

Reports from the Department of Philosophy
University of Turku

TRUE COLORS, TIME AFTER TIME
Essays Honoring Valter Arstila

**Edited by Alexander D. Carruth, Heidi Haanila,
Paavo Pylkkänen & Pii Telakivi**



Reports from the Department of Philosophy
Vol. 52

TRUE COLORS, TIME AFTER TIME
Essays Honoring Valteri Arstila

Edited by
Alexander D. Carruth,
Heidi Haanila,
Paavo Pylkkänen,
Pii Telakivi

University of Turku
Finland

Copyright © 2024 Authors

Cover Image: Polaris Koi

SERIES EDITORS:

Olli Koistinen
Juha Räikkä

Department of Philosophy
University of Turku
FI-20014 Turku
Finland

ISSN 1457-9332
ISBN 978-951-29-9959-0 (PRINT)
ISBN 978-951-29-9960-6 (PDF/Online)

Painosalama Oy, Turku 2024

Contents

Foreword	5
I MIND AND ACTION	9
Hume on Phenomenal Consciousness <i>JANI HAKKARAINEN</i>	10
The Internal-External Divide and Husserl’s Phenomenology <i>ILPO HIRVONEN</i>	20
Theories of Consciousness Should Not Be Afraid of Dreams <i>MICHAL KLINCEWICZ</i>	53
What Theory Did Mary Develop? <i>KRISTJAN LOORITS</i>	70
Gesture and Literary Experience <i>KALLE PIHLAINEN</i>	83
Chalmers, Scrutability, and Phenomenal Truths <i>PANU RAATIKAINEN</i>	102
From Predicting Brains to Normative Authority <i>JANI SINOKKI</i>	106
Why Should I Care about My Choices? <i>SUSANNE UUSITALO</i>	132
Jenann Ismael on Causation and Free Will <i>SAARA WUOKKO</i>	138
II TIME AND TEMPORAL EXPERIENCE	149
Explaining Temporal Phenomenology: Hume’s Extensionalism and Kant’s Apriorism <i>ADRIAN BARDON</i>	150
From Here to Eternity: Two Absolutists (Newton, Clarke) and Two Relationalists (Leibniz, Kant of New Elucidation) on the Essence of Time <i>VISA HELENIUS</i>	167

Write Down Every Moment: An Argument for Perspectivalism about A- and B-Theories of Time	
<i>VILLE V. KOKKO</i>	187
Kant on Time	
<i>HEMMO LAIHO</i>	204
Dynamic Snapshots and Artificial Temporality	
<i>DAN LLOYD</i>	215
What If Time Were Like Other Minds	
<i>BRUNO MÖLDER</i>	228
On the Real Possibility of O’Brien: Historicized Transcendental Conditions and the Contingency of Sanity	
<i>SAMI PIHLSTRÖM</i>	240
III OTHER PHILOSOPHICAL REFLECTIONS	255
Quantum Properties as Potentialities in Bohm’s 1951 Book <i>Quantum Theory</i>	
<i>ALEXANDER D. CARRUTH & PAAVO PYLKKÄNEN</i>	256
Some Thoughts on Descartes on Immortality	
<i>OLLI KOISTINEN</i>	273
On Voluntary Servitude	
<i>EERIK LAGERSPETZ</i>	282
Frequently Asked Questions about Metaphilosophical Naturalism	
<i>HENRI PETTERSSON</i>	299
Understanding, Charity, and Interpretation	
<i>DAVID ROSENTHAL</i>	312
Can Computers Reason Like Medievalists? Building ‘Formal Understanding’ into the Chinese Room	
<i>LASSI SAARIO-RAMSAY</i>	332
Some Reflections on the Notion of Arbitrary Function	
<i>GABRIEL SANDU</i>	359
Tabula Gratulatoria	382

Foreword

Professor Valtteri Arstila will turn 50 years on 13th December, 2024. We know Valtteri not only as an excellent philosopher whose interests range widely and probe deeply, but also as a friendly colleague who likes to bring people together. To celebrate his birthday, we decided to edit this *Festschrift* in his honour. Valtteri has recently become a full professor of theoretical philosophy at the University of Turku, so this book is also to congratulate him for this great achievement.

Valtteri specialises in philosophy of mind, so his professorship means a significant strengthening of this area of philosophy in Finland. Philosophy of mind has been one of the most active areas of philosophy internationally during recent decades, with topics such as consciousness and artificial intelligence being actively discussed. Valtteri's particular interests lie in subjective time, theories of consciousness, time-consciousness, and philosophy of colour, and his publications on these topics have attracted international attention and appreciation. What distinguishes him from many other philosophers of mind is his strong engagement with empirical studies. Valtteri has excelled in this demanding field, which requires expertise in both philosophy and empirical research, with his work recognized as "a particularly successful example of empirically informed philosophy" (in earning him the Best Paper Award 2018 from *Australasian Journal of Philosophy*). He also collaborates with scholars from diverse disciplines. We will not provide a detailed description of Valtteri's many academic achievements here, but refer the reader to his home page <https://valtteri-arstila.net> which speaks for itself.

Valtteri's talents and achievements are not limited to his published contributions to academic philosophy. He is an inspiring teacher and has acted as mentor for many up-and-coming researchers in Finland. He designs his teaching to foster deep learning, provides personalised feedback, and encourages students to engage in discussion. As a PhD supervisor, he provides guidance in philosophical research and academic practices, while also offering essential support throughout the academic journey. In addition, Valtteri has played a significant role in organising events related to the philosophy of mind, creating and strengthening connections within the philosophy community. He is a trusted person to whom many turn to for academic advice. Altogether, Valtteri possesses an exceptional talent for building bridges—not only between academic disciplines but also between people. He is a dedicated and accomplished long-distance runner,

as well as a knowledgeable connoisseur of fine wines, champagnes and espresso martinis. Rumour has it that, on occasion, he has been known to entertain the audiences of karaoke bars in Turku, Helsinki and beyond!

While Valtteri has been based in Turku for most of his career, in recent years he was a University Researcher in the Mind and Matter profiling action at the University of Helsinki. It was in this context in Helsinki that many of the editors interacted especially closely with him. We held regular reading groups and seminars and typically continued the discussion afterwards in the beer house Kaisla in Kaisaniemi. Hopefully this *Festschrift* conveys to others some of the enthusiasm and intensity of the discussions we had with Valtteri. The editors of this volume are, at the time of publication, based at the universities of Turku and Helsinki. Those amongst us in Turku feel honoured to have Valtteri as a colleague and collaborator; those based in Helsinki miss him, but trust that our cooperation will continue in the future, *time after time*.

This book is structured in three sections. The first two—‘Mind and Action’ and ‘Time and Temporal Experience’—include papers focussed on issues particularly close to Valtteri’s own research specialisation. The final section contains papers on various further philosophical issues. The first section, ‘Mind and Action’, collects together contributions on a variety of topics such as consciousness, content, agency and normativity; encompassing approaches from within analytic philosophy, phenomenology and the history of philosophy. **Jani Hakkarainen**’s chapter explores a surprisingly under-studied element of Hume’s philosophy: his account of phenomenal consciousness. **Ipo Hirvonen** analyses the internalism-externalism debate in Husserl’s transcendental phenomenology with the help of two phenomenological ways to understand internality and externality. **Michal Klincewicz**’s paper engages with both empirical and philosophical evidence concerning whether dreams should be understood as phenomenally conscious, arguing that several disanalogies between normal wakeful consciousness and dreaming suggest that a *sui generis* category is required to account for dreams. **Kristjan Loorits**’s essay provides a critical reflection on what sort of lesson we ought to take from Jackson’s famous ‘Mary’ thought experiment, by focussing closely on the nature and content of the kind of theory of colour experience Mary could develop inside her black and white room. **Kalle Pihlainen** draws on phenomenology to explore the role of gesture, embodiment and agency in literature and literary engagement. **Panu Raatikainen** analyses David Chalmers’s *scrutability thesis*, which plays a key role in his

epistemic two-dimensional semantics. Raatikainen argues that phenomenal truths are given too prominent a role in Chalmers's theory, which makes the scrutability thesis less plausible. **Jani Sinokki**'s paper explores the links between the Predictive Processing framework and how norms gain their authority, arguing that the former—in combination with some relevant psychological explanation—can help shed light on the latter. **Susanne Uusitalo** examines the importance of agency and autonomy in human action, examining concrete examples from gambling behaviour, and linking these issues to increasingly pressing questions concerning artificial intelligence. The final paper in this section, written by **Saara Wuokko**, focuses on the role of causation in Jenann Ismael's arguments for free will and shows that some of Ismael's causation-related arguments for free will are insufficient.

The second section, 'Time and Temporal Experience', collects together papers on topics including the nature of time itself; of our experience of time and of historicity and temporality more broadly; approaching these issues from a variety of perspectives including historical approaches. **Adrian Bardon** examines Hume's and Kant's responses to the question of how empiricists can explain temporal phenomenology and indicates how Arstila's dynamic snapshot theory provides both an alternative to Humean extensionalism and an intriguing way to address the gap left by Kantian apriorism. **Visa Helenius**'s contribution compares and contrasts different accounts of the nature of time from two absolutists—Newton and Clarke—and two relationists—Leibniz and Kant. **Ville V. Kokko** draws from Stephen King's novella *The Langoliers* to explore philosophical theories of time. He argues for a deflationary view about A- and B-theories of time, suggesting that we simply have A- and B-perspectives, which describe the same reality but highlight different aspects of it. **Hemmo Laiho**'s paper provides an overview of Kant's account of the nature of time, with a focus on the Transcendental Aesthetic and the idea that time is a subjective representation of temporal order. **Dan Lloyd**'s paper takes Arstila's model of subjective time—the *dynamic snapshot view*—as a starting point and applies it to analyse artificial temporality in a current large language model, GPT-4. **Bruno Mölder** draws on resources from social cognition and other minds to frame one possible approach to clarifying the nature of temporal experience. **Sami Pihlström** examines the transcendental conditions of truth, coherent thought, and sanity, drawing on an example of O'Brien, a figure from George Orwell's dystopian novel *Nineteen*

Eighty-Four, to demonstrate that without the concept of truth, we cannot make sense of our lives or remain sane.

The final section brings together papers that touch on issues within philosophy of science, logic, philosophy of language, political philosophy and more besides. **Alexander Carruth** and **Paavo Pylkkänen**'s joint paper examines the potentialities-centred interpretation of quantum theory developed by David Bohm in his 1951 book *Quantum Theory* and aims to situate it within a general ontological framework. **Olli Koistinen** scrutinises Descartes's idea of immortality: the non-destructibility of the soul depends on its being a pure substance, and pure substances are, by nature, indestructible. Koistinen delves into what Descartes means by a pure substance and why it should be indestructible. **Eerik Lagerspetz** discusses the question of whether one can freely give up one's autonomy, that is, freely choose to place oneself in absolute slavery, exploring the place this puzzle has held throughout the history of modern political thought. **Henri Pettersson**'s chapter presents an outline of metaphilosophical naturalism, aiming to clarify this much-discussed framework and dispel some common misconceptions concerning its nature and commitments. **David Rosenthal** explores what it takes to understand and misunderstand one another in communication. He argues that to understand accurately, one must interpret what the other person says as charitably as possible—but the principle of charity is more complicated than often recognized. **Lassi Saario-Ramsay** asks whether computers can distinguish valid from invalid arguments, drawing on medieval theories of formal consequence, contemporary methods of formalising conceptual-formal validity, and John Searle's Chinese Room Argument. **Gabriel Sandu**'s paper builds on his earlier work (both individually and with Jaakko Hintikka) in the debate surrounding Frege's and Russell's foundational program (logicism) and the notion of an arbitrary function.

The editors wish to express their gratitude to the Philosophy Unit of the University of Turku for supporting the publication of this volume. It has been a pleasure to bring these papers together, and have the opportunity to see the *true colours* of the various topics discussed herein *shining through*, and so of course we also extend our thanks to the authors for making this volume possible, providing us with a *dynamic snapshot* of philosophy in Finland and beyond.

In Turku and Helsinki, November 2024

Alexander D. Carruth, Heidi Haanila, Paavo Pylkkänen & Pii Telakivi

I

MIND AND ACTION

Hume on Phenomenal Consciousness

JANI HAKKARAINEN

Introduction

From the present point of view, it might be a surprising observation that Hume says very little about phenomenal consciousness and does not use even the term “consciousness” very often. Nonetheless, he is clearly aware of the concept of phenomenal consciousness. His awareness is already shown by three key passages from the first Book of Hume’s first work, *A Treatise of Human Nature* (1739):

1. For since all actions and sensations of the mind are known to us by consciousness, they must necessarily appear in every particular what they are, and be what they appear. Every thing that enters the mind, being in *reality* a perception, ’tis impossible any thing shou’d to *feeling* appear different. This were to suppose, that even where we are most intimately conscious, we might be mistaken. (T 1.4.2.7; SBN 190).¹
2. consciousness never deceives (EHU 7.14; SBN 66).
3. The only existences, of which we are certain, are perceptions, which being immediately present to us by consciousness ... no beings are ever present to the mind but perceptions (T 1.4.2.47; SBN 212).

This raises the question of Hume’s view of phenomenal consciousness. What is his account of it? What could we say about his view even though Hume says so little about phenomenal consciousness? Accordingly, the question of this paper is, how to understand this aspect of Hume’s philosophy?

¹ Following standard conventions, references to Hume’s *A Treatise of Human Nature* will be given in the text in two parts. The first is the Norton and Norton edition (New York: Oxford University Press 2000), abbreviated as ‘T’ followed by book, part, section, and paragraph numbers. The second is the L. A. Selby-Bigge edition, revised by P. H. Nidditch (Oxford: Clarendon Press, 1978), abbreviated as ‘SBN’ followed by page number.

There is surprisingly little scholarship done on this question, even though it is connected to many central issues of Hume scholarship, such as the much-discussed Hume's theory of belief, as will be seen below. Perhaps this state of affairs is due to the scant textual evidence we have about Hume's account of phenomenal consciousness. Hume scholars have had so little text to work on here. Therefore, any interpretation of Hume on phenomenal consciousness has to be more like a rational reconstruction than a historical reconstruction.

Furthermore, I need to admit that accounting for phenomenal consciousness is a problem for an interpretation of what Hume calls "perceptions" that I have defended earlier (Hakkarainen 2012a, b, 2015). As is well-known, Hume divides "perceptions" into "impressions" and "ideas": roughly, sense-perceptions, bodily feelings (bodily pains and pleasures), passions (e.g., love and hate), and thoughts about them. Another distinction of his is between "simple" and "complex perceptions": perceptions that do not divide and do divide into parts (T 1.1.1.2; SBN 2). My interpretation of perceptions is limited to simple perceptions: they are particular simple properties or qualities, such as a navy blue or an intense pain in one's tooth, that is to say, tropes in the present day terms.

My elaboration of this interpretation, proposed in this paper, is, initially, that phenomenal consciousness is an aspect of perceptions (i.e., impressions and ideas). Nowadays, I also prefer emphasizing the point that perceptions are qualities to them being properties (cf. Keinänen & Hakkarainen 2024). Here I restrict my discussion to simple perceptions; I put complex perceptions aside.

The structure of the paper is as follows. First I summarise my published interpretation that simple perceptions are tropes and show its problems. In the second chapter, I discuss Donald C. Ainslie's (2015, sec. 6.6) "package view of perceptions". I finish the paper by proposing a solution to the problem that my published interpretation faces. I elaborate upon it with the help of Ainslie's interpretation.

1. Perceptions Are Particular Properties

Earlier (Hakkarainen 2012a, b, 2015) I have defended the interpretation that perceptions are nothing but particular entities in the category of properties. They are qualities and qualitatively simple: they present only a single determinate uniform quality. To illustrate, if I have a simple visual impression of navy blue, there is a determinate uniform navy-blue present to me. To put it in present-day terms, perceptions are mereologically simple or mereological atoms: they do not

divide into parts. Recall that I am talking about simple perceptions. So perceptions are particular simple qualities, that is to say, “tropes” in our terms (see Keinänen & Hakkarainen 2024). I have also argued for an interpretation that perceptions are causally inert (Hakkarainen 2011). Illustrations of existing Humean perceptions would be the following: there is a navy blue, there is a crimson, there is a hot, there is a rose aroma, there is an intense toothache and there is (a feeling of) pride. Hume believes, indeed, that these kinds of perceptions are entities (e.g., T 1.4.2.47; SBN 212).

Now I see, however, that there are severe problems for this interpretation. First, how to account for Hume’s constant talk about the *actions* of the mind, perceptions as actions and especially beliefs as *acts*? A telling instance of this is what Hume says about the latter: “belief is an act of the mind arising from custom” (T 1.3.9.13; SBN 114; see also T 1.3.5.7 (SBN 86) and T 1.3.16.8; SBN 178). Second, how to account for the fact that perceptions are *conscious* when a mind consists ultimately of nothing but perceptions, which are rather objects or contents than acts? There are no other mental item types in Hume’s bundle theory of the mind than perceptions and minds:

But setting aside some metaphysicians of this kind, I may venture to affirm of the rest of mankind, that they are nothing but a bundle or collection of different perceptions, which succeed each other with an inconceivable rapidity, and are in a perpetual flux and movement. (T 1.4.6; SBN 252).

Hence there are no acts, actions, states of consciousness or their subjects *in addition to* perceptions and the mind as the bundle of perceptions in this theory. If there are acts, actions or states of consciousness in Hume’s view, they are not distinct from perceptions.

But are there then two kinds of perceptions: mental acts and their objects or contents? Third, how to account for “force and vivacity”? As is well-known, force and vivacity are those which, according to Hume (at least in the *Treatise* 1), explain the difference between impressions and ideas, on the one hand (T 1.1.1.3, SBN 2-3), and believed and merely entertained perceptions, on the other:

Thus it appears, that the *belief* or *assent*, which always attends the memory and senses, is nothing but the vivacity of those perceptions they present; and that this alone distinguishes them from the imagination. (T 1.3.5.7; SBN 86)

This third problem is reinforced by the fact that in my published view perceptions are nothing but contents or objects and Hume holds that the content of an idea does not change along with its force and vivacity:

All the perceptions of the mind are of two kinds, *viz.* impressions and ideas, which differ from each other only in their different degrees of force and vivacity. Our ideas are copy'd from our impressions, and represent them in all their parts. When you wou'd any way vary the idea of a particular object, you can only encrease or diminish its force and vivacity. If you make any other change on it, it represents a different object or impression. [...] So that as belief does nothing but vary the manner, in which we conceive any object, it can only bestow on our ideas an additional force and vivacity. (T 1.3.7.2; SBN 94-5).

Functionalism might be an option to explain the distinction between impressions and ideas. However, a functionalist account of force and vivacity is not very satisfactory, since it has the problem of explaining Hume's constant phenomenal description of force and vivacity:

I conclude, by an induction which seems to me very evident, that an opinion or belief is nothing but an idea, that is different from a fiction, not in the nature, or the order of its parts, but in the *manner* of its being conceiv'd. But when I wou'd explain this *manner*, I scarce find any word that fully answers the case, but am oblig'd to have recourse to every one's feeling, in order to give him a perfect notion of this operation of the mind. An idea assented to *feels* different from a fictitious idea, that the fancy alone presents to us: And this different feeling I endeavour to explain by calling it a superior *force*, or *vivacity*, or *solidity*, or *firmness*, or *steadiness*. (T 1.3.7.7; SBN 629)

There is a further complication of explaining the fact that in Hume's view we can be mistaken whether a perception is an impression or an idea:

Thus in sleep, in a fever, in madness, or in any very violent emotions of soul, our ideas may approach to our impressions: As on the other hand it

sometimes happens, that our impressions are so faint and low, that we cannot distinguish them from our ideas. (T 1.1.1.1, SBN 2)

When any of these passions [impressions] are calm, and cause no disorder in the soul, they are very readily taken for the determinations of reason [ideas] (T 2.3.3.8; SBN 417).

It is force and vivacity, however, that is the only circumstance that distinguishes impressions and ideas. Hume is also committed to the view that at least the content of perceptions is transparent to the mind when we are conscious of it, as was seen in the first quotes of the paper. So how could one be mistaken about the degree of the force and vivacity of a perception if it were part of the content of the perception?

My published interpretation faces therefore not only the problem of phenomenal consciousness but also the problems of explaining actions or acts of the mind, force and vivacity and the possibility of making a mistake as to whether a perception is an impression or an idea.

2. The Package View

Partly to solve these problems, Donald C. Ainslie (2015, sec. 6.6) has proposed the package view of perceptions (cf. Broughton 1992). According to Ainslie, sensory perceptions, for instance seeing navy blue, are packages consisting of the actions of awareness and the image contents of awareness. In his view, a perception is the awareness-of-an-image-content package, in which the action of awareness or phenomenal consciousness and its image content (object) are *aspects* of the perception. These aspects are not numerically or really but only rationally distinct from each other and the perception. They are really one and the same although we can distinguish them in our thinking.

Ainslie suggests that both the image content and the action of awareness are determinate: for example, both a navy blue and the visual sensory awareness of it are determinate. When we are talking about beliefs, the force and vivacity of believed perceptions is a *dimension* of manner of awareness. So Ainslie accounts for belief by a dimension of manner of awareness. There can be more than one such a dimension, such as the dimension of voluntariness and involuntariness.

Ainslie disagrees with Wayne Waxman (1994, 10 and 18), according to whom perceptions are objects of the acts of consciousness that are not perceptions. Waxman distinguishes acts from perceptions. Nonetheless, Ainslie thinks that

Waxman's account is a minority view. According to Ainslie, the dominating account is the object view, according to which perceptions are objects (e.g., Garrett 1997, Baxter 2008). Anyway, it seems to me that Waxman's interpretation is not compatible with Hume's official statement of the bundle theory, which I quoted above.

There is some textual evidence for Ainslie's package account, especially the slightly obscure passage from T 1.3.8.16 (SBN 106):

In thinking of our past thoughts we not only delineate out the objects, of which we were thinking, but also conceive the action of the mind in the meditation, that certain *je-ne-scai-quoi* [I-don't-know-what], of which 'tis impossible to give any definition or description, but which every one sufficiently understands.

In this passage, Hume suggests that the action of the mind is something we do not know what. He also states about belief as follows: "So that as belief does nothing but vary the manner, in which we conceive any object, it can only bestow on our ideas an additional force and vivacity." (T 1.3.7.5; SBN 96; see also T 1.3.7.7 quoted above). One might suggest, perhaps, that Hume's talk about the manner of conception refers to Ainslie's awareness aspect.

It seems to me, however, that this textual evidence is suggestive at best. Ainslie's account is therefore rather rational than historical reconstruction. It "saves" Hume from problems internal to his system like accounting for mental acts, beliefs and the possibility of confusing impressions and ideas with each other.

