jo et al. 2013). Considering also that in Libet's experimental setup, the decision to move the hand appears to occur later than it does in reality, the conscious decision and readiness potential occur almost simultaneously. Consequently, the claim that unconscious processes preceding actions determine actions is no longer tenable — Raatikainen might have written nothing about free will in the first place.

Summary

The question of free will is one of the most classical and contested issues in philosophy. In recent decades, the discussion has centered on two key themes: Kim's exclusion argument and the scientific epiphenomenalism inspired by Libet's experiments. Raatikainen has addressed both themes but focused mainly on the exclusion argument. He has aimed to demonstrate that the argument fails within the framework of any theory of causation. The causation-as-transmission view favored by Kim renders the argument circular, while the assumptions underlying the exclusion argument are flawed in the context of the interventionist theory of causation. Raatikainen has engaged less extensively with Libet's experiments, but he is critical of the generalizability of the results and the causal relevance of the readiness potential.

I have presented critical observations regarding Raatikainen's solutions to both themes in this paper. Concerning the exclusion argument, I argued that the interventionist theory fails to address the traditional problem of the determined nature of our actions and renders it impossible to test whether mental states are causally efficacious. Furthermore, I suggested that the solution relies on an assumption that presupposes its conclusion, similar to how Kim's understanding of the exclusion argument relied on the causation-as-transmission view of causation, as criticized by Raatikainen. As for Libet's experiments, I contended that Raatikainen's critique is problematic in three ways, as it does not take into account the context in which the studies were conducted, nor how this context affects the possible interpretations of the results. Nonetheless, this is unlikely to substantially alter the situation as regards scientific epiphenomenalism, as the results of Libet and Libet-style experiments are nowadays regarded as problematic, irrespective of how they are interpreted.

If my critical claims about Raatikainen's view on the exclusion argument are sound, then it is reasonable to doubt whether he could have written differently than he did. The challenge from scientific epiphenomenalism does not constrain the act of writing, however. The final twist in the narrative lies in the observation that people tend to exhibit compatibilist intuitions when presented with concrete scenarios, and many believe that individuals are morally responsible in concrete situations where they could not have acted otherwise (Nichols 2011). Hence, given that the topic of this paper has not been some abstract scenario but Raatikainen's extensive work on the question of free will, which has enriched and expanded the discussion on the topic, I believe he deserves praise for this work (too), regardless of whether he might have written differently.

References

Baumgartner, Michael (2018): 'The Inherent Empirical Underdetermination of Mental Causation', *Australasian Journal of Philosophy* 96 (2): 335–50. URL = https://doi.org/10.1080/00048402.2017.1328451.

Dominik, Tomáš, Alfred Mele, Aaron Schurger & Uri Maoz (2024): 'Libet’s Legacy: A Primer to the Neuroscience of Volition', *Neuroscience & Biobehavioral Reviews* 157: 105503. URL = https://doi.org/10.1016/j.neubiorev.2023.105503.

Hall, Ned (2004): 'Two Concepts of Causation', in John Collins, Ned Hall, and L. A. Paul (eds.), *Causation and Counterfactuals*, The MIT Press, 225–76. URL = https://doi.org/10.7551/mitpress/1752.003.0010.

Jo, Han-Gue, Thilo Hinterberger, Marc Wittmann, Tilmann Lhündrup Borghardt & Stefan Schmidt (2013): 'Spontaneous EEG Fluctuations Determine the Readiness Potential: Is Preconscious Brain Activation a Preparation Process to Move?', *Experimental Brain Research* 231(4): 495–500. URL = https://doi.org/10.1007/s00221-013-3713-z.

Kim, Jaegwon (1998): *Mind in a Physical World: An Essay on the Mind-Body Problem and Mental Causation*, Cambridge, Mass.: MIT Press.

Kim, Jaegwon (2005): *Physicalism, or Something Near Enough*, Princeton University Press.

Lavazza, Andrea (2019): 'Why Cognitive Sciences Do Not Prove That Free Will Is an Epiphenomenon', *Frontiers in Psychology* 10: 326. URL = https://doi.org/10.3389/fpsyg.2019.00326.

Lewis, David (1973): 'Causation', *The Journal of Philosophy* 70 (17): 556–567. URL = https://doi.org/10.2307/2025310.