Before I go on, a few things need to be clarified about Ainslie's view. The talk about mental "actions" or "acts" does not imply that there is an agent who commits these acts or actions: there are only the awareness-of-an-image-content packages and the mind as their bundle. Neither does it mean that all perceptions are active, not passive in the sense of being receptive. Ainslie does not intend to imply either that at least some perceptions involve genuine causal powers, whether active or passive. Rather, perceptions are determinate actualizations of the different "manners" of being aware or conscious: sensing, feeling and thinking. These may be passive in the sense of being receptive like sensing. Some perceptions, impressions in particular, are involuntary, which is a dimension of being aware in Ainslie's view.

Finally, awareness or consciousness as such does not involve any self-consciousness. There is not always awareness of awareness. Awareness of awareness might require a *secondary idea* about the perception, that is to say, an idea that is about another perception:

we can form secondary ideas, which are images of the primary; as appears from this very reasoning concerning them. [...] Ideas produce the images of themselves in new ideas; but as the first ideas are supposed to be derived from impressions, it still remains true, that all our simple ideas proceed either mediately or immediately, from their correspondent impressions. (T 1.1.1.11; SBN 7-8)

So, Ainslie suggests that the awareness aspect of perceptions is not always transparent to the mind in Hume's view. This kind of transparency requires consciousness of the awareness aspect of perceptions.

3. Perceptions Are Particular Properties View Elaborated

On Ainslie's account, perceptions are *episodes* of awareness (i.e., mental events) or mental states rather than objects (of mental acts) (Ainslie 2015, 211). This seems to be at odds with my view that they are particular qualities and properties. However, next I shall show that I can elaborate on my view with the help of the package view: to have its advantages with the merits of my earlier view.

The idea is that I can hold onto my view that perceptions are nothing but particular simple qualities when I qualify this statement: perceptions are particular simple qualities *qua* contents (under this aspect of perceptions). I can complement this with the awareness aspect. Under that aspect, perceptions are *actualizations of different determinate manners of being aware*: forcible and lively manners actualized (impressions, believed ideas of the memory and imagination), manners that are not such actualized (ideas of the imagination), involuntary manners actualized (impressions in given conditions) and voluntary manners actualized (e.g., ideas in calculations). This is how Hume could account for phenomenal consciousness.

I model this elaboration on Hume's view of the simplicity of simple ideas:

And of this we may be certain, even from the very abstract terms simple idea. They comprehend all simple ideas under them. These

resemble each other in their simplicity. And yet from their very nature, which excludes all composition, this circumstance, in which they resemble, is not distinguishable nor separable from the rest. (T 1.1.7.7n5; SBN 637)

To illustrate Hume here, the simplicity of a navy-blue idea is an aspect of this navy-blue idea, since it resembles every other simple idea. In a similar manner, I can explain the things that have been extremely difficult to account for in my particular properties view. Conscious perceptions are perceptions that involve an awareness aspect. Beliefs are construed as perceptions having the forcible and lively awareness aspect. Confusing impressions and ideas is possible, for instance confusing hallucinating a tiger with seeing a tiger, since the relevant aspect of perception is not always transparent to the mind, even though the content would be always transparent. Finally, it is coherent for me to say that at least sometimes being aware of the awareness aspect of perceptions requires a secondary idea.

What should I say now about the ontological category of perceptions? For Hume, perceptions are not *modes* of the thinking substance as they are for Descartes and Locke, for instance. In Hume's view, perceptions are existentially independent entities and modes are existentially dependent on substances (T 1.4.5.5; SBN 234). My view is that perceptions are properties (qualities). Ainslie thinks that perceptions are episodes of awareness (e.g., seeing navy blue): mental events. My new proposal is that it is not necessary to construe perceptions as episodes or events even if they have the aforementioned aspects.

Accordingly, I suggest that each conscious perception is being aware in a certain determinate manner of a content: for instance, being aware in a forcible and lively manner of a navy blue. It is a dual-aspect particular property (possibly multi-aspect). Aspect 1 is actualization of being aware in the determinate forcible and lively manner without this awareness being part of the content of the perception.² The second aspect of this perception is its content: a determinate navy blue (a quality and property).

To distinguish these two aspects requires abstraction: making a distinction of reason (T 1.1.7.17-8). Following Hume's view of the simplicity of simple ideas, the aspects are really inseparable or not really distinct: one cannot be had without the other (Hume's Separability Principle ("Our perceptions are all really different, and separable, and distinguishable from each other, and from every thing else

² Still, it can be part of the content of a secondary idea of the perception. For example, when I remember being aware of a toothache in the forcible and lively manner.

which we can imagine" (T 1.4.5.27; SBN 245-46)) does not apply to the aspects). The perception as a dual-aspect, or possibly multi-aspect, property is mereologically simple, and its content is qualitatively simple because awareness is not in the content. Its two aspects are abstract in this specific sense. Still abstracting does not create them: they are real aspects of perceptions independent of abstraction. They can be distinguished due to the different resemblance relations that hold amongst perceptions.

This metaphysics of perceptions fits better with Hume's philosophical landscape in early modern philosophy, where one does not find ontologies in which episodes (or events) are a more fundamental category than objects. The metaphysical picture is thus as follows: there are perceptions as these dual-aspect particular simple properties, which are existentially independent entities, for example there is a being aware in a forcible and lively sensory manner of a navy blue or a toothache. Minds are composed of them. Yet one can say that perceptions are had or acted by the mind, when one understands this along the lines of the bundle account of the mind (vs the substance-mode view of minds and ideas).

Tampere University

References

- Ainslie, D. (2015), *Hume's True Scepticism*, Oxford University Press, Oxford.
- Baxter, D. L. M. (2008), *Hume's Difficulty: Time and Identity in the Treatise*, Routledge, London and New York.
- Broughton, J. (1992), "What Does the Scientist of Man Observe?", *Hume Studies* 18, 155–168.
- Fisher, A. R. J. and A-S. Maurin (eds.) (2024), *The Routledge Handbook of Properties*, Routledge, London.
- Garrett, D. (1997), *Cognition and Commitment*, Oxford University Press, Oxford.
- Guigon, G. and G. Rodriguez-Pereyra (eds.) (2015), *Nominalism about Properties*, Routledge, London.
- Hakkarainen, J. (2011), "Hume's Argument for the Ontological Independence of Simple Properties", *Metaphysica* 12 (2), 197-212.
- Hakkarainen, J. (2012), "A Third Type of Distinction in the Treatise", *Hume Studies* 38(1), 55–78.
- Hakkarainen, J. (2012), "Hume as a Trope Nominalist", *Canadian Journal of Philosophy* 42(S1), 55–66.
- Hakkarainen, J. (2015), "Hume on Spatial Properties", in G. Guigon and G. Rodriguez-Pereyra (2015), 79–94.
- Hume, D. [1739–40] (1975), *A Treatise of Human Nature: Second Edition*, Edited by L. A. Selby–Bigge and revised by P. H. Nidditch, Clarendon Press, Oxford.

- Hume, D. (2007), *A Treatise of Human Nature: A Critical Edition*, 2 vols, Edited by D.F. Norton and M.J. Norton, Clarendon Press, Oxford.
- Keinänen, M. and Hakkarainen, J. (2024), “Trope Bundle Theories of Substance”, in A. R. J. Fisher and A-S. Maurin (2024), 239-249.
- Waxman, W. (1994), *Hume’s Theory of Consciousness*, Cambridge University Press, Cambridge.

The Internal-External Divide and Husserl's Phenomenology¹

ILPO HIRVONEN

Various interpretations of Husserl have been presented in relation to the internalism-externalism debate. The debate concerns the question whether linguistic and mental content can be determined by features that are not only internal but also external to the subject. Besides different internalist and externalist interpretations of Husserl, there are interpretations that reject both internalism and externalism as frameworks for understanding Husserl. The main reason not to commit to either externalist or internalist interpretations seems to be that the internal-external divide, which is pertinent to internalism and externalism, is purportedly inapplicable to the transcendental framework of Husserl's phenomenology. Yet, it is not obvious how the concepts of internality and externality are understood when they are deemed inapplicable in this context, which reflects the ambiguity of the concepts in the internalism-externalism debate itself. The aim of this article is to clarify the internal-external divide and its rejection in these interpretations of Husserl. The reason why the internal-external divide has been deemed inapplicable to the transcendental framework of Husserl's phenomenology is first located to a spatial sense in which the concepts of internality and externality are usually understood. It is then argued that the internal-external divide can be rendered applicable to the transcendental framework of Husserl's phenomenology by reconfiguring internality and externality in relation to the subject's point of view. In addition to Farkas' proposal that conceives the internal-external divide in phenomenal terms, two phenomenological ways of understanding internality and externality are considered. The main problem in both Farkas' proposal and the two phenomenological proposals is that they preclude some forms of externalism. This

¹ This paper has undergone peer review.

article solves the problem by using the two phenomenological proposals to modify Farkas' proposal so that a broader range of externalism is enabled.

1. Introduction

There have been various attempts to understand Husserl in relation to the internalism-externalism debate in philosophy of mind and language. The debate typically concerns the question whether the meaning or content of the subject's linguistic expressions and mental states may depend on features that are external to the subject. Traditionally, Husserl has been interpreted as an internalist within this debate (see e.g., McIntyre 1984; 1986; Dreyfus & Hall 1984; Keller 1999, pp. 3–9; Carman 2003, pp. 83–84ff.; Spear 2013). Although many contemporary Husserl scholars have abandoned the internalist interpretation, they remain divided when it comes to the question whether Husserl is then better understood as a kind of externalist (Beyer 2000; 2013; Smith 2008; Crowell 2008; 2013) or as a philosopher whose thinking cannot be understood within the internalism-externalism framework (Zahavi 2004; 2008; 2017, pp. 115–120; Murchadha 2003; 2008; Man-To 2014). The main reason for endorsing the latter alternative rather than committing to either internalist or externalist interpretations seems to be the idea that the internal-external divide, which is pertinent to the internalism-externalism framework, cannot be applied to the transcendental framework of Husserl's phenomenology. Yet, it is not clear what is meant by the notions of internality and externality when they are abandoned in this context. This unclarity seems to reflect the ambiguity of the internal-external divide in the internalism-externalism debate itself (see Farkas 2003; 2008; Gertler 2012; Gomes & Parrott 2021; Bruckner 2022). The aim of this article is to first identify the reason why the internal-external divide has been deemed inapplicable to the transcendental framework of Husserl's phenomenology and then clarify the internal-external divide for both Husserl's phenomenology and the internalism-externalism debate. The reason for rejecting the internal-external divide is located to a spatial sense in which the internal-external divide has usually been understood. It is then proposed that to be rendered applicable to the transcendental framework of Husserl's phenomenology the divide should rather be understood in a sense that reconfigures internality and externality in relation to the subject's point of view. After introducing Farkas' (2003; 2008) proposal that conceives the internal-external divide in relation to the subject's point of view, two phenomenological ways of understanding internality and externality are considered (Smith 2008; Crowell

2008; 2013). The key problem in both Farkas' proposal and the two phenomenological proposals is that they preclude some forms of externalism, indicating that they cannot provide general criteria for the internal-external divide in the internalism-externalism debate. To solve this problem, the two phenomenological ways of understanding the internal-external divide are used to modify Farkas' proposal. As a result, this article introduces a modified proposal for defining the internal-external divide that enables a broader range of externalism than the two phenomenological proposals and Farkas' proposal.

2. Internalism and Externalism

Internalism and externalism designate many different views in philosophy. In this article, they are understood in the context of philosophy of language and mind as opposing views about meaning or content. The reason is that these are the kinds of internalism and externalism mainly under discussion in the literature about Husserl's relation to the internalism-externalism debate (see e.g., McIntyre 1984; 1986; Keller 1999; Beyer 2000; 2013; Murchadha 2003; 2008; Zahavi 2004; 2008; 2017; Rowlands 2003, pp. 56–63; Crowell 2008; 2013; Smith 2008; Alweiss 2009; Spear 2013). Most importantly, the argument that Husserl cannot be understood within the internalism-externalism framework is primarily directed at these forms of internalism and externalism (Zahavi 2017, pp. 115–119).

At its heart, the internalism-externalism debate about content is a philosophical disagreement about individuation (see e.g., Kallestrup 2012, p. 60; Woodfield 1982, p. v; McGinn 1989, pp. 9–11; Boghossian 1997, p. 163). Internalism and externalism are sometimes defined as opposing views about the determination of meaning or content, but talk about determination can be misleading because it might give the false impression that externalism is a claim about causation (see Kallestrup 2012, p. 60), while causal theories of reference only make up one portion of externalism (see e.g., Kripke 1972; Donnellan 1972). The internalism-externalism debate is therefore better framed more generally in terms of individuation. Where internalism claims that meaning or content can be solely individuated by features internal to the subject, externalism claims that sometimes meaning or content can be individuated not only by features internal but also external to the subject².

² I am aware that this is not the only way to define internalism and externalism. An anonymous referee noted that perhaps internalism should be rather defined as the view that content is *generally*

The classic arguments for such externalism come from Putnam (1975) who makes two important claims. First, Putnam claims that even if I was not able to distinguish between two kinds of trees, elms and beeches, the meanings of the terms “elm” and “beech” would differ in my usage of the terms insofar as there were people in my linguistic community who could draw the distinction (Putnam 1975, pp. 143–145). “Thus the way of recognizing possessed by these ‘expert’ speakers,” Putnam writes, is “possessed by the collective linguistic body, even though it is not possessed by each individual member of the body” (ibid., p. 145). This “way of recognizing” can determine the meaning of a word “while being unknown to almost all speakers who acquire the word” (ibid., p. 145). This idea is known as the division of linguistic labor.

Second, Putnam makes the Twin Earth argument. Putnam asks us to imagine a planet called Twin Earth, a nearly exact atom-by-atom copy of Earth, which also includes a twin for every inhabitant of Earth. Twin Earth is a “nearly exact” replica of Earth because there is one subtle yet crucial difference between the two seemingly identical planets. Where the molecules that make up the water that fills the lakes and falls from the sky on Earth are H₂O, the molecules that make up a qualitatively identical kind of liquid that fills the lakes and falls from the sky on Twin Earth are different. Call them XYZ. Although the same word “water” is used

individuated by internal features. However, since this definition would require me to elaborate in which cases (that are not “general”) internalism would maintain that content is not individuated solely by internal features, I have decided to maintain the stricter definition of internalism. This way of distinguishing externalism from internalism also exemplifies at least one commonly accepted view in the literature. For example, in his book *Semantic Externalism*, Kallestrup characterizes internalism as the view that “meaning is fully determined by features that are internal to the speaker” and externalism as the view that “meaning is determined at least in part by features that are external to the speaker” (Kallestrup 2012, p. 2). More precisely, externalism is defined as the view “that such content is in part determined or individuated by features external to the individuals who are in states with that content, i.e. that such content supervenes on the conjunction of internal features (intrinsic physical, experiential, psychological properties) and external features” (ibid., p. 62), whereas internalism is defined as “the negation of semantic externalism,” namely “the view that propositional content does supervene on internal features (intrinsic physical, experiential, psychological properties) – or that such content is fully determined by such internal features” (ibid., p. 63). For other examples of defining internalism more strictly than externalism, see e.g., Brown (2007, p. 17) and Wikforss (2008, p. 161). This kind of “uncompromisingly internalist conception” is supported for example by Farkas (2008, p. vii). It is of course possible to be, say, an internalist about natural kind terms while accepting externalism about indexical terms (see e.g., Segal 2000, p. 19), but if one tries to define internalism and externalism generally, I think this way of defining them is more helpful. Furthermore, this stricter way of understanding internalism (as the view that content or meaning is never individuated by anything external to the subject) is also emblematic of how internalism is understood when it is attributed to Husserl (e.g., McIntyre 1986, p. 102; Carman 2003, p. 83; see also Smith & McIntyre 1982, pp. 225–226).

on both planets to refer to the two kinds of liquid, the liquids are distinct: on Earth, it is water; on Twin Earth, it is twin water. Due to this difference, Putnam argues, the meaning of the term “water” is different when I use the word “water” on Earth and when my twin uses the word on Twin Earth. They are different even if neither my twin nor I had any knowledge of the molecules that differentiate (i.e., individuate) water from twin water. (Ibid., pp. 139–142.) In fact, Putnam goes as far as claiming that the meanings of “water” would differ even in the year 1750 when no one knew anything about molecules (ibid., p. 141).

Both of Putnam’s arguments reach the same conclusion that the meaning of a word is not only individuated by “the psychological state of the individual speaker” (ibid., p. 192). Yet, where the argument for the division of linguistic labor maintains that meaning can also be individuated by “the way of recognizing” possessed by other people, the Twin Earth argument yields the idea that meaning can also be individuated by hidden structures of the real world. “Traditional philosophy of language, like much traditional philosophy,” Putnam claims, “leaves out other people and the world” (ibid., p. 193), but externalism tries to encompass both.

While Putnam’s arguments concern the meaning of natural kind terms (e.g., “water,” “gold,” “elm” etc.), the externalist thesis has since then been extended to the meaning of non-natural terms (Burge 1979) as well as the content of mental states (see e.g., McGinn 1977; 1989; Burge 1979; 1982; Davidson 1987; McDowell 1992). The reasoning behind this extension from language to the content of mental states is based on the idea that the meaning of what the subject says (e.g., “water is wet”) is the content of the subject’s belief (e.g., that water is wet). If the content of what the subject says can change due to changes in the subject’s environment, then the content of the subject’s belief seems to be able to change as well. (See e.g., Segal 2000, p. 24; Farkas 2006, p. 328; Kallestrup 2012, pp. 70–71.) It is sometimes customary to reserve the term “meaning” for the meaning of linguistic expressions and the term “content” for the content of mental states in the literature, but in this article both terms are used interchangeably. Although the distinction between meaning and content is philosophically significant, it is not relevant for the purposes of this article, in which internalism and externalism are discussed broadly as opposing views about the individuation of content.

The central conceptual distinction employed in internalism and externalism is the division between internality and externality. Although Putnam does not use

the terms “internal” and “external” in his first papers on the topic, the wording of his famous slogan “‘meanings’ just ain’t in the head” (Putnam 1975, p. 144) exemplifies a division between what is in the head and what is not. Despite its wording, the slogan is not meant to imply that internalism, contra Putnam’s externalism, conceived meanings as entities located inside the head (see *ibid.*, p. 138); rather, the point is that where internalism maintains that meanings are only individuated by features that are located inside the head, externalism claims that meanings can also be individuated by features that are located outside the head. For example, reflecting on his slogan many years later, Putnam writes that “[m]eanings aren’t in the head’ does not mean that the brain has nothing to do with semantic competence. But what fixes the meanings of a speaker’s words is not just the state of her brain; the reference of our terms is generally fixed by two things that classical philosophy of language either ignores or mentions only as an afterthought: *other people and the world*” (Putnam 2013, p. 197). Although Putnam does not describe internalism as the view that meanings are located inside the head, Putnam thus describes internalism as the view that meaning is individuated, determined, or fixed by features that are located inside the head.

Contributions to the internalism-externalism debate after Putnam have often either taken up this division between what is inside the head and what is outside the head or elaborated it further into a slightly broader division between what is inside the subject’s body and what is outside the subject’s body. For example, Burge describes the twin thought experiments in which the subject’s “bodily motion, surface stimulations, and internal chemistry” remain the same but “the environment” is changed (Burge 1988, p. 650) or in which the twins’ “internal physical occurrences” (Burge 1979, p. 76) and “internal qualitative experiences” (*ibid.*, p. 79) remain the same while “the differences [in content] seem to stem from differences ‘outside’” (*ibid.*, p. 79). Furthermore, externalism has been described as “the view that facts external to a thinker’s skin are relevant to the individuation of (certain of) his mental content” (Boghossian 1997, p. 163), whereas internalism has been described as the view according to which “content is not relational, does not depend on anything outside the subject” (Segal 2000, p. 12). “Clearly enough,” McCulloch writes, “the ‘external’ means *external to the human individual*, so that anything beyond the skin is external in this sense” (McCulloch 1995, p. 189).

The sense in which the terms “internal” and “external” are used in these familiar descriptions of internalism and externalism seem to denote concepts

understood in a spatial sense. What is located inside the subject's body is internal, whereas what is located outside their body is external. This common-sensical idea can be formulated as a general criterion for internality and externality in the following way: *x* is internal if *x* is spatially inside the subject's body, whereas *x* is external if *x* is spatially outside the subject's body. I call this *the spatial criterion*. If internalism and externalism are defined according to the spatial criterion, we acquire the following definitions: internalism is the view that content can only be individuated by what is inside the subject's body, whereas externalism is the view that content can also be individuated by what is outside the subject's body. Among things outside the subject's body are typically features in the subject's social and physical environment, such as social practices, the beliefs of other people, as well as e.g., molecular structures of physical reality, whereas among features inside the subject's body are typically brain states and bodily functions.

Although the spatial criterion might remain the accepted notion in the internalism-externalism debate to this day, it has also been heavily criticized. One of the main criticisms against the spatial criterion stems from the rather simple observation that there is water inside the human body, which seems to undermine the argument based on the Twin Earth thought experiment. Putnam writes that the twin is "an identical copy" (Putnam 1975, p. 139) that is "molecule for molecule 'identical' with me (in the sense in which two neckties can be 'identical')" (ibid., p. 144), but if water consists of different molecules on Earth and Twin Earth, the twins cannot be "molecule for molecule" identical with one another because the water inside their bodies has different molecules (see e.g., Farkas 2003, pp. 188–193; 2008, p. 77; Wikforss 2008, pp. 161–162; Gertler 2012, p. 54ff.; Jacquette 2013, pp. 69–70; Gomes & Parrott 2021, p. 321). One might at first think of responding to this objection by saying that maybe Putnam's example of water was just a bad example, but if the Twin Earth thought experiment is supposed to drive home the externalist thesis about all natural kind terms, and "water" is a natural kind term, then the problem persists.

Farkas writes that the objection that there is water inside the human body has often not been taken very seriously because "the general feeling was that we could find a better example," but Farkas claims that "our willingness to overlook this problem in the original argument is better explained by the fact that the point of externalism is not whether the individuating facts are inside or outside the body" (Farkas 2003, p. 191; 2008, p. 77). To make this argument, Farkas devises her own Twin Earth thought experiment about meningitis. Suppose that while meningitis

is caused by the bacterium meningococcus on Earth, a similar disease called “meningitis” is caused by a different bacterium XYZ on Twin Earth. If I say “meningitis is dangerous” on Earth in 1750 and my twin utters the same sentence on Twin Earth in 1750, our sentences and thoughts seem to have different contents, according to externalism. (Farkas 2003, pp. 190–191, 2008, p. 76.) While this conclusion is “very similar in spirit to the externalist thesis, (...) the relevant individuating facts in this case are inside the body” (Farkas 2008, p. 76) because meningitis is a disease of the brain.

Following Farkas’ meningitis argument, some philosophers have gone as far as claiming that the internalism-externalism debate cannot be understood in an intelligible way at all (Gertler 2012), while others have kept a more ambivalent relation to precise definitions of the internal-external divide (see e.g., Kallestrup 2012, pp. 62–63, 230–231n10; Williamson 2006, p. 292; Raatikainen 2020, p. 60). On the other hand, some philosophers have, like Farkas (2008, pp. 80–99), sought alternative ways of drawing the distinction between internality and externality (see Gomes & Parrott 2021; Bruckner 2022; see also Wikforss 2008, pp. 161–162). The research concerning Husserl’s relation to the internalism-externalism debate exemplifies these different strategies: where some abandon both internalism and externalism as unintelligible, others seek alternative ways of understanding these positions. In the next section, I show how the former strategy, that is, the rejection of the internalism-externalism framework when it comes to understanding Husserl, evokes this problem concerning the internal-external divide.

3. Husserl on the Internal-External Divide

Although many Husserl scholars have rejected internalist ways of reading Husserl, some remain suspicious about the plausibility of providing an externalist interpretation as well. Zahavi claims that “[t]he very alternative between internalism and externalism remains bound to a Cartesian inner/outer division, but this division is precisely one with which phenomenology plays havoc” (Zahavi 2017, p. 118). Zahavi cites three sources from Husserl to support this claim (ibid., p. 119). First, in *Logische Untersuchungen*, Husserl refers to “[a]ntiquated talk of external and internal senses” which is “plainly stemming from the naïve metaphysic and anthropology of daily life” (Husserl 1984, pp. 673/281–282)³.

³ References to Husserl include both the German text and its English translation, if such a translation is available (the translated edition can be found after the German edition in the list of

Second, in an unpublished manuscript which Zahavi cites, Husserl writes that “[t]he transcendental ego is neither in the world nor outside it, and the world, in turn, is neither in it nor outside of it” (quoted in Zahavi 2017, p. 119). Third, in a text from 1921 published posthumously in the collection *Transzendentaler Idealismus*, Husserl writes that “[t]he transcendental I has no exterior; that is completely senseless” (Husserl 2003, p. 179). Therefore, Zahavi concludes:

Given the transcendental framework of Husserl's thinking, the internalism/externalism divide loses its relevance. Indeed, one might say that the main lesson of the reduction is the rejection of the dualism between a self-enclosed mind and a mindless world. Ultimately, we should appreciate that the phenomenological investigations of the structure of phenomenality are antecedent to any divide between psychical interiority and physical exteriority, since they are investigations of the very dimension in which any object – be it external or internal – manifests itself (Zahavi 2017, pp. 119–120).

I think Zahavi correctly uncovers a problem with applying the internalism-externalism framework to Husserl. Where internalism and externalism seem to be bound to the internal-external divide, Husserl rejects it. However, I think the sense in which the concepts of internality and externality are understood here is somewhat ambiguous. Zahavi talks about the “Cartesian inner/outer division” as well as the “divide between psychical interiority and physical exteriority,” which seems to designate the Cartesian divide. The divide is not defined, but it seems to express the equation of internality with the psychical or mental and the equation of externality with the physical. Although Zahavi of course rejects this Cartesian divide, the psychical-physical divide seems to be how he understands the internal-external divide. Yet, this way of drawing the line between internality and externality does not seem to work within the internalism-externalism debate for at least two reasons. First, some psychical features might be physical if one, for example, identified psychical states with neural states. This would mean that some internal features can be physical, that is, external. Second, Putnam's division of linguistic labor appeals to the possession of concepts by other people as an external

references). The first pagination refers to the German text, the second refers to the English translation. In case no second pagination is given, then there is no English translation available, and the translation is my own.

factor, which can individuate meaning, but the possession of concepts by other people might be nonphysical (e.g., that some experts could possess the concept of elm even if I did not). This would mean that some external features may be non-physical, that is, not external. Given these problems with drawing the line between internality and externality in psychical and physical terms, a more plausible way of understanding the internal-external divide along traditional lines is to see it as a spatial divide irrespective of the physical-psychical divide (see Gertler 2012, pp. 55–58), i.e., according to the spatial criterion defined in the previous section. In the end, I think this is what Zahavi has in mind, as he talks about how, for Husserl, it would be misleading “to conceive of consciousness as somehow *located inside* an interior sphere” (Zahavi 2017, p. 119, my italics).

In fact, the spatial criterion seems to exemplify how Husserl uses the internal-external divide in the passages from the textual material that Zahavi cites. Husserl’s distinction between external and internal perception in *Logische Untersuchungen* is a distinction between sensory experiences of spatially external stimuli and reflective experiences of one’s own experiences. Husserl might not consider experiences to be within the skin or the physical body of the subject, but this distinction is precisely one that he faults, as deriving from “naïve metaphysic.” The claim that “the transcendental I has no exterior” is different in the sense that it concerns the I as conceived of within the transcendental attitude. Briefly put, the transcendental attitude is the attitude in which the world and the objects of experience “become our focus in a new way” in contrast to the common way in which they are for us in the so-called natural attitude of ordinary experience (Husserl 1989, p. 173/494). The shift of attitude from the natural to the transcendental attitude “does not at all signify a turning away from the world or a transition into a theoretical area of speciality that is estranged from the world and of no interest,” but rather, “it is this turn that makes possible a really radical investigation into the world” (Husserl 1989, p. 178/497). “What we have renounced, then,” Husserl writes, “is only the naivete by which we allow the common experience of the world to be already given to us” (Husserl 1989, p. 173/494). The I grasped within this transcendental attitude, that is, the transcendental I, is not ontologically distinct from the I grasped within the natural attitude, the empirical I. “As transcendental ego, after all,” Husserl claims, “I am the same ego that in the worldly sphere is a human ego” (Husserl 1976b, pp. 267–268/264). The distinction between the transcendental I and the empirical I is rather, as Carr puts it, “a distinction between two ways of considering one and the same

self” (Carr 2003, p. 183). In light of this distinction, I would suggest reading Husserl’s statement that “the transcendental I has no exterior” as claiming that there is no sense in conceiving anything external to the transcendental I because the transcendental I is not spatial; it is, as Husserl says in the unpublished manuscript that Zahavi cites, “not in the world nor outside it.” In other words, given that the transcendental I does not designate a separate entity from the empirical I but a different way of considering the same self, the statement that the transcendental I has no exterior could be formulated as claiming that within the transcendental way of considering the self, there is no sense in conceiving anything external to the self because the transcendental way of considering the self is not concerned with spatial distinctions. That Husserl rejects the internal-external divide according to the spatial criterion is even more evident in *Ideen I* §49 where Husserl writes that “consciousness, considered in ‘purity,’ has to hold as (...) a context that has no spatiotemporal outside” (Husserl 1976a, p. 105/90). Therefore, the spatial internal-external divide seems to be inapplicable to the transcendental framework of Husserl’s phenomenology, making the internalism-externalism debate irrelevant for Husserl. If there is nothing spatially inside or outside consciousness, then it does not make sense to inquire whether content can be individuated by features that are spatially outside consciousness.

Since Husserl rejects the internal-external divide, understood according to the spatial criterion, whereas internalists and externalists employ the divide according to the spatial criterion, there lacks a common measure for understanding Husserl in relation to internalism and externalism. This suggests that Zahavi’s argument against internalist and externalist Husserl interpretations might be best understood as an argument about incommensurability. Where some internalist interpretations of Husserl seem to call attention to an incompatibility between externalism and Husserl’s phenomenology, stemming from a contradiction between disparate metaphysical commitments⁴, Zahavi’s argument seems to call attention to the difficulties of putting together disparate theoretical frameworks. However, since Husserl rejects the internal-external divide specifically according

⁴ For example, Dreyfus and Hall claim that “[s]ince the conditions of satisfaction of an intentional state are internal to the intentional state they define, phenomenological analysis need not concern itself with whether or not the object aimed at by the intentional state exists” (Dreyfus & Hall 1984, pp. 6–7). This aspect of Husserl’s account of intentionality is what Smith and McIntyre designate by calling intentional relations “existence-independent” (Smith & McIntyre 1982, p. 90). They too rule out externalist interpretations of Husserl by arguing that “the ‘transcendental’ foundation of Husserl’s phenomenology (...) is incompatible with letting the object of perception, or any other part of the external world, play a role in perceptual intention” (ibid., pp. 225–226).

to the spatial criterion, then Zahavi's argument should be framed in the following way: the internalism-externalism framework and the transcendental framework of Husserl's phenomenology are incommensurable because the internal-external divide, understood according to the spatial criterion, cannot be applied to the transcendental framework of Husserl's phenomenology.

The above analysis has revealed that while Husserl rejects the internal-external divide, he only rejects it in the specific sense of the spatial criterion. This suggests that it might be possible to render the internal-external divide applicable to the transcendental framework of Husserl's phenomenology if the internal-external divide is defined in an alternative way. If the concepts of internality and externality are understood in a different, non-spatial way, then the issue of incommensurability between the internalism-externalism framework and Husserl's transcendental phenomenology might be resolved, meaning that Husserl could be understood within the internalism-externalism debate. Existing externalist interpretations of Husserl have at least indirectly employed alternative definitions of internality and externality, but an alternative to the spatial criterion can also be discovered within the internalism-externalism debate itself. Before analyzing the budding definitions of the internal-external divide in the externalist interpretations of Husserl in Section 5, I will first look at this alternative definition from the internalism-externalism debate proposed by Farkas (2003; 2008).

4. Internality and Externality in Relation to the Subject's Point of View

As said in Section 2, Farkas' meningitis argument shows that the spatial criterion is an inadequate way of drawing the line between internality and externality in the internalism-externalism debate. Rejecting the spatial criterion, Farkas sets out to seek another criterion based on what the twins have in common in the Twin Earth thought experiment. Farkas suggests that the twins' common factor is the "subjective indistinguishability of their situations" (Farkas 2003, p. 196). My twin and I could not distinguish water and twin water from one another if we were to switch places. The features that individuate the meanings of our terms and the contents of our mental states about water and twin water are subjectively indistinguishable; they are "external to the subject's point of view" (Farkas 2008, p. 82).

Farkas explains the subjective indistinguishability of the twins' situations in terms of "the sameness of the phenomenal character of all their conscious thought and experience" (Farkas 2008, p. 91). What the twins share is "that things appear

(look, taste, smell, sound) the same for them; or the world is (and has always been) the same from their subjective viewpoint” (ibid., p. 82). Since things appear the same for the twins, the phenomenal properties of the twins’ intentional states are the same. For example, even if the meanings of the twins’ judgments about the stuff that they call “water” were different, as externalism claims, the phenomenal characters of their intentional states would nevertheless be identical, according to this way of understanding externalism.

Farkas’ proposal of first identifying the common factor between the twins as the subjective indistinguishability of their situations and then explaining it in terms of the sameness of their phenomenal properties can be captured in the following criterion for internality and externality: *x* is internal if *x* either appears or has appeared to the subject, whereas *x* is external if *x* does not appear and has not appeared to the subject. I call this *the phenomenal criterion*. The phenomenal criterion understands internality and externality in relation to the subject’s point of view. What appears and has appeared to the subject is within their point of view, while what does not appear and has not appeared to the subject falls outside their point of view. Farkas writes that “facts individuate mental content contents only insofar as they *make a difference* to the way things appear to us”, meaning “that any difference in the content of thoughts should be distinguishable from the subject’s point of view” (Farkas 2003, p. 203). If a change in an object’s properties is subjectively distinguishable, then those properties are apparent and thus internal to the subject’s point of view. If a change in an object’s properties is conversely subjectively indistinguishable, then those properties are not apparent and thus external to the subject’s point of view (Farkas 2003, pp. 196–198). Based on the phenomenal criterion, “the internalism/externalism controversy should be formulated with reference to the relation of things appearing the same, or situations being indistinguishable” (Farkas 2008, p. 84). Thus, internalism becomes the view that meaning or content is individuated by features that appear or have appeared to the subject, whereas externalism becomes the view that meaning or content can be individuated not only by features that appear or have appeared to the subject but also features that do not appear and have not appeared to the subject.

If internalism and externalism are understood according to the phenomenal criterion, the issue of incommensurability between the transcendental framework of Husserl’s phenomenology and the internalism-externalism framework is dissolved because the phenomenal criterion removes the internal-external divide

from the spatial framework, in which Husserl rejects the divide, and reconfigures the internal-external divide in relation to the subject's point of view. It thus renders the internal-external divide applicable to the transcendental framework of Husserl's phenomenology and hence makes it possible to understand Husserl in relation to the internalism-externalism debate. Given that Husserl's transcendental phenomenology investigates the structure of phenomenality from the first-person perspective, the conclusion seems almost self-evident. The phenomenal criterion brings the philosophical question at stake in the internalism-externalism debate back to the subject's perspective, as "externality" is understood in the sense of "external to the subject's point of view" (Farkas 2008, p. 82). Even though Husserl rejects the idea of something spatially external to consciousness, he still must deal with the distinction between what is subjectively distinguishable and what is not or between what appears to the subject and what does not.

5. Phenomenological Ways to Understand the Internal-External Divide

The phenomenal criterion thus makes it possible to understand Husserl in relation to the internalism-externalism debate. However, the understanding of Husserl in relation to the internalism-externalism debate remains difficult even with this new criterion. It seems that Husserl, to borrow Zahavi's phrase, also "plays havoc" with the internal-external divide according to the phenomenal criterion as Husserl seems to reject the idea of something completely external to the subject's point of view. For example, consider the following passage from Husserl's *Cartesianische Meditationen*:

Every conceivable sense, every conceivable being, whether the latter is called immanent or transcendent, falls within the domain of transcendental subjectivity, as the subjectivity that constitutes sense and being. The attempt to conceive the universe of true being as something lying outside the universe of possible consciousness, possible knowledge, possible evidence, the two being related to one another merely externally by a rigid law, is nonsensical. (...) If transcendental subjectivity is the universe of possible sense, then an outside is precisely – nonsense. (Husserl 1973, p. 117/84, translation modified)

There are two important claims in this passage. First, Husserl claims that "[e]very conceivable [*erdenkliche*] sense, every conceivable being" is "within the domain

[*Bereich*] of transcendental subjectivity” or “the universe of possible sense” [*das Universum möglichen Sinnes*]. Second, Husserl claims that there is no sense in conceiving anything “outside” [*außerhalb*] or external to the domain of transcendental subjectivity or the universe of possible sense. In a word, Husserl claims that everything conceivable is within the domain of transcendental subjectivity and hence there is nothing conceivable outside that domain. If everything conceivable falls within the domain of transcendental subjectivity or the universe of possible sense, then there is no sense in conceiving anything outside such a domain. In the rest of this article, I use the term “transcendental domain” to designate this all-encompassing dimension that Husserl describes.

The passage from *Cartesianische Meditationen* quoted above is somewhat similar to Husserl’s assertion that “the transcendental I has no exterior” because in the passage Husserl seems to reject the notion of externality in the spatial sense; the transcendental domain is not spatial. However, this spatial reading does not hold up for the passage in its entirety. Even if the term “outside” in the claim that there is no sense in conceiving anything outside the transcendental domain could be understood according to the spatial criterion (i.e., that Husserl rejected the idea of something spatially external to the transcendental domain), the “within” in the claim that everything conceivable is within the transcendental domain could not be understood according to the spatial criterion (i.e., that Husserl would suggest that everything conceivable is spatially inside the transcendental domain). As said, the transcendental domain is not spatial. However, it also does not seem to make sense to understand the “within” in this claim according to the phenomenal criterion. The phenomenal criterion would make the claim that everything conceivable is within the transcendental domain into the claim that everything conceivable appears or has appeared to transcendental subjectivity, but clearly everything conceivable does not and has not appeared to transcendental subjectivity. The phenomenal criterion would also render the claim that there is no sense in conceiving anything outside the transcendental domain into the absurd claim that everything conceivable, which has not yet appeared to the subject, is beyond the universe of possible sense (i.e., “nonsensical”).

Given these complexities, the passage from *Cartesianische Meditationen* seems to engender a challenge for understanding Husserl in relation to the internalism-externalism debate. Although the phenomenal criterion makes it possible to understand Husserl as an internalist or an externalist, the internal-external divide should be drawn within the transcendental domain, meaning that

even features that are external should be within this transcendental domain. The aim should be to redefine the notions of internality and externality in this way without rendering any forms of either internalist or externalist theories of content incoherent. That is to say, a successful criterion for the internal-external divide in this context should, on the one hand, redefine the notions of internality and externality so that both internal and external things are within the transcendental domain and, on the other, this redefinition should not render theories of internalism and externalism incoherent.

In this section, I show how two externalist interpretations of Husserl, by Smith (2008) and Crowell (2008), introduce phenomenological criteria for the internal-external divide. They redefine “internality” and “externality” so that both internal and external things remain within the transcendental domain. It is important to emphasize, however, that Smith and Crowell are not primarily interested in proposing alternative criteria for the internal-external divide; rather, they are interested in presenting plausible interpretations of Husserl. Nevertheless, in interpreting Husserl as a kind of externalist, both interpretations end up proposing useful ways of thinking about the internal-external divide in a phenomenological context.

To reconcile externalism with Husserl, Smith (2008) proposes to “construe externalism as dealing with what is external to any *individual* consciousness” (Smith 2008, p. 317). This in itself might not sound that different from the spatial criterion, but Smith elaborates by stating that even if “the physical world may ultimately ‘reduce’ to consciousness, your consciousness cannot be reduced to mine, or mine to yours” and “[e]ach individual consciousness, therefore, is ‘external’ to every other” (ibid.). Furthermore, Smith remarks that “everything, the constitution of which involves other consciousnesses, will also count as external to any given consciousness” (ibid.). “In particular,” Smith writes, “real, objective elements of the physical world count as external to any individual consciousness, because they have an essentially inter-subjective constitution” (ibid.).

The key concept in Smith’s proposal is the constitution which “involves other consciousnesses,” i.e., intersubjective constitution. There are many ways to understand Husserl’s concept of constitution, but one somewhat common way to understand it is as a process in which something is enabled to manifest as something (see e.g., Zahavi 2003, p. 73; Ströker 1993, pp. 104–107). If something is constituted by consciousness, it is enabled to manifest as something to

consciousness. For example, material things around us are constituted as real. As intersubjective, this process involves other consciousnesses. Husserl writes that “everything else that is also called transcendent, such as the objective world, rests upon the transcendence of foreign subjectivity” (Husserl 1959, p. 495). The idea is that things are enabled to manifest as real due to the subject’s “experience of the transcendence (and inaccessibility) of foreign subjectivity” (Zahavi 2003, p. 115). Since the subject encounters other people who also have their own experiences of the world and the things around them, the subject experiences the world and the things around them as exceeding their subjective point of view. The constitutive significance of other consciousnesses stems from the idea that “objects cannot be reduced to being merely my intentional correlates if they can be experienced by Others as well” (Zahavi 2003, p. 115). The table that I see before me is external to my perspective because it is not a private object; others can see it too. Given this “intersubjective experienceability of the object” (ibid., p. 115) and the inaccessibility of others’ experiences of the object for me, the object is constituted by intersubjective consciousness. Thus other consciousnesses are also a part of the transcendental domain.

Smith understands the internal-external divide in relation to constitution. Smith does not explicitly define how he uses the terms “internal” and “external,” but he says that in addition to “other consciousnesses,” all things “the constitution of which involves other consciousnesses” are external; they are external “because they have an essentially inter-subjective constitution” (Smith 2008, p. 317). This seems to yield a starting point for redefining the internal-external divide in terms of intersubjective constitution. Based on Smith’s proposal, one can derive the following criterion for internality and externality: x is internal if the constitution of x does not involve other consciousnesses, whereas x is external if the constitution of x does involve other consciousnesses. Since the constitution of tables involves other consciousnesses, for example, they are external. I call this *the constitution criterion*. According to the constitution criterion, we acquire the following definitions of internalism and externalism: internalism becomes the view that content or meaning is solely individuated by features whose constitution does not involve other consciousnesses, whereas externalism becomes the view that the individuation of content or meaning may sometimes require features whose constitution involves other consciousnesses.

The constitution criterion differs from both the spatial and the phenomenal criterion. For example, the constitution of my body as a physical body with neural

and biological processes involves other consciousnesses: there are aspects of my body that I cannot experience but others can. The physical features inside my body, for example, are therefore external according to the constitution criterion. Although the constitution criterion seems to exemplify a similar attempt to reconfigure the internal-external divide in relation to the subject's point of view as the phenomenal criterion, the constitution criterion also differs from the phenomenal criterion. Some features of physical things spatially outside my body are internal according to the phenomenal criterion, insofar as any changes in those features would be subjectively distinguishable (e.g., the shape of the table in front of me), but those features would count as external according to the constitution criterion because their constitution involves other consciousnesses.

The main difference between the constitution criterion and the spatial and the phenomenal criteria is that the constitution criterion defines externality so that external things are within the transcendental domain. Although things such as the table in front of me are external according to the constitution criterion, Smith's emphasis that they are external only to an "individual consciousness" (Smith 2008, p. 317) seems to imply that they are nevertheless internal to what might be called intersubjective consciousness.

The constitution criterion inevitably brings forth questions about the conditions for something to be intersubjectively constituted (or "external," according to the constitution criterion). The way of framing the internal-external divide according to the constitution criterion works well enough for Smith's interpretation that Husserl is a disjunctivist, but the constitution criterion loses some of its strength when it is more broadly employed. To Smith, Husserl is a disjunctivist in the sense that veridical perceptions and hallucinations are different kinds of experiences because their horizons are different. The horizon is something that all experiences have (Husserl 1973, p. 82/44), and it designates how experiences extend beyond what is immediately given. "For example," Husserl writes in *Cartesianische Meditationen*, "there belongs to every external perception its reference from the 'genuinely perceived' sides of the object of perception to the sides 'also meant' – not yet perceived, but only anticipated and, at first, with a non-intuitional emptiness" (Husserl 1973, p. 82/44). For example, my perceptual experience of a house extends beyond the side of the house before me to encompass other sides of the house currently unseen. Smith claims that veridical perceptions and hallucinations have different horizons because hallucinations have disharmonious horizons, which include an experience that eventually discloses the

object of the hallucinatory experience as unreal, whereas veridical perceptions have harmonious horizons, which contain no such experiences that would reveal the object of experience as unreal (Smith 2008, pp. 329–330). Since the reality of the object is, at least partly, intersubjectively constituted to Husserl (1959, pp. 495n2/619n2), something “external” (i.e., the reality of the object), according to the constitution criterion, individuates veridical perceptions from hallucinations, making Husserl an externalist in this sense (of disjunctivism). However, when it comes to twin cases, where two veridical experiences seemingly intend the same object but actually intend different objects (e.g., distinct kinds of liquids), the reality of the objects cannot individuate the experiences because the objects of both experiences are real. If confronted with twin cases, the question for the constitution criterion is whether something like molecules could be intersubjectively constituted, even if the subject was not aware of molecules. This question cannot be answered without defining the conditions for something to be intersubjectively constituted more precisely.