Libet, Benjamin (1985): 'Unconscious Cerebral Initiative and the Role of Conscious Will in Voluntary Action', *The Behavioral and Brain Sciences* 8(4): 529–566. URL = https://doi.org/10.1017/s0140525x00044903.

Libet, Benjamin, Curtis A. Gleason, Elwood W. Wright & Dennis K. Pearl (1993): 'Time of Conscious Intention to Act in Relation to Onset of Cerebral Activity (Readiness-Potential)', *Brain* 106(3): 623–642. URL = https://doi.org/10.1007/978-1-4612-0355-1_15.

Libet, Benjamin, E. W. Wright, B. Feinstein, and D. K. Pearl (1979): 'Subjective Referral of the Timing for a Conscious Sensory Experience: A Functional Role for the Somatosensory Specific Projection System in Man', *Brain* 102(1): 193–224. URL = https://doi.org/10.1093/brain/102.1.193.

List, Christian & Peter Menzies (2009): 'Nonreductive Physicalism and the Limits of the Exclusion Principle’, *Journal of Philosophy* 106(9): 475–502. URL = https://doi.org/10.5840/jphil2009106936.

Mackie, J L. (1965): 'Causes and Conditions', *American Philosophical Quarterly* 2(4): 245–264.

Mele, Alfred R. (2014): *Free: Why Science Hasn’t Disproved Free Will*, New York: Oxford university press.

Menzies, Peter & Christian List (2010): 'The Causal Autonomy of the Special Sciences', in Cynthia Macdonald & Graham Macdonald (eds.), *Emergence in Mind*, Oxford University Press, 108–128. URL = https://doi.org/10.1093/acprof:oso/9780199583621.003.0008.

Milner, A. D. & M. A. Goodale (1995): *The Visual Brain in Action*, Oxford: Oxford University Press.

Nichols, Shaun (2011): 'Experimental Philosophy and the Problem of Free Will', *Science* 331(6023): 1401–1403. URL = https://doi.org/10.1126/science.1192931.

Putnam, Hilary (1967): 'Psychological Predicates', in W.H. Capitan & D.D. Merrill (eds.), *Art, Mind, and Religion*, Pittsburgh: University of Pittsburgh Press, 37–48.

Raatikainen, Panu (2007): 'Reduktionismi, Alaspäinen Kausaatio Ja Emergenssi', *Tiede & Edistys* 32(4): 1–12.

Raatikainen, Panu (2010): 'Causation, Exclusion, and the Special Sciences', *Erkenntnis* 73(3): 349–363. URL = https://doi.org/10.1007/s10670-010-9236-0.

Raatikainen, Panu (2013): 'Can the Mental Be Causally Efficacious?', in Konrad Talmont-Kaminski & Marcin Milkowski (eds.), *Regarding the Mind, Naturally: Naturalist Approaches to the Sciences of the Mental*, Newcastle upon Tyne: Cambridge Scholars Publisher, 138–166.

Raatikainen, Panu (2015): 'Materialismi, neurotiede ja tahdon vapaus', *Ajatus* 75: 173–197.

Raatikainen, Panu (2017): 'Kumoaako tiede vapaan tahdon?', *Niin & Näin* 93(2): 144–145. URL = https://philarchive.org/rec/RAAKTV.

Raatikainen, Panu (2018): 'Kim on Causation and Mental Causation', *E-LOGOS* 25(2): 22–47. URL = https://doi.org/10.18267/j.e-logos.458.

Sober, Elliott, Lawrence Shapiro, G Wolters & P Machamer (2007): 'Epiphenomenalism: The Do's and the Don'ts', in *Studies in Causality: Historical and Contemporary*, Pittsburgh, PA: University of Pittsburgh Press, 235–264.

Soon, CS, M. Brass, H.J. Heinze, and John-Dylan Haynes (2008): 'Unconscious Determinants of Free Decisions in the Human Brain', *Nature Neuroscience* 11(5): 543–545. URL = https://doi.org/10.1038/nn.2112.

Vaassen, Bram (2021): 'Causal Exclusion without Causal Sufficiency', *Synthese* 198(11): 10341–53. URL = https://doi.org/10.1007/s11229-020-02723-y.