Another phenomenological way of understanding the internal-external divide can be derived from Crowell’s (2008) externalist interpretation of Husserl, which conceives the internal-external divide with Husserl’s concepts of intentional immanence and transcendence. Husserl uses these concepts in various ways (see e.g., Husserl 1976a, pp. 78/66, 87–94/74–80, 96–98/82–83), but here I am concerned with how Husserl uses them in the text *Die Idee der Phänomenologie* where he introduces them for the first time. In *Die Idee der Phänomenologie*, Husserl introduces first a basic distinction between real or genuine immanence and transcendence, which is then complicated by a further distinction between intentional immanence and transcendence (also called “phenomenological” immanence and transcendence). The former seems to coincide with the spatial internal-external divide, as Husserl for example writes “[t]he immanent is in me (...) and the transcendent is outside of me” (Husserl 1950, pp. 5/62–63), but the latter distinction concerns the way something is given. Intentional immanence is “absolute and clear givenness, self-givenness in the absolute sense,” whereas intentional transcendence is givenness in which “we go beyond what is *given in the genuine sense*, beyond *what can be directly seen and apprehended*” (ibid., pp. 35/27–28). Although the profile of a house in front of me is transcendent in the real or genuine sense, it is immanent in this intentional sense because it appears to me absolutely and clearly. The further profiles of the house, currently unseen by me, are transcendent in the intentional sense. In his interpretation, Crowell claims

that “[intentional] immanence does not invoke an ontological distinction between regions – between a *forum internum* and an external world – but a normative relation between presence and absence” (Crowell 2008, p. 346). The gist of Crowell’s interpretation is that the normative links between the subject’s intentional implications (e.g., expectations about further, currently unseen aspects of the object) and the object determine content. However, the details of Crowell’s interpretation are not important in this article as I am mainly interested in the incipient idea in the interpretation that the internal-external divide could be understood with the distinction between intentional immanence and transcendence.

Based on the intentional distinction between immanence and transcendence, one can derive the following criterion for internality and externality: *x* is internal if *x* is given in intentional immanence, whereas *x* is external if *x* is given in intentional transcendence. I call this *the givenness criterion*. According to the givenness criterion, the two kinds of liquid that my twin and I perceive are “internal” in the sense of intentional immanence (they are present, they appear directly and clearly), but the two kinds of liquid have further unknown properties that are not self-given and thus “external” in the sense of intentional transcendence (they are absent, they do not appear directly and clearly). According to the givenness criterion, we acquire the following definitions of internalism and externalism: internalism becomes the view that content or meaning can only be individuated by features that are intentionally immanent, whereas externalism becomes the view that content or meaning can also be individuated by features that are intentionally transcendent.

Although Farkas’ phenomenal criterion might seem to fit Husserl’s phenomenology perfectly at first glance, there is a simplicity to the criterion that does not capture the more complex structure of phenomenality in Husserl’s transcendental phenomenology. The phenomenal criterion understands the internal-external divide in terms of what appears to the subject and what does not, but this distinction seems too simplistic within Husserl’s phenomenology whose concept of appearance includes an interplay between presence and absence, i.e., what appears directly in intentional immanence and what co-appears indirectly in intentional transcendence. The givenness criterion takes this richer concept of appearance into account by understanding the internal-external divide as a distinction between two modes of givenness, intentional immanence and transcendence.

Like the constitution criterion, the givenness criterion redefines “externality” so that external things are within the transcendental domain. This is evident, for example, in Crowell’s claim that in the year 1750 “neither H₂O nor XYZ are even phenomenologically [i.e., intentionally] transcendent” because they do not belong to the “evidential horizon of phenomenological [i.e., intentional] immanence” (Crowell 2008, p. 346). Yet, in the course of time, they will become intentionally, or phenomenologically, transcendent. “[B]ecause the water I perceive is not an inert datum but an element within a normative space of intentional implications (teleological interconnections),” Crowell writes, “it adumbrates avenues along which further aspects can become given – if/then structures that indicate a course of possible empirical investigation that could eventually disclose something like molecular structure” (ibid., p. 346). The intentional implications (i.e., expectations about the object) available to my twin and I in 1750 “will be the same,” but “[w]hen molecular theory develops, together with the relevant technology and practices, more of those aspects will be available,” and then our intentional implications will be different (ibid., pp. 346–347). These statements make clear that, in contrast to the spatial criterion and the phenomenal criterion, the givenness criterion redefines “externality” so that external things are retained within the transcendental domain. Unlike the constitution criterion, the givenness criterion seems to provide at least a provisional characterization of the conditions for something to be intentionally transcendent: it needs to be available. Given Crowell’s references to scientific, technological, and practical advancements, this notion of availability seems to be understood in the sense that such advancements enable the intentional transcendence of e.g., molecular structures.

The givenness criterion seems to run into a problem similar to the constitution criterion. Construed according to the givenness criterion, the internalism-externalism debate would become a dispute about the question whether content or meaning can be determined by intentionally transcendent factors. However, this might not be that helpful when it comes to framing the debate. Husserl has many remarks about intentional implications being co-meant. For example, Husserl talks about “sides which were co-meant in a more or less indeterminate way” (1952, p. 35/38), “sides ‘also meant’” (1973, p. 82/44), and how “non-visible sides are certainly also there somehow for consciousness, ‘co-meant’ as co-present” (1966, p. 4/40). Since these unknown intentionally transcendent aspects are co-meant in experience, they seem to determine meaning

or content. In fact, Husserl writes about how intentional implications make experience go “beyond what is at any given time grasped in singularity and relative determinateness and accepted as such” and how, therefore, intentional implications engender “a transcendence of sense” that “clings to every particular apperception, to every complex of particular apperceptions” (Husserl 1939, p. 30/34). If sense or meaning becomes partly transcendent by means of the intentional implications co-meant, then meaning is individuated by external factors, understood according to the givenness criterion. Given Husserl’s remarks about intentional implications being co-meant and them engendering a transcendence to sense, it would seem self-evident that, according to the givenness criterion, Husserl is an externalist.

Yet, this way of framing the debate gives a false sense of simplicity because even if meaning was transcendent, it remains debatable whether “the transcendence of sense” designates just a general openness of meaning rather than a more specific relation of individuation dependence that externalists have in mind. For this general openness of meaning might also be just as akin to the internalist idea, sometimes called the common concept strategy, according to which the twins in the Twin Earth thought experiment share the same concept of water and therefore the meaning of the term “water” can mean the same thing on Earth and Twin Earth; that is, the meaning of the term includes both H₂O and XYZ (see e.g., Zemach 1976, p. 118; Mellor 1977, p. 303; Segal 2000, p. 19; Jackson 2003, p. 61). Even if one employed the givenness criterion, the seemingly externalist statement that meaning is individuated by intentionally transcendent features thus remains open to both internalist and externalist readings. It remains open whether those intentionally transcendent features that individuate meaning are in turn determined by the expectations and hypothetical beliefs of the subject’s cognitive community (where there might be uncertainty whether the molecular structure of water is H₂O or something else) or real-world structures despite how things seem to the community. This remains ambiguous due to a similar shortcoming as in the constitution criterion: the conditions for something to be intentionally transcendent are unclear.

As said, Crowell’s reference to availability can be understood as an attempt at clarifying those conditions. Due to scientific and technological advancements, for example, molecules can become available and thus intentionally transcendent. However, consider a situation where such scientific and technological advancements have been made that the molecular structure of water can be

discovered, but people have not yet made the discovery. The question is whether H₂O would then be intentionally transcendent. This is basically the internalism-externalism question just about intentionally transcendent features. Rather than asking whether the term “water” means different things on Earth and Twin Earth, one ends up inquiring what determines intentional transcendence. Neither of these questions can be answered unless the conditions for intentional transcendence are more specifically defined.

Furthermore, even if these questions about the conditions for something to be intentionally transcendent were adequately answered, Crowell’s proposal seems to be committed to the idea that the meaning of the term “water” changes when molecular theory and scientific technology develop. This is of course one possible conception, but if this conception is written into the very criterion for the internal-external divide, then some basic forms of externalism are precluded at the outset by the criterion. For one, Putnam’s semantic externalism maintains that the meaning of scientific terms such as “atom” and “kinetic energy” retain the same meaning despite historically changing definitions (see Putnam 2013, p. 194). The reason is that the meaning of these terms is externally individuated by the stuff that is there in the real world.

6. Reconfiguring the Internal-External Divide

The main issue in these two phenomenological ways of understanding the internal-external divide is that in redefining the internal-external divide so that external things are within the transcendental domain they end up precluding some forms of externalism. Namely, the givenness criterion seems to rule out Putnam’s externalism, which is a problem if one wants to use the criterion as a general way of construing the internalism-externalism debate. The same problem seems to emerge for the constitution criterion because if externality is associated with intersubjectivity, then forms of externalism that refer to external things beyond the community of subjects seem to be precluded. Yet, whether this problem emerges for the constitution criterion is somewhat difficult to decide without defining the conditions for something to be intersubjectively constituted. Regardless of these issues, I think both the constitution criterion and the givenness criterion provide useful concepts for improving the phenomenal criterion so that the phenomenal criterion could be adjusted to take the transcendental domain into account.

When it comes to the constitution criterion, the main takeaway is the role of intersubjectivity. By combining the constitution criterion with the phenomenal

criterion, one can derive the notion of intersubjective distinguishability rather than constitution that involves other consciousnesses. Something that Farkas does not discuss, but which seems important for defining externalism, is that some changes might be subjectively indistinguishable yet intersubjectively distinguishable. Some forms of externalism appeal to subjectively indistinguishable changes in the real world that are also intersubjectively indistinguishable. For example, in the year 1750, nobody could distinguish water and twin water. In turn, other forms of externalism might appeal only to subjectively indistinguishable changes in the real world which are nevertheless intersubjectively distinguishable. For example, there are experts in botany who can tell elms and beeches apart. Based on the phenomenal criterion, the question between internalism and externalism becomes the question whether content can only be individuated by features whose changes are subjectively distinguishable or also features whose changes are subjectively indistinguishable. If one opts for the latter, that is, externalist, position, then the following question is whether content can be individuated by features whose changes are subjectively indistinguishable but still intersubjectively distinguishable or whether content can also be individuated by features whose changes are not even intersubjectively distinguishable (e.g., the individuating features between water and twin water in the year 1750).

Where the constitution criterion can be used to complement the phenomenal criterion with the notion of intersubjective distinguishability, the givenness criterion can be used to improve the phenomenal criterion by introducing a more nuanced conception of appearance. Rather than simply operating with a distinction between what appears and what does not, the givenness criterion introduces a phenomenologically richer distinction between what appears clearly and directly (in intentional immanence) and what co-appears unclearly and indirectly (in intentional transcendence). With this more nuanced conception of appearance, I think the givenness criterion can be used to solve a problem in the phenomenal criterion. The solution to this problem is also helpful in spelling out a better criterion that eludes many of these problems.

One of the criticisms against the phenomenal criterion is that it seems to make a view known as phenomenal externalism incoherent (see Gertler 2012, p. 61; Gomes & Parrott 2021, pp. 323–324). This is a problem insofar as a successful criterion should provide definitional criteria for “internality” and “externality” that can be employed in plausible definitions of internalism and externalism without precluding any forms of them. Phenomenal externalism is a view that extends the

externalist thesis about the content of judgments and beliefs to the character or qualia of perceptual experiences (Dretske 1996; Lycan 2001). Although judgments, beliefs, and sensory experiences might be (spatially) internal to the subject, phenomenal externalism maintains that the facts that can individuate them are (spatially) external to the subject. These individuating facts are phenomenal characters or qualia. According to phenomenal externalism, not only are beliefs and judgments about water and twin water different judgments and beliefs (because they have different contents) but also sensory experiences about water and twin water are different sensory experiences (because they have different phenomenal characters or qualia). This means that, unlike it is conventionally presumed, water and twin water do not even appear the same to the twins.

Dretske has an example that is useful in illustrating phenomenal externalism (Dretske 1996, pp. 153–154). Suppose my friend and I are going to listen to a symphony. I am musically ignorant, but my friend is more familiar with music. When my friend hears a change of key in the symphony, he turns to me and asks whether I heard it. Given my lack of musical expertise, I would not be able to distinguish between series of sounds with changes of key and without them. Regardless of whether there was a change of key or not, I would not notice it. In one sense of “appearance” (which Dretske calls “doxastic”), the change of key, if there was one, does not therefore appear to me. In another sense of “appearance” (which Dretske calls “phenomenal”), however, the change of key seems to appear to me. Yet, even if it did appear to me, Dretske claims, I could not be made aware of this fact because I lack the concept for change of key (*ibid.*, p. 154).

Although phenomenal externalism seems to be a coherent view when understood according to the spatial criterion, phenomenal externalism seems to become incoherent if it is formulated according to the phenomenal criterion. If externality is understood according to the phenomenal criterion, then phenomenal characters or qualia could not be external, unlike phenomenal externalism claims; they are, by definition, phenomenal, that is, apparent to the subject and therefore internal. According to the phenomenal criterion, internal factors are those that “make a difference” to the way things appear (Farkas 2003, p. 203). Phenomenal characters seem to be internal in this sense as they make a difference to the way things appear. If there was a change in the phenomenal characters of the subject’s intentional states, the subject would notice it and be able to distinguish between the intentional states. Phenomenal externalism, construed according to the phenomenal criterion, becomes incoherent because phenomenal characters or

qualia – the ways things appear – cannot be external according to the criterion (i.e., features that do not appear or have not appeared to the subject).

One solution to this problem is to refrain from giving the subjective indistinguishability of the twins' situation the explanation that Farkas does; that is, the explanation that the twins have the same phenomenal properties. If one rather just says that the twins are similar in the sense that their situations are subjectively indistinguishable, one can avoid this problem (see Gomes & Parrott 2021, p. 325). The problem stems from claiming that the twins are phenomenally identical, which is something phenomenal externalism denies. The fact that the problem can be solved this easily might imply that it is not a very serious problem in the first place. In Dretske's example, the symphony with the change of key is for me subjectively indistinguishable from the symphony without the change of key. It is external to the subject's point of view. In this sense, phenomenal externalism seems coherent when formulated with the phenomenal criterion; phenomenal externalism maintains that sensory experiences can be individuated by something external and the term "external" can be understood here in the general sense that Farkas proposes (that changes in *x* would be subjectively indistinguishable). The problem only emerges when one adds the explanation that equates subjective indistinguishability with phenomenal identity.

Therefore, the natural solution is to either refrain from making this further explanation, as Gomes and Parrott (2021) do, or modify the explanation. I think the givenness criterion provides an avenue for modifying the explanation so that phenomenal externalism remains coherent. If internality is understood as intentional immanence and externality as intentional transcendence, then phenomenal externalism becomes the view that sensory experiences can be individuated not only by intentionally immanent but also intentionally transcendent characters. Phenomenal characters can be (partially) something that do not appear clearly and directly but co-appear unclearly and indirectly. In the sense of intentional immanence, the change of key does not appear to me. It is not given clearly and directly to me. The change between a change of key and its absence would be subjectively indistinguishable to me. On the other hand, in the sense of intentional transcendence, the change of key does co-appear to me. It is given indirectly and unclearly. Even though I may not be aware of it, the change of key is there, and I could notice it. As a result, phenomenal externalism remains a coherent view when construed according to the givenness criterion: the

phenomenal characters of sensory experiences can be individuated by features that do not appear in one sense but appear in another sense (that is, they co-appear).

Although the givenness criterion solves an issue in the phenomenal criterion, it has its own problems. As said, the givenness criterion, like the constitution criterion, is more restricting than the phenomenal criterion. While the givenness criterion can allow phenomenal externalism, it seems to preclude other forms of externalism. Namely, if molecular structures are not even intentionally transcendent in the year 1750, as Crowell claims, then Putnam's externalist claim that the meaning of "water" is different on Earth and Twin Earth in the year 1750 is ruled out. In this sense, Putnam's externalist claim would be appealing to features that are not even external; they are beyond the internal-external divide as it is understood according to the givenness criterion.

However, I think this problem can be solved by looking at the solution to the issue of phenomenal externalism and the phenomenal criterion above. One crucial aspect in regard to Dretske's symphony example is that the change of key *could* appear to me in intentional immanence. It could appear to me directly and clearly if conditions were altered. It could appear to me thus if, for example, I were to study music so that I could distinguish series of sounds with and without changes of keys. In fact, one way to understand intentionally transcendent features is to understand them as potentially apparent. They are of course "apparent" in the sense that they co-appear with what appears in intentional immanence, but as such they are also not "apparent" in the sense of intentional immanence. Nevertheless, they *may* appear in intentional immanence if conditions were altered.

In order for the concept of potential appearance not to run into the same problem of ambiguity as intentional transcendence, however, its conditions must be clear. I think potential appearance is best understood as not just determined by my capabilities and the capabilities of others but also how things are in the real world. In order for the change of key to be even potentially apparent to me, I of course need to have the capabilities to discover it, the possibilities for acquiring such capabilities, or at least someone else needs to have such capabilities or possibilities of acquiring such capabilities, but the symphony also needs to have a change of key so that it can become apparent. Viewed this way, it seems plausible to understand Husserl's claim in *Cartesianische Meditationen* that there is nothing external to the transcendental domain as the claim that it is nonsensical to postulate

entities that are not even potentially apparent⁵. In fact, in the passage from *Cartesianische Meditationen*, Husserl talks about how there is no sense in postulating anything “outside the universe of possible consciousness, possible knowledge, possible evidence” (Husserl 1973, p. 117/84). The insight taken from this quote is that perhaps “externality” should be redefined so that external things are nevertheless within the transcendental domain as they are potentially apparent.

I am using the notion of potential appearance in the broad sense that it also encompasses indirect observations such as seeing entities under a microscope or seeing evidence of subatomic particles in empirical measurements. All these are instances where something “appears” in one sense of the term. Even though people in the year 1750 might have lacked the capabilities (that is, the scientific instruments and the know-how to employ them) to distinguish between water and twin water, water still was water (rather than twin water) and therefore its nature as water (its molecular structure) was potentially apparent. It was possible for people, in the long run, to acquire the capabilities needed for water as H₂O to become apparent to them.

Combining these two takeaways from the givenness criterion (potential appearance) and the constitution criterion (intersubjective distinguishability), the phenomenal criterion can be modified in the following way: x is internal if x either appears or has appeared to the subject, whereas x is external if x does not appear and has not appeared to the subject but is potentially apparent to a subject. I call this *the modified phenomenal criterion*. Based on the modified phenomenal criterion, we acquire the following definitions of internalism and externalism. Internalism becomes the view that meaning or content can only be individuated by features that either appear or have appeared to the subject, whereas externalism becomes the view that meaning or content can be individuated not only by features that either appear or have appeared to the subject but also features that might appear to the subject or someone else.

The modified phenomenal criterion is very similar to the phenomenal criterion, but it adds the phenomenological condition on externality: in order for an external feature to be individuating, it must be potentially apparent to a subject,

⁵ This seems to be how Ameriks, for example, reads the passage about “the universe of possible sense”. He says that to Husserl it is “‘absurd’ though not necessarily logically contradictory to posit beings which are in principle uncognizable by any mind” (Ameriks 1977, p. 510). Although Ameriks is initially talking about another passage from Husserl, he then says that “[p]recisely the same point is expressed in the claim (...) that it is ‘nonsense’ to speak of something outside ‘transcendental subjectivity.’” (ibid.)

i.e., either the subject or someone else at some point in time. Although intersubjectivity plays a role in the criterion, the modified phenomenal criterion does not require that changes in external features need to be actually intersubjectively distinguishable so that the features qualify as external. Changes can be intersubjectively indistinguishable as long as they are potentially intersubjectively distinguishable.

The modified phenomenal criterion fulfils both requirements set for a successful criterion: it redefines “externality” within the transcendental domain without rendering forms of internalism and externalism incoherent. Unlike the constitution and the givenness criterion, the modified phenomenal criterion enables Putnam’s externalist claim: in the year 1750, the difference between water and twin water would not be subjectively nor intersubjectively distinguishable but it would be potentially subjectively and intersubjectively distinguishable. The features of water that individuate it from twin water would be potentially apparent even in the year 1750. Unlike the phenomenal criterion, the modified phenomenal criterion enables phenomenal externalism. The presence or absence of the change of key in the symphony is potentially apparent to the subject as are the individuating features of water. Construed according to the modified phenomenal criterion, phenomenal externalism becomes the view that phenomenal characters or qualia can be features that do not appear to the subject insofar as they are potentially apparent.

7. Conclusion

Some have argued that Husserl cannot be understood within the internalism-externalism framework. In this article, I have identified the motivation for this argument in a spatial way of understanding the internal-external divide, which is pertinent to internalism and externalism but rejected by Husserl. I have argued that the internal-external divide can be rendered applicable to the transcendental framework of Husserl’s phenomenology if the internal-external divide is understood alternatively in relation to the subject’s point of view. In addition to Farkas’ attempt to reconfigure the internal-external divide in phenomenal terms (i.e., the phenomenal criterion), I have investigated two phenomenological ways of understanding the internal-external divide: the constitution and the givenness criterion. These phenomenological ways of understanding differ from the phenomenal criterion mainly by redefining the internal-external divide within the so-called transcendental domain. The problem with these two phenomenological

criteria is that, on the one hand, they leave the conditions for something to be “external” (either in the sense of intersubjective constitution or intentional transcendence) somewhat ambiguous and, on the other, they preclude some forms of externalism, but I have claimed that they can nevertheless be used to improve the phenomenal criterion. Where the constitution criterion can be used to derive the notion of intersubjective distinguishability, the givenness criterion can be used to introduce a more nuanced conception of appearance for the purposes of modifying the phenomenal criterion. The upshot is a modified phenomenal criterion for the internal-external divide, which both redefines the internal-external divide within the transcendental domain and enables forms of externalism that the other criteria preclude.

University of Helsinki

References

- Alweiss, L. (2009). Between Internalism and Externalism: Husserl’s Account of Intentionality. *Inquiry*, 52, 53–78. <https://doi.org/10.1080/00201740802661494>
- Ameriks, K. (1977). Husserl’s Realism. *The Philosophical Review*, 86(4), 498–519
- Beyer, C. (2000). *Intentionalität und Referenz: Eine sprachanalytische Studie zu Husserls transzendentaler Phänomenologie*. Mentis.
- Beyer, C. (2013). Noema and Reference. In M. Frauchiger (Ed.), *Reference, Rationality, and Phenomenology: Themes from Føllesdal* (pp. 73–88). De Gruyter.
- Boghossian, P. (1997). What the Externalist Can Know A Priori. *Proceedings of the Aristotelian Society*, 97, 161–175. <https://doi.org/10.2307/1522971>
- Brown, J. (2007). Externalism in Mind and Epistemology. In S. C. Goldberg (Ed.), *Internalism and externalism in semantics and epistemology* (pp. 13–34). Oxford University Press.
- Bruckner, M. (2022). Internal identity is (partly) dispositional identity. *Synthese*, 200(4). <https://doi.org/10.1007/s11229-022-03747-2>
- Burge, T. (1979). Individualism and the Mental. In P. French, T. Uehling, W. Wettstein (Eds.), *Midwest Studies in Philosophy vol. IV* (pp. 73–121).
- Burge, T. (1982). Other bodies. In A. Woodfield (Ed.), *Thought and object: Essays on intentionality* (pp. 97–120). Oxford University Press.
- Burge, T. (1998). Individualism and Self-Knowledge. *Journal of Philosophy*, 85, 649–663. <https://doi.org/10.5840/jphil1988851112>
- Carman, T. (2003). *Heidegger's Analytic: Interpretation, Discourse and Authenticity in Being and Time*. Cambridge University Press.
- Carr, D. (2003). Transcendental and Empirical Subjectivity: The Self in the Transcendental Tradition. In D. Welton (Ed.), *The New Husserl: A Critical Reader* (pp. 181–198). Indiana University Press.

- Crowell, S. (2008). Phenomenological immanence, normativity, and semantic externalism. *Synthese*, 160(3), 335–354. <https://doi.org/10.1007/s11229-006-9083-3>
- Crowell, S. (2013). *Normativity and Phenomenology in Husserl and Heidegger*. Cambridge University Press.
- Davidson, D. (1987). Knowing One's Own Mind. *Proceedings and Addresses of the American Philosophical Association*, 60(3), 441–458. <https://doi.org/10.2307/3131782>
- Donnellan, K. (1972). Proper Names and Identifying Descriptions. In D. Davidson & G. Harman (Ed.), *Semantics of Natural Language* (pp. 356–379). D. Reidel.
- Dretske, F. (1996). Phenomenal externalism or if meanings ain't in the head, where are qualia? *Philosophical Issues*, 7, 143–158. <https://doi.org/10.2307/1522899>
- Dreyfus, H. L., Hall, H. (1984) Introduction. In H. L. Dreyfus & H. Hall (Ed.), *Husserl, Intentionality, and Cognitive Science* (pp. 1–27). The MIT Press.
- Farkas, K. (2003). What Is Externalism? *Philosophical Studies*, 112, 187–208. <https://doi.org/10.1023/a:1023002625641>
- Farkas, K. (2006). Semantic Internalism and Externalism. In B. C. Smith & E. Lepore (Ed.), *The Oxford Handbook of the Philosophy of Language* (pp. 323–340). Oxford University Press.
- Farkas, K. (2008). *The Subject's Point of View*. Oxford University Press.
- Gertler, B. (2012). Understanding the Internalism-Externalism Debate: What Is the Boundary of the Thinker? *Philosophical Perspectives*, 26(1), 51–75.
- Gomes, A., Parrott, M. (2021). On Being Internally the Same. In U. Kriegel (Ed.), *Oxford Studies in Philosophy of Mind Volume 1* (pp. 315–340). Oxford University Press.
- Husserl, E. (1939). *Erfahrung und Urteil: Untersuchungen zur Genealogie der Logik*. Rev. and ed. by L. Landgrebe. Academia Verlagsbuchhandlung Prag. *Experience and Judgment: Investigation in a Genealogy of Logic*, trans. J. S. Churchill & K. Ameriks. Routledge & Kegan Paul, 1973.
- Husserl, E. (1950). *Die Idee der Phänomenologie*. Husserliana II, ed. W. Biemel. Nijhoff. *The Idea of Phenomenology*, trans. L. Hardy. Springer, 1999.
- Husserl, E. (1952). *Ideen zu einer reinen Phänomenologie und phänomenologischen Philosophie, Zweites Buch: Phänomenologische Untersuchungen zur Konstitution*. Husserliana IV, ed. M. Biemel. Nijhoff. *Ideas Pertaining to a Pure Phenomenology and Phenomenological Philosophy, Second Book: Studies in the Phenomenology of Constitution*, trans. R. Rojcewicz & A. Schuwer. Nijhoff, 1989.
- Husserl, E. (1959). *Erste Philosophie, Zweiter Teil: Theorie der phänomenologischen Reduktion*. Husserliana VIII, ed. E. Boehm. Kluwer. *First Philosophy: Lectures 1923/24 and Related Texts from the Manuscripts (1920–1925)*, trans. S. Luft & T. M. Naberhaus. Springer, 2019.
- Husserl, E. (1966). *Analysen zur passiven Synthesis: Aus Vorlesungs- und Forschungsmanuskripten 1918–1926*. Husserliana XI, ed. M. Fleischer. Nijhoff. *Analyses concerning passive and active synthesis: lectures on transcendental logic*, trans. A. J. Steinbock. Kluwer, 2001.
- Husserl, E. (1973). *Cartesianische Meditationen und Pariser Vorträge*. Husserliana I, ed. S. Strasser. Nijhoff. *Cartesian Meditations*, trans. D. Cairns. Kluwer, 1999.

- Husserl, E. (1976a). *Ideen zu einer reinen Phänomenologie und phänomenologischen Philosophie, Erstes Buch: Allgemeine Einführung in die reine Phänomenologie*. Husserliana III/1, ed. K. Schuhmann. Nijhoff. *Ideas for a Pure Phenomenology and Phenomenological Philosophy, First Book: General Introduction to Pure Phenomenology*, trans. D. Dahlstrom. Hackett, 2014.
- Husserl, E. (1976b). *Die Krisis der europäischen Wissenschaften und die transzendente Phänomenologie*. Husserliana VI, ed. W. Biemel. Nijhoff. *The Crisis of European Sciences and Transcendental Phenomenology*, trans. D. Carr. Northwestern University Press, 1970.
- Husserl, E. (1984). *Logische Untersuchungen, Zweiter Band, Zweiter Teil: Untersuchungen zur Phänomenologie und Theorie der Erkenntnis*. Husserliana XIX/2, ed. U. Panzer. Nijhoff. *Logical Investigations, Volume 2*, trans. J. N. Findlay, ed. D. Moran. Routledge, 2001.
- Husserl, E. (1989). Phänomenologie und Anthropologie. In T. Nenon & H.R. Sepp (Eds.), *Aufsätze und Vorträge (1922–1937)*, Husserliana XXVII. Kluwer. Phenomenology and anthropology. In *Psychological and transcendental phenomenology and the confrontation with Heidegger (1927–1931)* (Trans. and ed. T. Sheehan & R. E. Palmer). Springer, 1997.
- Husserl, E. (2003). *Transzendentaler Idealismus, Texte aus dem Nachlass (1908–1921)*. Husserliana XXXVI, ed. R. D. Rollinger in cooperation with E. Sowa. Kluwer
- Jackson, F. (2003). Narrow Content and Representation, or Twin Earth Revisited. *Proceedings and Addresses of the American Philosophical Association*, 77(2), 55–70. <https://doi-org.libproxy.helsinki.fi/10.2307/3219741>
- Jacquette, D. (2013). Phenomenological thought content, intentionality, and reference in Putnam's Twin Earth. *The Philosophical Forum*, 44(1), 69–87. <https://doi.org/10.1111/phil.12002>
- Kallestrup, J. (2012). *Semantic Externalism*. Routledge.
- Keller, P. (1999). *Husserl and Heidegger on Human Experience*. Cambridge University Press.
- Kripke, S. (1972). *Naming and Necessity*. Blackwell Publishing.
- Lycan, W. G. (2001). The case for phenomenal externalism. *Philosophical Perspectives*, 15(15), 17–35. <https://doi.org/10.1111/0029-4624.35.s15.2>
- Man-To, T. (2014). Husserl's Transcendental Idealism and Its Way Out of the Internalism-Externalism Debate. *Meta*, 6, 463–483.
- McCulloch, G. (1995). *The Mind and Its World*. Routledge.
- McDowell, J. (1992). Putnam on Mind and Meaning. *Philosophical Topics*, 20(1), 35–48.
- McGinn, C. (1977). Charity, Interpretation, and Belief. *The Journal of Philosophy*, 74, 521–535. <https://doi.org/10.2307/2025795>
- McGinn, C. (1989). *Mental Content*. Basic Blackwell.
- McIntyre, R. (1984). Searle on Intentionality. *Inquiry*, 27(1-4), 468–483. <https://doi.org/10.1080/00201748408602044>
- McIntyre, R. (1986). Husserl and the Representational Theory of Mind. *Topoi*, 5, 101–113. <https://doi.org/10.1007/BF00139224>
- Mellor, D. H. (1977). Natural kinds. *British Journal for the Philosophy of Science*, 28, 299–312. <https://doi.org/10.1093/bjps/28.4.299>

- Murchadha, F. (2003). Pierre Keller, Husserl and Heidegger on Human Experience. *Husserl Studies*, 19, 93–100. <https://doi.org/10.1023/a:1022260212405>
- Murchadha, F. (2008). Reduction, externalism and immanence in Husserl and Heidegger. *Synthese*, 160, 375–395. <https://doi.org/10.1007/s11229-006-9085-1>
- Putnam, H. (1975). The meaning of “meaning.” *Minnesota Studies in the Philosophy of Science*, 7, 131–193.
- Raatikainen, P. (2020). In defense of semantic externalism. *E-Logos – Electronic Journal for Philosophy*, 27(2), 57–70. <https://doi.org/10.18267/j.e-logos.479>
- Rowlands, M. (2003). *Externalism: Putting Mind and World Back Together Again*. McGill-Queen’s University Press.
- Segal, G. (2000). *A Slim Book about Narrow Content*. The MIT Press.
- Smith, A. D. (2008). Husserl and Externalism. *Synthese*, 160, 313–333. <https://doi.org/10.1007/s11229-006-9082-4>
- Smith, D. W., McIntyre, R. (1982). *Husserl and Intentionality*. D. Reidel.
- Spear, A. D. (2013). Intentionality and Indexicality: Content Internalism and Husserl’s Logical Investigations. In C. Svennerlind, R. Ingthorsson & J. Almäng (Eds.), *Johanssonian Investigations: Essays in Honour of Ingvar Johansson on His Seventieth Birthday* (pp. 574–607). De Gruyter.
- Ströker, E. (1993). *Husserl’s Transcendental Phenomenology*. Trans. L. Hardy. Stanford University Press.
- Wikforss, Å. (2008). Semantic Externalism and Psychological Externalism. *Philosophy Compass*, 3, 158–181. <https://doi.org/10.1111/j.1747-9991.2007.00107.x>
- Williamson, T. (2006). Can cognition be factorised into internal and external components? In R. Stainton (Ed.), *Contemporary Debates in Cognitive Science* (pp. 291–306). Blackwell Publishing.
- Woodfield, A. (1982). Foreword. In A. Woodfield (Ed.), *Thought and Object: Essays on Intentionality* (pp. v–xi). Oxford University Press.
- Zahavi, D. (2003). *Husserl’s Phenomenology*. Stanford University Press.
- Zahavi, D. (2004). Husserl’s Noema and the Internalism-Externalism Debate. *Inquiry*, 47, 42–66. <https://doi.org/10.1080/00201740310004404>
- Zahavi, D. (2008). Internalism, externalism, and transcendental idealism. *Synthese*, 160, 355–374. <https://doi.org/10.1007/s11229-006-9084-2>
- Zahavi, D. (2017). *Husserl’s Legacy: Phenomenology, Metaphysics, and Transcendental Philosophy*. Oxford University Press.
- Zemach, E. (1976). Putnam’s theory on the reference of substance terms. *The Journal of Philosophy*, 73(5), 116–127. <https://doi.org/10.2307/2025740>

Theories of Consciousness Should Not Be Afraid of Dreams

MICHAL KLINCEWICZ

This paper reviews some reasons for us believing that dreams are phenomenally conscious experiences. I point out that if dreams would be such then one should be able to draw analogies between them and waking conscious experiences at the level of folk psychology, behavior, and neuroanatomy. Instead, we find disanalogies at all these levels, so while we may have no strong reason to accept that dreams are unconscious, we have good reason to think they are different from conscious experiences. A safer bet is to think that they form a distinct *sui generis* psychological category. Sleep science should wait for consciousness science and until then treat its subject matter as distinct.

1. Introduction

We all seemingly know what conscious experiences are like and for those that do not there is a quip from Ned Block: ‘if you got to ask, you ain’t never gonna get to know’ (Block 1978, 281). Unfortunately, a general theory of consciousness is not around the corner. It is also far from settled which neural state or network is a necessary or sufficient condition for conscious experience (Giacino et al., 2014; Baars et al., 2003). We also do not have a settled criterion for detecting consciousness behaviorally (Seth et al 2005; Seth 2009). Verbal reports remain the gold standard and probably will remain in this role for the foreseeable future. This might be some consolation to cognitivist theorists of consciousness, for whom conscious experience is tied closely to subjective reports and thus to cognitive access (Cohen and Dennett 2011, Rosenthal 2007, Brown 2012). Conversely, this would be unsatisfactory to anyone that emphasizes the ineffability of aspects of conscious experience or puts a lot of stock in phenomenology (Block 2007; Block

2011). In other words, what one takes to be evidence of consciousness in dreams often depends on one's theoretical commitments (Sebastián 2014).

One such claim is the claim that dreams are phenomenally conscious experiences (Windt and Noreika 2010). The best philosophical argument for this claim, is that shortly after dreamers are woken up they can report about them (Windt 2013). There has been considerable debate since that argument was presented, but not outside the theoretical context of consciousness studies. When we take the argument outside of that context what we are left with are mere analogies. This is where we take the issue up.

This paper offers a challenge to thinking that there are *analogies* between dreams and waking conscious experience at the level of folk psychology, behavior, and brain activity. It also offers support for the view that it is best to taxonomize them as belonging to a *sui generis* category rather than as being in the same category as waking conscious experiences. Section 2 lays out the folk-psychologically framed analogy between conscious experiences and dreams and the challenge to it. Section 3 focuses on observed behavioral analogies between waking and dream states. Section 4 challenges the purported isomorphisms between brain activity during wakefulness and dreaming. Section 5 addresses lucid dreams. Finally, it briefly offers an alternative view based on Melanie Rosen's narrative fabrication theory of dream recall (Rosen 2013), which is supported by the disanalogies uncovered in Sections 2, 3, and 4. Section 6 addresses two objections.

2. Folk-psychological analogy between dreams and conscious experiences

The claim that dreams are phenomenally conscious has been widely discussed in philosophy¹ and it reached foundational status with Descartes' *Meditations*. Descartes tells us that 'as I think about it more carefully, I see plainly that there are *never* any sure signs by means of which being awake can be distinguished from

¹ Arguably, this assumption starts at least with Plato, when Socrates asks Theaetetus 'what proof you could give if anyone should ask us now, at the present moment, whether we are asleep and our thoughts are a dream, or whether we are awake and talking with each other in a waking condition' Plato (1921) *Theaetetus*, Cambridge: Harvard. To which Theaetetus replies that indeed there is no way of knowing and that 'the likeness between the one talk [in reality] and the other [in a dream] is remarkable' *ibid*. Theaetetus thinks that he might as well be dreaming of having that very conversation with Socrates.

being asleep' (Descartes 2008, 13; my italics). In this certainty begins Descartes' doubt about knowledge arrived at by the senses—in the purportedly obvious truth that being awake and being asleep are subjectively indistinguishable. The idea that this is obvious continues, even today in sophisticated recent treatments of the issue (Windt, Nielsen, and Thompson 2016).

There is at least one important folk-psychologically framed analogy that may support this latter claim, outside of the philosophical debate about consciousness: dreams and conscious experiences are remembered and forgotten for roughly the same reasons. For future reference, let us call dreams that are or may be remembered *r-dreams* and those that are forgotten *f-dreams* and their waking counterparts *r-experiences* and *f-experiences*.² If the analogy were true, the folk could attribute r-dreams/f-dreams using the same folk taxonomy and conditions that they use to attribute r-experiences/f-experiences.

The analogy is relatively simple. First, co-occurrence of strong emotions typically correlates with r-experiences (Cahill and McGaugh 1998). Emotionally salient experiences function like attractors for other memories, which create clusters of episodes that can be recalled with more fidelity. By analogy, dreams that involve extraordinary content are also more likely to be remembered.

It is perhaps not even controversial that some of our conscious experiences are generic and forgettable—these are the f-experiences. Whatever is not salient does not grab our attention and does not warrant extra resources. The most economical solution for our brain is to render most experiences in a generic way as 'some of that stuff over there' instead of kicking it up to full-blown, memorable, r-experiences (see: Dennett 1991, 354-356). Presumably, a similar process is responsible for r-dreams.

However, there are reasons to think that f-dreams are significantly different than f-experiences, at the level of folk attributions, even if the underlying causes of r-dreams and r-experiences are similar. Frequency and quality of dream recall depends on the quality of sleep, time of awakening, and other factors that have nothing to do with content and emotional saliency (Eichenlaub et al 2018,

² R-dreams are especially common in laboratory settings when participants are explicitly instructed to remember their dreams. In this context, the rate of recall can reach as high as 80% when participants are woken up during REM sleep and 50% when they are woken up during non-REM sleep: Nielsen T.A., Laberge L., Paquet J., et al. (2000), "Development of Disturbing Dreams During Adolescence and their Relation to Anxiety Symptoms", *Sleep* 23(6), 727-737. The distinction should also be distinguished from David Foulkes' notion of A- and B-dreams.

Giesbrecht and Merckelbach 2006). In other words, f-dreams correlate with factors that are different from those that correlate with f-experiences.

The second disanalogy between r-experiences and r-dreams is temporal. In scientific terms, normal verbal reports about r-experiences engage the contents of the sensory memory system in which representations decay rapidly. The special status of subjective verbal reports as criteria of conscious experience is connected to their engaging this aspect of the memory system. When r-experiences involve things that are significantly further in the past, they involve the long-term memory system and cannot play that particular role of criteria for consciousness (Craik and Tulving 1975).

Verbal behavior during sleep is sometimes treated by us folk *as if* it was a clear indication of conscious experience. But it can equally be interpreted as merely mouthing words without any connection to conscious experience. What speaks to the latter interpretation is that we often cannot follow up verbal behavior during sleep with some questions to settle whether these are merely mouthings of words or genuine reports. Following up with questions can and has been done after the person wakes up. But here, unlike in the case of conscious experiences while awake, the report is not engaging short-term memory, but long-term memory systems, so we lose the relevant scientific connection to consciousness.

To sum up, we have at least two reasons to not take verbal reports about dreams in the same way that we take verbal reports about normal conscious experiences. First, the factors contributing to the occurrence of f-dreams and f-experience are different. Second, we do not follow up on verbal utterances made during sleep in the same way that we could with typical verbal reports. If the folk analogy between r-dreams and r-experiences is not strong, then it is not as obvious why—at the level of folk psychology—people like Descartes claim that they are indistinguishable.

3. Behavioral analogies between dreams and conscious experiences

Another way to support the analogy between r-dreams and r-experiences is to argue that some behavior during sleep has a functional role similar enough to subjective verbal reports, even though in itself it is not actually a subjective verbal report itself. This role is what makes the two indistinguishable, in a Cartesian sense. For example, in the same way in which being in pain is manifested

differently across species or even individual humans, reporting on conscious experiences is manifested differently depending on whether it involves sleep or wakefulness.

There are some straight-forward candidates for evidence of this commonality. If patterns of speech during sleep match the content of retrospective dream reports after the sleeper wakes up, that is an independent reason to suppose that speech behavior during sleep is functionally equivalent to a subjective report—it gives us evidence that the dreamer was consciously aware of their dreams. A person that screams about birds during sleep and then reports that they had a dream about birds can be said to engage in something like reporting on the contents of their conscious experience during sleep by screaming: “birds!” Non-pathological sleep talking is fairly common (66.8% in a sample of 1000 Norwegian adults) (Bjorvatn et al 2010), even though coherent speech patterns during sleep occur most often in people with neurodegenerative disorders (Tachibana et al 1997; Abe et al 1984).

Further examples along these lines can be found in people with REM sleep disorder, which is a disturbance or failure of sleep paralysis. Typically, REM sleep disorder results in a variety of non-verbal behaviors, some of which are reflections of dream content (Schenck et al 1986, Chen et al 2013, Borek et al 2007). For example, one may be swatting at the air during sleep. Swatting at the air is the sort of thing one would do, if one had a conscious experience of being attacked by birds. So, if the dreamer reports that they dreamed of being attacked by birds and they swatted at the air during sleep we have evidence that the dream was a conscious experience. The swatting was, functionally, like reporting.

Corroborating evidence lies in a number of studies that show how the content of dreams can be manipulated (Schredl 2010). For example, proper names presented during REM sleep cause the sleeper to incorporate them into their dream (Berger 1963). We can easily imagine how sitting by the shore, we find ourselves thinking about birds and then realize that the seagulls have been squawking around the whole time.

However suggestive, all this does not suffice to support the functional analogy. Examples of complex sleep behavior, verbal utterances, and content-influences are all coupled with retrospective reports—the golden standard for evidence of consciousness. Without these, we would not be in a position to pronounce in any way about the psychological nature of what happens when we

sleep. Verbal reports import the thing that is at stake, namely, conscious experience—being a gold standard and all.

Of course, the charge of begging the question can be made back against the claim that dreams are not conscious experiences. Proponents of the claim that dreams are conscious could argue that denying or minimizing the significance of the connection between complex sleep behavior and retrospective reports builds in the assumption that dreams are not conscious. And so it would seem that we end up in a stalemate, in a situation that Daniel Dennett characterized as an ‘open, and theoretical question whether dreams fall inside or outside the boundary of experience’ (Dennett 1976, 170-1). In other words, we have theoretical assumptions doing most of the work in the interpretation of the phenomenon.

What speaks against there being a stalemate is the magnitude of difference between admittedly complex dream behavior and the behavior typically associated with waking conscious experiences. First, people that are awake can answer questions giving further detail about their current state of mind and alter their behavior if asked to do so. This form of interaction does not often happen with people that are sleep-talking, sleepwalking, or experiencing night terrors. It does not help to say to a dreamer: “there are no birds attacking you, it is a dream.”