Wegner, Daniel M. (2017): *The Illusion of Conscious Will*, MIT press.

Woodward, James (2008): 'Mental Causation and Neural Mechanisms', in J. Hohwy and J. Kallestrup (eds.), *Being Reduced: New Essays on Reduction, Explanation, and Causation*, Oxford: Oxford University Press, 218–262.

List of Contributors

Valtteri Arstila, Professor of Philosophy, University of Turku
ORCID: https://orcid.org/0000-0001-6838-9946

Michael Devitt, Professor of Philosophy, Graduate Center,
City University of New York
ORCID: https://orcid.org/0000-0003-2298-2490

Zachary W. Goodsell, Assistant Professor of Philosophy, National University of Singapore

Jani Hakkarainen, Senior Lecturer, Tampere University
ORCID: https://orcid.org/0000-0002-7337-8034

Aleksi Honkasalo, Tampere University
ORCID: https://orcid.org/0000-0002-7721-1960

Markku Keinänen, PhD, Tampere University
ORCID: https://orcid.org/0000-0003-2289-0721

Inkeri Koskinen, Academy Research Fellow, University of Helsinki
ORCID: https://orcid.org/0000-0002-9060-7011

Tomi Kokkonen, Senior Lecturer, University of Helsinki

Anssi Korhonen, PhD, University of Helsinki
ORCID: https://orcid.org/0000-0002-7399-7281

Jaakko Kuorikoski, Professor of Philosophy, University of Helsinki
ORCID: https://orcid.org/0000-0001-5676-717X

Arto Laitinen, Professor of Philosophy, Tampere University
ORCID: https://orcid.org/0000-0002-4514-7298

Markus Lammenranta, PhD, University of Helsinki
ORCID: https://orcid.org/0000-0001-6976-8083

Genoveva Martí, Professor of Philosophy, ICREA and University of Barcelona
ORCID: https://orcid.org/0000-0003-1269-4655

Ilkka Niiniluoto, Professor of Philosophy Emeritus, University of Helsinki
ORCID: https://orcid.org/0000-0003-3162-5970

Renne Pesonen, PhD, Tampere University
ORCID: https://orcid.org/0000-0001-6425-5772

Sami Pihlström, Professor of Philosophy of Religion, University of Helsinki
ORCID: https://orcid.org/0000-0002-6410-8382

Jaakko Reinikainen, PhD, Tampere University
ORCID: https://orcid.org/0000-0002-4409-4964

David P. Schweikard, Professor of Philosophy, Europa-University Flensburg
ORCID: https://orcid.org/0000-0003-4851-2972

Teemu Toppinen, Associate Professor of Philosophy, Tampere University

Pasi Valtonen, PhD, Tampere University
ORCID: https://orcid.org/0000-0002-9470-1486

Vilma Venesmaa, Tampere University

Timothy Williamson, Emeritus Wykeham Professor of Logic, University of Oxford
ORCID: https://orcid.org/0000-0002-4659-8672

Juhani Yli-Vakkuri, Helsinki
ORCID: https://orcid.org/0000-0001-6929-0913

Mikko Yrjönsuuri, Professor of Philosophy, University of Jyväskylä
ORCID: https://orcid.org/0000-0003-0961-4853

Tampere University Press
Kalevantie 5 (PL 617)
33014 Tampereen yliopisto
FINLAND

email: tup@tuni.fi
https://edition.fi/tup
https://isni.org/isni/0000000507168743

Professor Panu Raatikainen's academic work in philosophy ranges from the philosophy of mind and language to truth and science, logic and mathematics. His signature contribution in any area is bringing argumentative transparency to foggy and complex debates.

This edited volume brings together leading philosophers who have encountered Panu Raatikainen in various academic occasions, as well as past and present colleagues. It celebrates Raatikainen's 60th birthday with a mosaic of papers discussing various topics handpicked from his long and prosperous career. The articles engage in critical, constructive dialogue with Raatikainen's key ideas and arguments focusing on three central topics in analytic philosophy: the nature of reality in which we live, the nature of truth of claims or thoughts about reality, and the nature of language expressing these thoughts and referring to reality. The volume also contains novel philosophical reflection on the nature of philosophy itself.

ISBN 978-952-359-073-1

TAMPERE
UNIVERSITY
PRESS

9 789523 590731