Second, in a related point, people that are asleep lack the kind of control over their behavior that accompanies normal waking conscious experience. We would expect dreamers to be able to at least sometimes retrospectively report that they were asked something and that they could not answer because they were paralyzed. Control is closely connected to the folk conception of intentional action and this is why violent behavior during sleep is not criminalized (Pressman 2007, Siclari et al 2010).

These limitations and differences contrast sharply with the wide range of complex behaviors available to people that are awake. They suggest that whatever is going on during episodes of complex behavior during sleep is significantly different from what happens during episodes of complex behavior when people are awake. In conclusion, the best explanation of the relationship between complex behaviors during sleep and subsequent retrospective reports cannot be that they are mediated by conscious experience. The better inference is to the claim that the mediator is something *other* than conscious experience.

4. Analogies in brain activity

The human brain undergoes significant changes as it enters the different stages of sleep (Nir et al 2011, Nir et al 2013; Gvilia 2010; Brown et al 2012). Nonetheless, during all stages the brain is active, including areas involved in sensory processing, spatial processing, and multisensory integration, among others (Hong et al 2009). It is taken for granted in sleep science that most dreaming occurs during rapid eye-movement (REM) sleep, when eye movements appear functionally similar to those we observe during wakefulness (Miyachi et al 2009, Leclair-Visonneau et al 2010). Eye-movements together with sensory brain activation parallels a coupling between the two processes that is observed during wakefulness.

In many other respects, our brain wakes up during sleep, while our body remains inert (Siclari et al 2017). There is a return to rapid low voltage oscillations in EEG recordings during this time, which matches the features of recordings during wakefulness (Crunelli and Hughes 2009). In addition, the cortico-thalamic system that is online during wakefulness also comes online at this time (Saper et al 2001, Cohen 1974). Recordings of brain activity during REM sleep show patterns of activation similar to activity during wakefulness (Schwartz and Maquet 2002, Hobson and Pace-Schott 2002). Particularly active are sensory areas and areas correlated with imagery (Nir and Tononi 2010). Perhaps most compelling to that last claim is the line of research that shows how the content of retrospective reports about dream narratives can be predicted by software trained on patterns of brain activity recorded using fMRI scans of the visual cortex during sleep (Horikawa et al 2013). These are all powerful reasons to think that dreams are conscious—they have a similar neural signature, including complexity, as normal waking states. The only apparent difference is behavior, which may be explained by sleep paralysis.

Compared to the folk analogy between r-dreams/r-experiences or the functional analogies between verbal and non-verbal behavior in parasomnias, the similarities in brain activity may indeed make a more compelling case that dreams are conscious, especially to people that are committed to a particular theory of consciousness that comes with neuro-anatomical claims. However, when we look more closely at the areas of the brain that are *not* active during REM sleep, we get yet another disanalogy.

For one, the prefrontal cortex is typically disengaged during REM sleep, except during lucid dreams, more about which later (Maquet et al 1996, Muzur et

al 2002). This is significant, since areas of the prefrontal cortex are standardly thought to be the seat of complex cognition (Stuss and Knight 2013) and, on some views, subjective verbal reports (Lau and Passingham 2006). Some cognitivist theorists of consciousness also take the pre-frontal cortex to be a good candidate for the neural correlate of consciousness (Dehaene et al 2003, Baars et al 2003).

Another neural level analogy that may help with the claim that dreams are conscious channels the classic Tulving distinction between episodic, semantic, and implicit memory (Tulving and Schacter 1990, Tulving 1985a). Episodes, so those memories that we store in the episodic long-term memory system necessarily involve conscious experience, as Tulving's theory predicts. We would not be able to recall ourselves as dreaming without having stored them in that system. Without consciousness during dreams, there is no remembering them at all—or so it may appear to the folk.

The main problem with this argument is that there is a lot of evidence—some of it already mentioned above—that sophisticated processing can occur without awareness (e.g., Dehaene and Changeux 2011). There is even some evidence that we process the semantic meaning of objects without these objects *ever* being conscious (Sanguinetti et al 2013). The sum-total of this evidence suggests that the inference from an instance of remembering something to phenomenal consciousness is too quick. We process and store things without consciousness all the time, so why not during sleep?

Furthermore, Tulving's view is not the only game in town. There are alternative theories of narrative or episodic memory, which do not make auto-noetic consciousness a necessary component (Martin 2001, Nelson and Fivush 2004). The most theory-neutral interpretation of the connection between auto-noetic consciousness and episodic memory makes the relation between them contingent (Klein 2013). This leaves the possibility of retrospective reports that masquerade as genuine cases of remembering a conscious episode, so we are back to the start.

The wide range of neural evidence about the neural signature of consciousness should give one pause before drawing analogies between brain activity in and out of sleep. It is not enough that the brain is in many respects functioning just as it does during wakefulness. What matters is that the extra neural ingredient that results in phenomenal consciousness is also active at that time. And, as it turns out, scientists disagree about that.

5. Lucid dreams and their kind

Some dreams are lucid, that is, they are accompanied by meta-cognitive awareness and a feeling of control over their content (Kahan and LaBerge 1994). It remains controversial what process lies behind them, but it is uncontroversial that they occur and that they appear to be phenomenally conscious. If the only thing missing in regular dreams is their lucidity, whatever that turns out to be, then we have an independent reason to think that regular dreams are also conscious.

Philosophical skeptics about dream consciousness, such as Norman Malcolm, Daniel Dennett, and Gilbert Ryle, argued that lucid dreams are not genuine cases of conscious experience on a priori grounds. The idea of lucid dreaming implies, on their view, something other than sleep, so lucid dreams fall under some other category than conscious experiences. While this sort of conceptual analysis may be a good way to defend the claim that lucid dreams are not conscious experience in the context of philosophy, we do not have to go that way ourselves here.

For one, lucid dreamers routinely report having control and display a good deal of control over their mental states. With appropriate training, people can even move their eyes as directed while lucid dreaming in a phenomenon called ‘lucid-control’ (Gackenbach 2009). These eye-movements function like subjective verbal reports about one’s conscious experience. Anti Revonsuo (Revonsuo, 1995) reviewed a great deal of similar empirical evidence that altogether strongly suggests that, contrary to a priori philosophical arguments, lucid dreams have all the signs of conscious experience, corroborated by classic studies on the issue (LaBerge 1988, LaBerge et al 1981, Worsley 1988, Schatzman et al 1988). Since 1995, that evidence has grown.

Indeed, this evidence is compelling: lucid dreams are probably conscious experiences. However, note that none of the disanalogies mentioned in sections 2, 3, and 4 hold either. People that are lucid dreaming are doing something significantly different from those that are non-lucid dreaming and significantly like what they are doing when they are awake.

This comes out in the neuroscientific evidence, too, speaking to the neural disanalogy. EEG recordings show significant increase of 40 Hz band oscillations and coherence at frontal brain regions during lucid dreams, as compared with REM sleep (Voss et al., 2009). This comports with the already mentioned decrease of neural activity or its complete absence in the dorsolateral pre-frontal cortex

during REM sleep (Muzur et al 2002). During lucid dreams, activity in that region is increased (Nofzinger et al 1997). Many other empirical findings can be brought up to make a similar point.

Neural activity in frontal regions and increased coherence of firings suggest that lucid dreaming involves a brain state that is a hybrid between sleep and wakefulness. Arguably, there are other hybrids of these states as well and a number of in-between states, such as the hypnogogic state or hypnoponic state (Vaitl et al 2005). In other words, lucid dreams may involve mental states that are distinct from both dreaming and regular waking conscious experiences. Non-lucid dreams may be an in-between state themselves, somewhere between unconscious states and lucid dreams.

So what is going on? Melanie Rosen has argued that the best way to interpret the process of dream recall is as a form of fabrication that mirrors our tendency to rationalize strange elements in waking reports. Subjects tend to leave out supernatural or bizarre elements when reporting waking memories of stories (...). Confabulation is exacerbated in dreams by rapid memory loss and bizarre dream content (Rosen 2013, 13).

Indeed, when we recall an episode of dreaming, we are able to impose the relevant causal structure on its content, very much like we do when we recall things of which we were not at all aware (Nisbett and Wilson 1977). We end up with a report that is ostensibly about a particular episodic memory, say, but could be based on the contents of other memory systems, partial memories, or just the innate need that our psychological system has to impose causal structure, full stop.

Confabulation is not uncommon, even with fully conscious people, such as those with Korsakoff's syndrome. These people routinely make what appear to be retrospective reports about episodes that never actually happened and stay convinced that they are right, even when confronted with evidence to the contrary (Kopelman et al 2009). We cannot infer anything about the state of mind of someone that confabulates. They could have been completely unconscious at the time or consciously experiencing something else.

In conclusion, if all of this is true, then it should be unclear what folk psychological category dreams belong to and it should be clear that it is probably not 'conscious experience.' We typically talk about unconscious mental states in contrast to conscious ones, with the sometimes tacit assumption that these unconscious states could at least in principle have become conscious (Searle 1991). The view presented here leaves dreams in some in-between category that

is neither phenomenally conscious nor unconscious in the sense of unconscious mental state.

The moral from these philosophical observations for dream science is that the concept of consciousness comes with difficulties and muddles that may be best set aside in the interest of making better empirical predictions and hypotheses. The only thing that you lose is the purported connection between the science of dreams and the science of consciousness, which can be brought back once we have a more theory-neutral grip on the neural correlates of conscious experience. Currently, the science of consciousness is not there to help, it is there to make things more difficult by adding an extra theory-laden variable.

6. Objections and replies

Malcolm redux objection. In *On Dreaming* Norman Malcolm claims that subjective verbal reports are the only criteria we have for conscious experience (Malcolm 1956, Malcolm 1962) and concludes from this that a person that is asleep cannot be said to have any conscious experiences. The view presented here offers the same sort of argument, so it opens itself up to the same criticism that has been levelled against Malcolm.

Crucial to Malcolm's view is the notion of a criterion, which in Malcolm's usage is closely tied to confirmation. According to Malcolm, a declarative sentence that cannot be confirmed to be either true or false is unintelligible and senseless. Sentences that assert that dreams are conscious are hence unintelligible and senseless, since the only criteria that we have for conscious experience are subjective verbal reports, which happen to be impossible during sleep.

As a number of philosophers have pointed out, there are serious problems with Malcolm's idiosyncratic use of 'criterion' (Dunlop 1977, Chihara 1965). Most important of these is that the close connection between verbal reports and conscious experience that Malcolm exploits does not support the claim that dreams are unconscious. At best, it tells us something about the way that we often use words such as "consciousness" and "sleep" in non-scientific discourse (Putnam, 1962). This confuses the role of inductive inference in our regular language use with its role in scientific practice. On Malcolm's view science should allow only:

- (1) Conclusive verification based on the application of 'criteria', and (2)
- inductive inference in the most restricted possible sense: induction by simple

enumeration. But, as Mill remarked, no developed social science (or any other science, one might add) will ever be possible on this basis (Putnam 1962, 317).

So, the fact that we cannot report during sleep only shows that there is no way to settle the issue of whether dreams are conscious by reference to verbal reports alone. That does not mean that when scientists attempt to discover via experiment and probabilistic reasoning whether dreams are conscious, they are engaging in something unintelligible and senseless.

Reply. The position defended here does not demand that we adopt verbal reports as a criterion (in a Malcolmian sense) of conscious experience, just that such reports are currently a gold standard to be superseded once we have a suitable independent way of testing for phenomenal consciousness. What supports the view are empirically observed disanalogies and a folk psychological taxonomy that is open to revision, not a priori conceptual analysis.

Ryle redux objection. One may also object that the argument of this paper is a re-warmed serving of Gilbert Ryle. Similarly to Malcolm, Ryle argued that dreams are unconscious based on a priori reasoning. According to Ryle's view, dreams are unconscious because the concept <dream> contains a negation of the concept <waking experience> (Ryle 1954, 93-95). By analogy, just as we understand what a counterfeit coin is because we understand what a coin is, we understand what it means to be dreaming, because we understand what it means to have normal waking experiences. If Ryle is right, it is *by definition* impossible for dreams to be like waking conscious experiences.

One problem with Ryle's argument is that it conflates the distinction between creature and state/phenomenal consciousness (Manson 2000, Rosenthal 1986). Creature consciousness is a state of the organism, typically covering a whole range of processes and changes, that we categorize with general state terms, such as "sleep," "coma," or "minimally-conscious state." The type of consciousness that is at stake in the question of conscious dreams is state/phenomenal consciousness, which is at the heart of the so-called hard problem of consciousness (Chalmers 1995). State consciousness is a property of a mental state of the organism and not the whole organism. Sleep is a paradigmatic example of creature unconsciousness. We understand what it means to be sleeping (creature unconscious), because we understand what it means to be awake (creature conscious)—not because we understand what it means to have a

conscious mental state. So, this paper (like Ryle) is wrong in making the analogies/disanalogies that it does—it is a sort of category mistake.

Reply. The dispute does not concern the concept <dream> and whether it contrasts in the Rylean sense with the concept <conscious experience>. If the disanalogies presented here are good then we should treat them as evidence of a sui generis category for dreams, distinct from the concept of conscious experience, not as evidence that they are not conscious.

Tilburg University, Netherlands

References

- Abe K., Amatori M. and N. Oda (1984), “Sleepwalking and Recurrent Sleepwalking in Children of Childhood Sleepwalkers”, *The American Journal of Psychiatry* 141(6), 800-801.
- Baars B.J., Ramsøy T.Z. and S. Laureys (2003), “Brain, Conscious experience and the Observing Self”, *Trends in Neurosciences* 26(12), 671-675.
- Berger R.J. (1963), “Experimental Modification of Dream Content by Meaningful Verbal Stimuli”, *The British Journal of Psychiatry* 109(463), 722-740.
- Bjorvatn B., Grønli J. and S. Pallesen (2010), “Prevalence of Different Parasomnias in the General Population”, *Sleep Medicine* 11(10), 1031-1034.
- Block N. (2007), “Consciousness, Accessibility, and the Mesh Between Psychology and Neuroscience”, *Behavioral and Brain Sciences* 30(5-6), 481-499.
- Block N. (2011), “Perceptual Consciousness Overflows Cognitive Access”, *Trends in Cognitive Sciences* 15(12), 567-575.
- Borek L.L., Kohn R. and J.H. Friedman (2007), “Phenomenology of Dreams in Parkinson's Disease”, *Movement Disorders* 22(2), 198-202.
- Brown R. (2012), “The Myth of Phenomenological Overflow”, *Consciousness and Cognition*, 21(2), 599-604.
- Brown R.E., Basheer R., McKenna J.T., Strecker, R., and R. McCarley (2012), “Control of Sleep and Wakefulness”, *Physiological Reviews* 92(3), 1087-1187.
- Cahill L. and J.L. McGaugh (1998), “Mechanisms of Emotional Arousal and Lasting Declarative Memory”, *Trends in Neurosciences* 21(7), 294-299.
- Chalmers D.J. (1995), “Facing up to the Problem of Consciousness”, *Journal of Consciousness Studies* 2(3), 200-219.
- Chen M.C., Yu H., Huang Z.-L. and J. Lu (2013), “Rapid Eye Movement Sleep Behavior Disorder”, *Current Opinion in Neurobiology* 23(5), 793-798.
- Chihara C.S. (1965), “What Dreams are Made On”, *Theoria* 31(3), 145-158.
- Cohen D.B. (1974), “Toward a Theory of Dream Recall”, *Psychological Bulletin* 81(2), 138.
- Cohen M.A. and D.C. Dennett (2011), “Consciousness Cannot be Separated from Function”, *Trends in Cognitive Sciences* 15(8), 358-364.

- Craik F.I. and E. Tulving (1975), "Depth of Processing and the Retention of Words in Episodic Memory", *Journal of Experimental Psychology: General* 104(3), 268-294.
- Crunelli V. and S.W. Hughes (2009), "The Slow (< 1 Hz) Rhythm of Non-REM Sleep: A Dialogue Between Three Cardinal Oscillators", *Nature Neuroscience* 13(1), 9-17.
- Dehaene S. and J.-P. Changeux (2011), "Experimental and Theoretical Approaches to Conscious Processing", *Neuron* 70(2), 200-227.
- Dehaene S., Sergent C. and J.-P. Changeux (2003), "A Neuronal Network Model Linking Subjective Reports and Objective Physiological Data During Conscious Perception" *Proceedings of the National Academy of Sciences of the United States of America* 100(14), 8520-8525.
- Dennett, D.C. (1976), "Are Dreams Experiences?", *The Philosophical Review*, 151-171.
- Dennett, D.C. (1991), *Consciousness Explained*, Penguin Books, New York.
- Descartes R. (2008), *Meditations on First Philosophy: With Selections from the Objections and Replies*, Oxford University Press, New York.
- Dunlop, C.E. (1977), *Philosophical Essays on Dreaming*, Cornell University Press, Ithaca NY.
- Eichenlaub J.B., van Rijn E., Gaskell M.G., Lewis P.A., Maby E., Malinowski J.E., Walker M.P., Boy F., and M. Blagrove (2018), "Incorporation of Recent Waking-life Experiences in Dreams Correlates with Frontal Theta Activity in REM Sleep", *Social Cognitive and Affective Neuroscience* 13(6), 637-647.
- Gackenbach, J. (2009), "Electronic Media and Lucid-control Dreams: Morning after Reports", *Dreaming* 19(1), 1-6.
- Giacino J.T., Fins J.J., Laureys S., and N. Schiff (2014), "Disorders of Consciousness after Acquired Brain Injury: The State of the Science", *Nature Reviews Neurology* 10, 99-114.
- Giesbrecht, T. and H. Merckelbach (2006), "Dissociatieve Symptomen en Slaap [Dissociative symptoms and sleep]", *Tijdschrift voor Psychiatrie* 48(3), 207-215.
- Gvilia I. (2010), "Underlying Brain Mechanisms that Regulate Sleep–Wakefulness Cycles", *International Review of Neurobiology* 93, 1-21.
- Hobson J.A. and E.F. Pace-Schott (2002), "The Cognitive Neuroscience of Sleep: Neuronal Systems, Consciousness and Learning", *Nature Reviews Neuroscience*, 3(9), 679-693.
- Hoerl C and T. McCormack (2001), *Time and memory: Issues in philosophy and psychology*, Oxford, Oxford UK.
- Hong C.C.H., Harris J.C., Pearlson G.D., Kim J.-S., Calhoun V.D., Fallon J.H., Golay X., Gillen J.S., Simmonds D.J., van Zijl P.C.M, Zee D.S. and J.J. Pekar (2009), "fMRI Evidence for Multisensory Recruitment Associated with Rapid Eye Movements During Sleep", *Human Brain Mapping* 30(5), 1705-1722.
- Horikawa T., Tamaki M., Miyawaki Y., and Y. Kamitani (2013), "Neural Decoding of Visual Imagery During Sleep", *Science* 340(6132), 639-642.
- Kahan T.L. and S. LaBerge (1994), "Lucid Dreaming as Metacognition: Implications for Cognitive Science", *Consciousness and Cognition* 3(2), 246-264.
- Klein S.B. (2013), "Making the Case that Episodic Recollection is Attributable to Operations Occurring at Retrieval Rather than to Content Stored in a Dedicated Subsystem of Long-term Memory", *Frontiers in Behavioral Neuroscience* 7.

- Kopelman M.D., Thomson A.D., Guerrini I. and J.E. Marshall (2009), "The Korsakoff Syndrome: Clinical Aspects, Psychology and Treatment", *Alcohol and Alcoholism* 44(2), 148-154.
- LaBerge S.P. (1988), "The Psychophysiology of Lucid Dreaming", *Conscious Mind, Sleeping Brain*, 135-153.
- LaBerge S.P., Nagel L.E., Dement W.C., and V.P. Zarcone Jr. (1981), "Lucid Dreaming Verified by Volitional Communication During REM Sleep", *Perceptual and Motor Skills* 52, 727-732.
- Lau H.C. and R.E. Passingham (2006), "Relative Blindsight in Normal Observers and the Neural Correlate of Visual Consciousness", *Proceedings of the National Academy of Sciences* 103(49), 18763-18768.
- Leclair-Visonneau L., Oudiette D., Gaymard B., Leu-Semenescu S. and I. Arnulf (2010), "Do the Eyes Scan Dream Images During Rapid Eye Movement Sleep? Evidence from the Rapid Eye Movement Sleep Behaviour Disorder Model", *Brain* 133(6), 1737-1746.
- Malcolm N. (1956), "Dreaming and Skepticism", *The Philosophical Review*, 14-37.
- Malcolm N. (1962) *Dreaming*, Routledge & Paul, London.
- Manson N. (2000), "State Consciousness and Creature Consciousness: A Real Distinction", *Philosophical Psychology* 13(3), 405-410.
- Maquet P., Péters J.-M., Aerts J., Delfiore G., Degueldre C., Luxen A. and G. Franck (1996), "Functional Neuroanatomy of Human Rapid-Eye-Movement Sleep and Dreaming", *Nature* 383(6596), 163-166.
- Martin M. (2001), "Out of the past: Episodic recall as retained acquaintance" in C. Hoerl and T. McCormack, 257-284.
- Miyauchi S., Misaki M., Kan S., Fukunaga T. and T. Koike (2009), "Human Brain Activity Time-locked to Rapid Eye Movements During REM Sleep", *Experimental Brain Research* 192(4), 657-667.
- Muzur A., Pace-Schott E.F. and J.A. Hobson (2002), "The Prefrontal Cortex in Sleep", *Trends in Cognitive Sciences* 6(11), 475-481.
- Nelson K. and R. Fivush (2004), "The Emergence of Autobiographical Memory: A Social Cultural Developmental Theory", *Psychological Review* 111(2), 486-511.
- Nielsen T.A., LaBerge L., Paquet J., Tremblay R.E., Vitaro F. and J Montplaisir (2000), "Development of Disturbing Dreams During Adolescence and their Relation to Anxiety Symptoms", *Sleep* 23(6), 727-737.
- Nir Y., Massimini M., Boly M., and G. Tononi (2013), "Sleep and Consciousness", *Neuroimaging of Consciousness*. Springer, 133-182.
- Nir Y., Staba R.J., Andrillon T., Vyazovskiy V.V., Cirelli C., Itzhak F. and G. Tononi (2011), "Regional Slow Waves and Spindles in Human Sleep", *Neuron* 70(1), 153-169.
- Nir Y. and G. Tononi (2010), "Dreaming and the Brain: From Phenomenology to Neurophysiology", *Trends in Cognitive Sciences* 14(2), 88-100.
- Nisbett R.E. and T.D. Wilson (1977), "Telling More than We Can Know: Verbal Reports on Mental Processes", *Psychological Review* 84(3), 231-259.
- Nofzinger E.A., Mintun M.A., Wiseman M., Kupfer D.J. and R.Y. Moore (1997), "Forebrain Activation in REM Sleep: An FDG PET Study", *Brain Research* 770(1), 192-201.

- Plato (1921), *Theaetetus*, Cambridge: Harvard.
- Pressman M.R. (2007), “Disorders of Arousal from Sleep and Violent Behavior: The Role of Physical Contact and Proximity”, *Sleep* 30(8), 1039-1047.
- Putnam H. (1962), “Dreaming and 'Depth Grammar'”, in R.J. Butler (ed.), *Analytical Philosophy*, Oxford University Press, Oxford.
- Revonsuo A. (1995), “Consciousness, Dreams and Virtual Realities”, *Philosophical Psychology* 8(1), 35-58.
- Rosen M.G. (2013), “What I Make Up When I Wake Up: Anti-experience Views and Narrative Fabrication of Dreams”, *Frontiers in Psychology* 4 (514), 1-15
- Rosenthal D. (1986), “Two Concepts of Consciousness”, *Philosophical Studies* 49(May), 329-359.
- Rosenthal D. (2007), “Phenomenological Overflow and Cognitive Access”, *Behavioral and Brain Sciences* 30(5-6), 522-523.
- Ryle G. (1954), *Dilemmas: The Tarner Lectures 1953*, Cambridge University Press, Cambridge UK.
- Sanguinetti J.L., Allen J.J. and M.A. Peterson (2013), “The Ground Side of an Object Perceived as Shapeless Yet Processed for Semantics”, *Psychological Science*, 256-264.
- Saper C.B., Chou T.C. and T.E. Scammell (2001), “The Sleep Switch: Hypothalamic Control of Sleep and Wakefulness”, *Trends in Neurosciences* 24(12), 726-731.
- Schatzman M., Worsley A. and P. Fenwick (1988), “Correspondence During Lucid Dreams Between Dreamed and Actual Events”, *Conscious Mind, Sleeping Brain*, 155-179.
- Schenck C.H., Bundlie S.R., Ettinger M.G., and M.W. Mahowald (1986), “Chronic Behavioral Disorders of Human REM Sleep: A New Category of Parasomnia”, *Sleep* 9(2), 293-308.
- Schredl M. (2010), “Characteristics and Contents of Dreams”, *International Review of Neurobiology* 92, 135-154.
- Schwartz S. and P. Maquet (2002), Sleep Imaging and the Neuro-psychological Assessment of Dreams”, *Trends in Cognitive Sciences* 6(1), 23-30.
- Searle J. (1991), “Consciousness, Unconsciousness and Intentionality”, *Philosophical Issues*, 45-66.
- Sebastián, M.Á. (2014), “Dreams: An Empirical Way to Settle the Discussion Between Cognitive and Non-cognitive Theories of Consciousness”, *Synthese* 191, 263–285.
- Seth A. (2009), “Explanatory Correlates of Consciousness: Theoretical and Computational Challenges”, *Cognitive Computation* 1(1), 50-63.
- Seth A., Baars B.J. and D.B. Edelman (2005), “Criteria for Consciousness in Humans and Other Mammals”, *Consciousness and Cognition* 14(1), 119-139.
- Siclari F., Khatami R., Urbaniok F., Nobili L., Mahowald M.W., Schenck C.H., Cramer Bornemann M.A. and C.L. Bassetti (2010), “Violence in Sleep”, *Brain* 133(12), 3494-3509.
- Siclari F., Baird B., Perogamvros L., Bernardi G., LaRocque J.J., Riedner B., Boly M., Postle B.R. and G. Tononi (2017), “The Neural Correlates of Dreaming”, *Nature Neuroscience* 20(6), 872-878.
- Stuss D.T. and R.T. Knight (2013) *Principles of Frontal Lobe Function*, Oxford University Press.

- Tachibana N., Kimura K., Kitajima K., Shinde A., Kimura J. and H. Shibasaki (1997), "REM Sleep Motor Dysfunction in Multiple System Atrophy: With Special Emphasis on Sleep Talk as its Early Clinical Manifestation", *Journal of Neurology, Neurosurgery & Psychiatry*, 63(5), 678-681.
- Tulving E. (1985a), "How Many Memory Systems are There?", *American Psychologist* 40(4), 385-398.
- Tulving E. (1985b), "Memory and Consciousness", *Canadian Psychology/Psychologie Canadienne* 26(1), 1-12.
- Tulving E. and D.L. Schacter (1990), "Priming and Human Memory Systems", *Science* 247(4940), 301-306.
- Vaitl D., Birbaumer N., Gruzelier J., Jamieson G.A., Kotchoubey B., Kübler A., Lehmann D., Miltner W.H.R., Ott U., Pütz P., Sammer G., Strauch I., Strehl U., Wackermann J. and T. Weiss (2005), "Psychobiology of Altered States of Consciousness", *Psychological Bulletin*, 1(131), 98-131.
- Voss U., Holzmann R., Tuin I. and J.A. Hobson (2009), "Lucid Dreaming: A State of Consciousness with Features of Both Waking and Non-lucid Dreaming", *Sleep* 32(9), 1191-1200.
- Windt J.M. (2013), "Reporting Dream Experience: Why (not) to be Skeptical About Dream Reports", *Frontiers in Human Neuroscience* 7.
- Windt J. M., Nielsen, T., and E. Thompson (2016), "Does Consciousness Disappear in Dreamless Sleep?", *Trends in Cognitive Sciences*, 20(12), 871-882.
- Windt J.M. and V. Noreika (2010), "How to Integrate Dreaming into a General Theory of Consciousness—a Critical Review of Existing Positions and Suggestions for Future Research", *Consciousness and Cognition*, 1091-1107.
- Worsley A. (1988), "Personal Experiences in Lucid Dreaming", *Conscious Mind, Sleeping brain*, 321-342.

What Theory Did Mary Develop?

KRISTJAN LOORITS

It's time to revisit Mary, the beloved fictional neuroscientist – an expert on human color vision who has spent, rather ironically, her entire life in a black and white research facility and has never seen colors (Jackson 1982; 1986). What the thought experiment invites us to believe is that regardless of the depth and details of Mary's scientific knowledge on human color vision, Mary does not know what it is like to see red.

The original thought experiment is framed even more boldly. Accordingly, Mary knows “all the physical facts about us and our environment” (Jackson 1986, 291). But let us be more realistic. Let us assume that Mary is smart enough to know, exceptional neuroscientist as she is, that the information processing capacities of the human brain (including hers) have their limits and knowing *all the physical facts* must be far beyond these limits.

A less fantastical scenario would keep the thought experiment interesting enough: Mary has developed a comprehensive theory that allows her to explain and predict human behavior, including people's responses to colors. Let us also assume (again, quite realistically) that Mary's theory is compatible with the laws of physics. Thus, when monitoring brain processes at different levels of electro-chemical interactions, Mary finds no anomalies – no “ghosts in the machine” or any kind of inexplicable primitive emergence. Everything runs according to strictly mechanical principles, just like in the good old Leibniz's mill.

Readers who are drawn towards quantum brain approaches are welcome to imagine that the “mill” also contains non-trivial quantum mechanical processes (i.e., quantum processes that are functionally and computationally relevant in the context of the theory). If that is the case, then Mary is also familiar with the nature of these processes and can confirm that all “quantum-enhanced” mechanisms in the brain follow the known laws of physics (at least known to her, if not to us). After all, if we were to detect in the brain “non-classical” effects that did not obey

the experimentally discovered laws of quantum mechanics, then on what ground could we even call them quantum effects?

Wittgensteinians and proponents of 4E approaches need not worry. Mary's theory is about *behavior* – not about consciousness or perception. Regardless of how the notions of perception, experience, consciousness, etc., ought to be analyzed, it is widely agreed that the proximate causes of our muscle movements are to be found in the workings of our nervous systems.

The bottom line is, Mary knows that nothing in the behavior-inducing brain processes violates the laws of physics. And although that piece of knowledge is not yet a theory, any comprehensive theory of behavior must be compatible with it.

That said, Mary is well aware that all special sciences are to some extent *autonomous* and do not describe their subject matter in terms of microphysical processes alone. She also expects this to be the case with the theory of behavior. Presumably, there are myriad patterns and regularities to be discovered in different levels of neurocognitive processing. Moreover, it is highly probable that many explanatorily relevant processes reach far beyond the limits of the skull – both in the spatial and the temporal sense – including the evolutionary history of our species and the complex fabric of social interactions.

Yet, despite this methodological, terminological, and theoretical autonomy, Mary knows that the final theory of human behavior must be metaphysically reductive in a loose (and perhaps old-fashioned) sense of the term: if she were to “zoom in” at whichever resolution to whichever section of the complex causal fabric between sense organs and behavior-executing muscles, she would never find any violations of the laws of physics.

In sum, what the thought experiment asks us to imagine is that Mary has developed a metaphysically reductionist theory (in the loose sense outlined above) that allows her to explain complex human behavior – including behavioral responses to colors. Since part of that complex behavior is philosophers' tendency to utter or write statements about the phenomenal aspect of color experiences, Mary's theory ought also to explain that. In other words, Mary's theory should provide us with nothing less than a solution to the so-called *meta-problem* of consciousness.

1. Approaching the meta-problem

While the hard problem is about explaining the existence of phenomenal experiences (e.g., why experiencing red feels the way it does), the meta-problem concerns the question as to why people express beliefs of having phenomenal experiences (e.g., Chalmers 2018). If such beliefs are conscious, then we may also call them *phenomenal judgements* (e.g., Chalmers 1996).

Since the meta-problem is about explaining behavior (behavioral expressions of phenomenal judgements), it belongs technically among the “easy” problems (easy in the sense of being approachable within the traditional framework of psychology and cognitive neuroscience). Mary is aware of that. As noted before, she has no reason to think that the muscle movements executing the expressions of phenomenal judgements are ultimately caused by anything other than some mechanical processes in the brain.

As mentioned earlier, Mary also knows that a comprehensive theory of behavior requires not only detailed neurobiological descriptions but also many higher-order terms, concepts, equations, models, etc. (just as a theory of life cannot be framed in terms of primitive chemical reactions but needs to outline the role of DNA, basic principles of evolution, specific mechanisms of metabolism, etc.).

When it comes to complex behavioral responses to colors, the theory must outline some higher-order psychological processes that would make those responses unsurprising and intelligible. And of course, a particularly interesting question for Mary (who has never seen colors) is why people express *phenomenal judgements* about perceived colors.

In the search for a representative example, Mary considers the sentence: “The redness I experience has a distinctive qualitative character that is impossible to describe.” But after familiarizing herself with the nuances of the hard problem debate, she ends up with the phrasing she finds more instructive: “The redness I experience has a distinctive qualitative character that cannot be described in fully structural terms.”

Mary relies here on Chalmers’s idea according to which the hard problem results from a tension between scientific structuralism and the nonstructural appearance of experienced qualities. Accordingly, all scientific descriptions (theories, hypotheses, statements, models, etc.) are ultimately about structure and dynamics, and phenomenal experiences are problematic precisely because their qualitative content (e.g., the experienced redness) cannot be described in fully structural terms (e.g., Chalmers 1996; 2003).

Although sympathetic towards Chalmers's take on scientific structuralism, Mary finds it unnecessarily strict and restrictive. Regardless of whether *all* scientific descriptions are *only* about structure and dynamics, it seems plausible enough that when it comes to the basic neurocognitive mechanisms that govern our behavior, all that matters is the functional structure those mechanisms make up and sustain.

What this means is that with regard to explaining behavior (not consciousness, remember), it does not matter whether the neurons or any other explanatorily relevant basic units have some intrinsic nonstructural nature or not, it only matters how those units interact with each other (that would be true even if some relevant interactions followed the “non-classical” laws of quantum mechanics).

A mechanical Leibniz's mill runs the way it does (i.e., *behaves*) due to the causal relations between its gears and springs. Perhaps those gears and springs have the most blissful and wonderful inner lives (or perhaps their inner lives are horrifying and dreadful), but that would make no difference on the functioning of the mill. And that *irrelevance of the intrinsic* is all the structuralism one needs to acknowledge in the present context.

In sum, what Mary's theory needs to explain is why the structurally (relationally, functionally) analyzable mechanisms that govern our behavior sometimes produce utterances like “the redness I experience has a distinctive qualitative character that cannot be described in fully structural terms.”

2. Towards the theory

The actual process of developing the theory is creative and messy: different bottom-up and top-down paths are explored (“upwards” from basic “gears and springs”, and “downwards” from the expressions of sophisticated beliefs); popular folk-psychological notions are analyzed (some rejected, some sustained, some modified, some given a more precise technical meaning); computational models of different cognitive capacities and psychological traits are developed, illuminating unifying stories are compiled, etc.

The result is an immensely complex multileveled description of human psychology. Since the theory is ultimately about behavior, it needs to specify how each acknowledged psychological notion manifests in behavioral tendencies. And since the theory is comprehensive (as assumed in the thought experiment), it needs to specify by which mechanisms any given psychological capacity is (typically)

implemented. In short, the theory is a textbook example of multilevel functionalist theory.

Mary is familiar with the philosophical arguments against functionalism, including the accusation that a purely functional account leaves out the subjectively felt phenomenal character – *what it is like* to be a (complex) psychological organism. But since her theory concerns *only* behavior, she can dismiss that line of criticism entirely.

Nevertheless, Mary's functionalist theory adopts many terms that are also used in everyday and philosophical language to depict subjectively felt phenomenal states. For instance, the theory acknowledges the existence of pains and pleasures, fear and anger, aversion and affection. But of course, in the context of the theory, all these (and other similar) notions are defined solely by their functional roles.

Mary's theory also acknowledges the distinction between conscious and nonconscious states. Within her functionalist framework, that distinction concerns mainly (but not only) reportability and communicability. In that regard, Mary finds many of Dennett's ideas illuminating and helpful (especially Dennett 1991; 2017).

Broadly speaking, a typical conscious state carries some informational content (about the environment and/or about the subject's own psychological or physiological states) that can be communicated to other people. The fact that it *can* be communicated but is not revealed automatically means that the behavioral manifestations of conscious states are highly context sensitive. That key feature of conscious behavior is manifest in all mundane situations where some valuable information is communicated to allies yet withheld from rivals, or when the information is memorized for long periods of time – ready to be communicated once (if ever) the “right” context is identified. (The communication-based analysis of consciousness is also compatible with the basic idea of the Global Workspace model: conscious information is widely accessible by different subsystems in the brain, including the ones that process information about different aspects of the overall situation and context.)

One interesting class of conscious states manifests in situations where the subject reports identifying a property yet fails to communicate its defining character to others. In those cases, the subject may insist that the identified property has a definite and easily recognizable character that is primitive and impossible to describe.

While these states are certainly interesting, their existence poses no deep mystery in the strictly functionalist framework of Mary's theory. At any given moment, our brain hosts a vast number of informational states that guide our behavior. Only a tiny fraction of that information is communicable to others. So, in the context of the theory, the "ineffable" property has two key features: first, the subject's ability to communicate the fact that the property is identified, and second, the subject's inability to communicate the nature of the cognitive mechanisms by which the property is identified.

Resulting phenomenal judgements are sometimes framed in a remarkably sophisticated philosophical language and expressed with firm conviction and powerful emotional charge. When applying her theory to such cases, Mary must constantly remind herself that she is only interested in behavior. Thus, when a subject insists that her experienced redness is a "very real" property whose existence and "primitive" nature is impossible to deny, Mary needs not to ponder whether the subject might be actually right. When in the role of researcher, her only task is to evaluate whether the theory can explain the corresponding behavior (i.e., the confident and emotionally charged *expression* of the phenomenal judgement).

Interestingly enough, Mary realizes that even if she is only interested in behavior, she must accept a class of psychological states that are surprisingly similar to the "ineffable" phenomenal experiences the naïve subjects talk about. The fact that she "must" accept them should be understood in the spirit of special science autonomy: removing those states from the theory would lead to an inevitable loss in predictive and explanatory power (e.g., Wallace 2003). But of course, in the context of the theory, also those states are ultimately grounded in functionally analyzable mechanisms.

The two key features of these states are already mentioned. The first is the subject's ability to communicate the fact that she has identified a property (e.g., the redness or the smell of a rose). The evolutionary rationale behind that ability is easy to understand: all the possible benefits of having the information that, say, the strawberries behind the nearby hill are red and sweet, can thereby be shared with an ally.

The second key feature is the subject's inability to communicate the nature of the process by which her cognitive apparatus has picked up and identified the property in question. Also here, it is easy to understand why evolution may have deprived us from that further ability. While being able to communicate

information about the environment and about our internal states is often useful, it can be expensive in terms of cognitive processing. So, without any significant selective pressure, why should evolution provide us with the extravagant capacity to communicate the detailed information about these cognitive processes? (As Dennett 1991 famously argues, for the purposes of communication, it is often more useful to form clear and simple “user illusions”.) A more interesting question is, why do some of us (especially philosophers) find such communicative limitations so fascinating and problematic.

Presumably, the reasons for that are varied and multiple. But among the simplest ones is that when we find phenomenal properties philosophically problematic (i.e., when we recognize the hard problem), we are not focusing merely on the communication barrier. Instead, in those situations our attention targets simultaneously both key aspects of these properties: their identifiability and indescribability. And apparently, when targeted that way, these properties tend to trigger emotional reactions.

Usually, we are not indifferent towards the “ineffable” properties but tend to like or dislike them (e.g., according to Humphrey 2006, the very function of phenomenal properties or “qualia” is to *matter to us*). And when we are fascinated by their ineffable nature and the resulting hard problem, that element of valence receives an extra spin: an emotional response towards the (readily communicable) fact that we are unable to describe an identified property.

Of course, in the context of Mary’s theory, emotions also ought to be analyzed via their functional roles. Very broadly speaking, their primary role is to motivate behavior (e.g., Baumeister et al. 2007). Simply put, each emotional reaction makes some behavioral responses more likely. And as before, to say that an emotion is consciously experienced means that the subject is able to communicate some aspects of the corresponding motivations and behavioral tendencies to other people.

When mapping the psychological space of emotions, there are obviously many other dimensions beside valence that need to be acknowledged. A state of sadness can be equally as unpleasant as a state of anger, but anger tends to lead to a faster and more aggressive response (among other differences). Similarly, feeling suspicious or amazed may both motivate explorative behavior, but the first motivates a more cautious approach. So, when a subject is *amazed* and *fascinated* (and perhaps a little frustrated) by the ineffable character of the experienced

redness, her overall emotional condition can be analyzed as a complex and nuanced functional state.

Thus, step by step, Mary's strictly functionalist theory that is ultimately concerned with mere behavior adopts more and more psychological terms that people use when talking about the puzzling nature of consciousness and the hard problem. A naïve subject or a philosopher may say, with great confidence and strong emotional commitment, that the redness she experiences cannot be explained in terms of neural processes because its qualitative nature is fundamentally different from the structurally analyzable neurocognitive mechanisms. And Mary's theory agrees that this is exactly how things appear from the perspective of the subject!

Of course, in the context of Mary's functionalist theory, the "subject" is not a primitive metaphysical entity but, roughly put, a subsystem that maintains information in an easily communicable form – ready to be released to the motor systems (for the execution of the communication) once the overall context (monitored by other subsystems) is suitable. Thus, when a person insists that a property appears to her as primitive, ineffable, and inexplicable, the theory agrees that this is exactly how the property is displayed in the communicative subsystem: easily identifiable as "this" rather than "that" (e.g., red rather than blue), yet indescribable in terms of the underlying cognitive mechanisms by which the property is identified and represented within the system.

As a sidenote, in practice most phenomenal properties are not indescribable in the absolute sense. At the very least, people are usually able to talk about similarity and difference relations between the experienced qualities. Most of us would agree that the taste of strawberries is closer to the taste of raspberries than it is to the taste of fried bacon. We may also use metaphorical and associative language. For instance, we may detect "yearning" in the sound of a trumpet or a "summery whiff" in the smell of lilacs. All these communicable similarity relations and associations may reveal something very important about the implicit informational content and cognitive functions of the apparently simple phenomenal properties (e.g., Koch 2004; Dennett 1988; 1991; Fink et al. 2021).

Nevertheless, a typical subject would insist that all these communicable relations do not *exhaust* the nature of how the experienced quality appears to her. In other words, the subject would insist that even if phenomenal qualities can be described to some extent, they are never *fully* describable. And if Mary's theory is supposed to explain *all* phenomenal judgements, then it must also explain the ones

that emphasize the existence of that indescribable aspect (consider: “the blueness I experience has a hint of turquoise in it and brings to mind the Mediterranean Sea on a warm sunny day, but of course, these connections do not exhaust its qualitative nature which is ultimately impossible to describe”).

In any case, at this point it would be misleading to say that the theory aims to explain phenomenal judgements in terms of simple mechanical interactions between neurons (or whatever basic units the theory relies upon) – just as it would be misleading to say that biology explains birds’ ability to fly in terms of simple chemical reactions. Thus, when the theory claims or predicts that from the perspective of a subject a certain property is easily identifiable yet (fully) indescribable in structural terms, it is not a metaphorical way to talk about some microlevel interactions. In the context of the theory, “identifiability” and “indescribability” are well-defined and perfectly legitimate higher-order notions whose meanings correspond quite closely to how those terms are used in everyday language. Put simply, the claim is meant to be taken *literally* – just as literally as the claim that birds need feathers in order to fly.

The more advanced the theory becomes, the harder it is for Mary to convince herself that the theory is *only* about behavior. While explaining behavior is its ultimate goal, the theory is also about different psychological states whose existence must be acknowledged in order to explain behavior. Regardless, Mary is still under the impression that her functionalist theory leaves out the “real” phenomenal experiences. Although she has never seen colors, she has heard music, smelled roses, tasted exquisite wines, etc. She feels that each of these phenomenal experiences has a “positive” qualitative character that exceeds the theory’s dry and hollow “something identifiable yet (fully) indescribable”.

Mary has also read Jackson’s articles about herself (Jackson 1982, 1986). She has often thought about Jackson’s prediction of what would happen if she left her black and white environment and saw colors for the first time. Would she learn a new fact? Mary is not sure how to answer that.

She knows she would be surprised to some extent, for that is what the theory predicts! She *really would be* in a new psychological state for the very first time. Just as she was when tasting her favorite Pinot Noir or hearing Mozart’s 40th symphony for the first time. Also in those situations she had familiarized herself with all the relevant theoretical predictions and descriptions (which turned out to be accurate) and knew what to expect, scientifically speaking. But among the

things to expect was also the sense of novelty and surprise (psychologically speaking).

It is often argued that functionalism has its limits and cannot account for the existence of phenomenal consciousness or solve the hard problem. But what Mary begins to understand is that her functionalist theory has led her to a more fundamental limit – a limit of what can be meaningfully said about phenomenal consciousness and the hard problem.

3. Reaching the horizon

It seems to be absolutely essential for the hard problem that it is not one of the easy problems (the very concept of the hard problem is often clarified by contrasting it with the easy problems). It seems equally clear that the meta-problem is technically one of the easy problems (Chalmers 2018). Yet, what Mary begins to understand is that once the meta-problem has undergone a detailed enough analysis and received a satisfactory scientific solution, it becomes impossible to articulate which aspect (if any) of the hard problem remains still unresolved.

It does not necessarily mean that solving the meta-problem automatically solves the hard problem. After all is said and done, Mary still feels that the theoretical account of her gustatory experience of Pinot Noir does not capture what it is “really” like for her to taste the wine. Yet, whenever she tries to *describe* the aspect that is not captured by the theory, she notices that her descriptions correspond to how the theory would describe her psychological reaction.

As a part of an experiment, Mary decides to listen to the sound of Stradivarius violin she has never heard before. Her theory has provided her with a detailed multilevel description of the psychological state she is about to enter. She has read and understood the description but is well aware that understanding the description does not generate the corresponding state within her brain (that too is predicted by the theory).

Once Mary hears the sound, she can confirm that the theory’s description was accurate – there were no “surprises” in the scientific sense of the term. Then again, at the human level, the experience was genuinely new and surprising (as the theory also predicted), and she cannot resist the feeling that she has thereby learned something new about her fellow human beings who have experienced the sound before. She has also no reason to doubt that she would have a similar reaction to seeing colors for the first time (as Jackson 1982 predicts).

Is such an impression of new knowledge illusory? Once again, Mary is not sure how to respond to that. On the one hand, entering a novel psychological state is a genuinely new situation for the subject – some new properties are identified in a manner the subject is unable to describe (at least not fully: as noted earlier, the subject may be able to provide partial descriptions, e.g., in terms of associations or similarity relations with other “ineffable” properties). Thus, in a way it is true that no amount of theoretical knowledge can fully replace the first-hand experience and the resulting sense of learning something new. On the other hand, it is equally true that such psychological reactions are predicted by the theory and do not provide any new data for the theory (Mary has no need to modify the theory after her novel experience).

Mary realizes that no further hypotheses or speculative additions to the theory could ever change the dynamics outlined above. If it is the case that understanding a description of a psychological state does not automatically generate that state in one’s brain, then no amount of speculative stories about some yet-to-be-discovered new laws of physics would make any difference. The perceived gap between the theory and the first-hand experience is not a flaw in the theory – it is a result of a *real difference* between two distinct psychological states: the state of perceiving a new property and the state of understanding the description of the first state.

Mary returns to the question: if the hard problem concerns the apparent existence of ineffable and structurally unanalyzable phenomenal properties, and if her functionalist theory provides a comprehensive account of the psychological states in which the subject can identify a property while being unable to describe or analyze the property in structural terms, then what aspect of the hard problem remains still to be explained? The “real” phenomenal aspect?

At that point Mary feels like she is driving in circles: whatever there is to be said about the “real” phenomenal aspect, the exact same thing can be said about its theoretical counterpart – with an additional clause that this is how the “ineffable” (i.e., identifiable yet indescribable) property appears to the subject (i.e., to the cognitive subsystem responsible for interpersonal communication). Mary realizes that from this point forward there is nothing anyone can say that would bring us any closer to resolving the long-lasting philosophical debate about the hard problem.

As a scientist, she is convinced that her theory is as complete and accurate as a theory can be – providing not only a comprehensive account of behavior, but

also of the higher-order psychological states whose scientific descriptions correspond to how people describe their immediate phenomenal experiences. Then again, when listening to the sound of the Stradivarius and enjoying her favorite Pinot Noir, she is equally convinced that her qualitative experiences cannot be fully reduced to the mechanical interactions her theory ultimately rests upon. The immediately experienced qualities feel so “meaty” and “substantial”, while the corresponding theoretical notions are nothing but higher-order features of a mechanical system (Leibniz’s mill) whose nature is fully defined by what it does.

Mary has developed a habit of keeping a journal where she writes her thoughts and feelings about the “real” phenomenal experiences. At other times she examines these entries through the lens of her theory and takes great joy in recognizing how well the theory explains the psychological states under which she has written the passages. In her final entry she writes:

No doubt that the theory is a great victory for reductionists! It seems clear now that for each clever argument against reductionism, my reductionist theory can provide a psychologically compelling explanation of why a person would put forward such an argument. It seems weird, though, that the theory hasn’t made a reductionist out of me yet. Then again, that’s exactly what the theory predicts. Apparently, for someone who is as aware of the theory’s mechanistic roots as I am, it is psychologically impossible to accept that all these wonderful sounds, smells, and emotions are nothing but an agglomeration of simple causal interactions between dumb gears and springs.

Anyway, enough of that already. I’m really excited about tomorrow! Looking forward to finally getting out and seeing all the colors! I have read so much about what to expect – including all these weird ways in which my expectations and theoretical knowledge will interfere with the upcoming experience. Still, as the updated model shows, the surprise and the sense of novelty will definitely be there! And the overall sensation is supposed to be very pleasant! Well, I guess I’ll find out soon enough.

University of Helsinki

References

- Baumeister, R. F., Vohs, K. D., DeWall, N., and Zhang, L. (2007), “How emotion shapes behavior: Feedback, anticipation, and reflection, rather than direct causation”, *Personality and Social Psychology Review* 11(2), 167–203.

- Chalmers, D. (1996), *The Conscious Mind: In Search of a Fundamental Theory*, Oxford: Oxford University Press.
- Chalmers, D. (2003), “Consciousness and its place in nature”, in S. Stich and F. Warfield (2003), 247–272.
- Chalmers, D. (2018), “The Meta-Problem of Consciousness”, *Journal of Consciousness Studies* 25(9–10), 6–61.
- Dennett, D. (1988), “Quining Qualia”, in A. Marcel and E. Bisiach (1988), 42–77.
- Dennett, D. (1991), *Consciousness Explained*, Boston: Little, Brown.
- Dennett, D. (2017), *From Bacteria to Bach and Back: The Evolution of Minds*, New York: W. W. Norton.
- Fink, S. B., Kob, L., and Lyre, H. (2021), “A structural constraint on neural correlates of consciousness”, *Philosophy and the Mind Sciences* 2:7. <https://doi.org/10.33735/phimisci.2021.79>
- Humphrey, N. (2006), *Seeing Red*, Cambridge, MA: Harvard University Press.
- Jackson, F. (1982), “Epiphenomenal qualia”, *Philosophical Quarterly* 32, 127–36.
- Jackson, F. (1986), “What Mary didn’t know”, *Journal of Philosophy* 83, 291–295. doi: 10.2307/2026143
- Koch, C. (2004), *The quest for consciousness: a neurobiological approach*, Englewood: Roberts and Co.
- Marcel, A. and E. Bisiach (eds.) (1988), *Consciousness in Contemporary Science*, Oxford University Press.
- Stich, S. and F. Warfield (eds.) (2003), *The Blackwell Guide to Philosophy of Mind*, Oxford: Blackwell.
- Wallace, D. (2003), “Everett and structure”, *Studies in History and Philosophy of Modern Physics* 34, 87–105.

Gesture and Literary Experience

KALLE PIHLAINEN

It has been my privilege to follow Valteri's work since his licentiate and doctoral theses, and I'm honoured to contribute to this collection in celebration of his 50th birthday, which also marks over a quarter century of his impressive philosophical contributions. Although our primary research topics have mostly been quite distant, our exchanges have been consistently rewarding, with each of us, I hope, finding relevance in the other's work. Notably, our ideas aligned most closely in the early 2000s when I began exploring "neurophenomenology," and we found ourselves studying many of the same scholars—several of whom I lean on here. During that period, Valteri's meticulous critiques of, for example, commonly inflated ideas relating to mirror neurons and empathy, helped keep my thinking more grounded. And, though we didn't publish jointly on these themes, we have since co-authored on topics like the causal theory of perception and mimesis in psychology. Currently, we continue to collaborate on predictive processing – another area that Valteri introduced to me. For me, this partnership remains as rewarding as ever.

In this essay, in great part inspired by years of conversation with Valteri, I undertake an exploratory effort somewhat removed from my main work, examining the ways in which literary texts evoke a sense of agency in readers—a phenomenon frequently ignored in literary theory and philosophy. I propose that this experience of agency, along with its emotional resonance, can be usefully approached through Maurice Merleau-Ponty's thinking about touch, gesture and language. Merleau-Ponty argued that gestures embody emotions directly, rather than simply representing them. Attempting to extend this idea, I argue that a form of "verbal gesticulation" within texts generates meaning not solely through intellectualized or acquired interpretations, but rather through an embodied, immediate engagement. This perspective challenges conventional approaches to reader engagement with literary texts. Rather than returning agency to fictional situations or character identification, I suggest that a reader's sense of agency

emerges in significant part through a dynamic, embodied interaction with the text itself.

For some, the notion that readers engage with texts in ways that involve a pre-reflective or pre-linguistic “knowing” or understanding shaped by fundamental forms and structures rather than explicit symbols or references may seem incongruent with established practices of literary analysis. However, I argue that such scepticism often reflects a lingering—and, admittedly, quite understandable—textualist bias in literary theory and philosophy. In contrast, foregrounding embodiment in reading points to a mode of engagement that operates alongside, or perhaps even precedes, traditional conceptions of language and meaning.¹ My hope is that questioning the general textualist bias might thus serve as a reminder of the indispensable role of embodied experience in the production of literary and aesthetic meaning. In this way, relating textuality to touch, gesture and embodied “knowing” could deepen understanding of the unique agency and experiential depth that writing can offer—an idea that, though admittedly exploratory for me here, has also found inspiration in many of our discussions with Valtteri over the years. This piece, then, is both a tribute to our intellectual friendship and an attempt to project from our numerous, often far-flung discussions into some new, experimental realms.

1. Touch and Textualist Thinking

Existential phenomenology has long positioned “touch” as a fundamental, even disruptive point of contrast to reflective thought. But it doesn’t automatically follow that we should assign the same primacy to the haptic in discussions of literary texts and their experiential qualities. It seems reasonable to begin by asking why the other senses—visual, auditory, olfactory, and gustatory, for

¹ Of course, there are clear precedents to be found for going this route outside the group of existential-phenomenological thinkers I rely on here, most notably, in the work of George Lakoff and Mark Johnson (see, centrally, Lakoff and Johnson 1980 and 1999). Johnson writes: “In the embodied mind, abstract reason is not separate from the sensorimotor system, but rather builds on it. Sensorimotor experience is schematized—as in image schemas (e.g. containers, paths, contact, balance, centrality) and motor schemas (grasping, pushing, pulling, moving) that have ‘logics’ that are regular consequences of perception and action. Our more abstract concepts are developed via metaphorical extensions of these basic sensorimotor structures, and our abstract reasoning involves inferences that are basically structure-preserving projections of sensorimotor inferences.” This, as he goes on to say, already “provides a major critique of any view that treats meaning and conceptual structure as radically ungrounded and arbitrary.” (Johnson 1999, 85; it is also worth comparing these ideas to those of Ruthrof 1997 and Gallagher 2005, all working in the same direction.)

example—couldn't hold equally valid roles in conveying the embodied experience of reading? These senses, too, stimulate moving representations and resonate with readers, invoking sympathetic engagement, and all have been present in discussions of experientiality.² The interest in touch here stems, however, from its arguably unique role in the experience of agency, a quality that is sometimes claimed to distinguish it from the other senses. Unlike sight or hearing, often viewed as remaining largely observational and receptive, touch establishes a reciprocal dynamic, suggesting a kind of agency unavailable through vision or sound alone. With touch, agency is exercised and perceived simultaneously, establishing an immediate feedback loop that is not equally present in the other senses—or, to be accurate, is present only when accompanied by proprioceptive and/or motorsensory feedback.

This circularity of touch is famously noted by Samuel Todes in *Body and World*, where he argues that feeling or touching includes a self-reflective component:

in the *very act* of feeling something, e.g. feeling a smooth surface by moving my hand across it, I feel myself feeling it. In vision as in thought, in order to catch sight of myself making sense of things, I must do so in a second-order act taking my first as its objective. I must do so, in short, reflectively. But feeling is *reflexive*; I make sense of my own making sense of something within the first-order act of sense-making. (Todes 2001, 266)

Whether things are as straightforward as Todes would have it remains debatable (the Gibsonian tradition would at least partially contest this, it seems; also see Derrida 2005, esp. 82 ff. for a more complicated, albeit parallel reading), the circularity of touch remains distinct in that all remains in one modality. And the particular immediacy suggested is certainly a tempting idea, framing it as uniquely resistant to the detached, analytical nature of reflective thought.

² Indeed, a question for further thought is whether literary experientiality even in the sense of “lived experience” and embodiment should be bound to the (five or more physical) senses alone. Beliefs and values, for example, seem to attach to embodied states that cannot be explained through the senses but instead manifest more as nonlocalized states or urges. This kind of visceral understanding might better be approached through other means, although the haptic seems to provide a good starting point in terms of its mechanisms, particularly its obvious circularity, as noted. (For an initial, exploratory sketch of a roadmap toward a more visceral approach, see Ball 2005.)

In a related observation (and in relation to another existential phenomenologist, Michel Henry), Dan Zahavi (1999, 192–193) emphasizes the key importance of mode of perception, explaining that “functioning subjectivity and radical alterity both belong to a totally different ontological dimension than the one dominated by vision.” He clarifies: “It is not because the Other is an Other, but because it is a self that I cannot perceive it directly.” To me, Zahavi’s point suggests that a sense or experience of agency—what else, after all, defines a “functioning subjectivity”?—might be better understood by exploring ontological dimensions not dominated by vision. And it is in this exploration that the haptic offers itself readily: touch involves a direct, embodied circularity, a continuous feedback loop of sensation that intertwines one’s body with the world, a characteristic that differentiates it significantly from the detached observational nature of vision.

Introducing a further challenge, however, this exploration of agency through the immediacy of touch also raises questions about how textuality might frame, or even limit, our understanding of experience. The interplay between language and experience is especially complex in academic dealings with texts, where experience often risks being reduced to language. Thomas J. Csordas (1999, 146) captures this tension with his concept of “hungry metaphor,” a term that critiques how extreme forms of textualism can subsume experiences under the dominion of text. In this textualist thinking—still so prevalent in many fields of cultural study reliant on literary theory and particularly on the ideal of deconstruction—the “beyond the text” that is denied by the methodological circumscription of texts as self-contained *from the point of view of strictly textual interpretation*, has been extended to encompass (or consume?) experiences and embodied knowledge essential to understanding the fundamental bodily attitudes on which texts also rely in their meaning production. Thus, the idea of there being “nothing outside the text” suggests not merely that interpretation must be textually grounded, but that our modes of understanding experience should be constrained within textual limits. Understood in this way, textuality appears to consume more than its fair share of the world.

To resist this “hungry” textuality, cultural analysis can instead frame experience as denoted by language rather than constituted by it, a shift that could rehabilitate lived experience as essential to understanding texts (cf. Csordas 1999, 145–147). Csordas’ (1999, 152) “cultural phenomenology” advocates for embodiment-focused analysis, an approach that foregrounds the sensory

modalities involved in perception. He emphasizes that a cultural-phenomenological approach allows for intersubjective perception that is attuned to embodied dimensions, such as the healer's attunement to another's distress. By shifting attention from the textual to the experiential, cultural analysis can recentre the role of the body in understanding texts. In a similar vein, my inquiry here into touch—and, by extension, into the embodied knowledge communicated through gesture—serves as a means of challenging the often-unquestioned primacy of vision in literary and theoretical discourse.

Through focusing on touch and gesture, we can examine how these embodied elements challenge traditional visual dominance and highlight the haptic's unique role in constructing knowledge (for excellent arguments in this same direction, see, for example Marks 2002; Fredriksson 2014 and 2022). Todes, for his part, critiques conventional theories of embodied subjectivity for their visually oriented biases, showing how such theories detach the perceiving self from bodily engagement with the world. For Todes (2001, 46),

When the self, becoming a pure inactive spectator, retreats from the body, detaching itself as a disembodied visual point of view, its body feelings become inert “in” the body. They lose the mobility that is characteristic of body feelings in the active body. [...] This involves loss of the basic sense of vital, corporate *unity* in one's body.

On this view, the motionlessness of the perceiving (evaluating rather than engaging) subject creates, then, the illusion of a detached cognitive faculty. Voluntarily withdrawing from their physical surroundings, perceivers adopt an attitude of objectivity and aloofness, which can easily lead to a generalizing view of human beings and their involvement in the world. In this way, perception becomes an act of observation rather than participation, creating a subjectivity that is more intellectual than embodied.

2. Reconsidering Literary Gesture

The question I want to begin my investigation into the limits of embodied “knowing” with is whether touch and gesture might be something that can be carried over into literary representation. Or, perhaps, more moderately, whether that kind of transfer of embodied meaning is something that we even want to

signify with our use of such terms. To clarify my intentions: there are certainly many ways to represent touch and gestures, and the more mimetic the means, the more unproblematic their “readings.” At the same time, any practice of gesture—or gesturing—in literary texts needs to be of a very different kind.

Indeed, even identifying gesture in literature is a fraught endeavour. As Merleau-Ponty (1962, 215) explains in his exploration of the communicative power of physical gestures: “The sense of the gestures is not given, but understood, that is, recaptured by an act on the spectator’s part. The whole difficulty is to conceive this act clearly without confusing it with a cognitive operation.” Yet, even suggesting that literary texts can be approached in these same terms introduces an apparent contradiction: When dealing with texts, how can anything be considered external to such cognitive operations? Where could this kind of non-cognitive meaning reside in the text? Or, from the point of view of a reader: How can something be available to us through a text—apparently a thoroughly cognized medium—without being subject to a cognitive operation on our part in the reading (and appropriation) of it?

Here, Merleau-Ponty’s (1962, 213–214) further observation that language and gesture are not distinct but intersect through shared bodily understandings comes to the rescue:

The spoken word is a genuine gesture, and it contains its meaning in the same way as the gesture contains its. This is what makes communication possible. In order that I may understand the words of another person, it is clear that his vocabulary and syntax must be “already known” to me. [...] The spoken word is a gesture, and its meaning, a world.

With this, it becomes easier to understand how Merleau-Ponty can claim that capturing the sense of a gesture is not a cognitive operation. What remains less obvious is why he also identifies the spoken word as “a genuine gesture.” On one hand, he appears to refer to the “act” of *speaking*, which obviously has a strong bodily component and aligns it with gesture. At the same time, he makes allowances for the cultural “world” of spoken language and “already known” meanings. Indeed, he sees speech acts as akin to gestures, something that we read fluently as part of our shared embodiedness *and* our embeddedness in cultural contexts. Yet they also have the capacity to point not only to commonly shared meanings but—specifically in their being “gestures” (and hence distinct from

signs)—to idiosyncratic positions and subjective interpretations; in short, as Merleau-Ponty eloquently summarizes, to a world. This does not mean, however, that we cannot relate to this world and reach some kind of understanding of it through communication; rather, this relationality is what communication ideally promotes.

Merleau-Ponty's perspective resonates with ideas from Nietzsche and Husserl, who also saw consciousness as deeply embodied and contextually embedded (for a useful take on Nietzsche as a philosopher of embodiment see Tambornino 2002; on Husserl in this connection, see Derrida 2005, esp. 348 ff.). As Merleau-Ponty (1962, 137) summarizes: "These elucidations enable us clearly to understand motility as basic intentionality. Consciousness is in the first place not a matter of 'I think that' but of 'I can'." This foundational claim points to an active, bodily intentionality that cannot be reduced to cognitive acts alone. This idea, taken further, implies that language itself operates as a kind of bodily gesture—a means of interacting with the world. Yet his comparison implies more than just a shared dynamic; it suggests that gestures carry a depth of meaning that words alone may not fully encapsulate:

The linguistic gesture, like all the rest, delineates its own meaning. [...] Verbal 'gesticulation' [...] aims at a mental setting which is not given to everybody, and which it is its task to communicate. But here what nature does not provide, cultural background does. Available meanings, in other words former acts of expression, establish between speaking subjects a common world, to which the words being actually uttered in their novelty refer as does the gesture to the perceptible world. And the meaning of speech is nothing other than the way in which it handles this linguistic world or in which it plays modulations on the keyboard of acquired meanings. (Merleau-Ponty 1962, 216–217)

With this statement, he unequivocally connects the gesturing ("gesticulating") of words to the "linguistic world" of communal, established and available meanings. Yet it seems odd that he would draw the parallel between gestures and speech if this were all there was to it. If the two were identical, the analogy would lose significance, which leads us to question how gestures might apply to complex verbal representations as well—despite any potential difficulties. Gestures and speech overlap but do not wholly merge, revealing a creative potential in language

that invites both cultural familiarity and subjective reinterpretation.³ In other words, by positioning gesture as a bridge between the “natural” and “cultural” worlds, Merleau-Ponty further underscores its power to integrate physiological functions, possibilities, and constraints with our cultural embeddedness. In this way, it permits the creation of new and spontaneous meanings through familiar and traditional means, which could, in turn, bear significance for understanding the functioning of literary texts.

3. Communication as Recognition

Merleau-Ponty’s exploration of gesture, especially as a bridge between bodily engagement and cultural meaning, also provides an essential perspective on freedom in communication. In this view, communication itself enacts a form of freedom—a common interpretation derived from the existential phenomenologists’ elaboration of the self–other relation. As Merleau-Ponty writes:

The communication or comprehension of gestures comes about through the reciprocity of my intentions and the gestures of others, of my gestures and intentions discernible in the conduct of other people. It is as if the other person’s intention inhabited my body and mine his. The gesture which I witness outlines an intentional object. This object is genuinely present and fully comprehended when the powers of my body adjust themselves to it and overlap it. The gesture presents itself to me as a question, bringing certain perceptible bits of the world to my notice, and inviting my concurrence in them. Communication is achieved when my conduct identifies this path with

³ Elsewhere Merleau-Ponty also introduces the distinction of “spoken language” and “speaking language.” (For a very rich and detailed investigation of all these issues, see Zlatev 2023.) As Merleau-Ponty explains these two modes of language operation: “one might draw a distinction between a *speaking word* and a *spoken word*. The former is the one in which the significant intention is at the stage of coming into being. Here existence is polarized into a certain ‘significance’ which cannot be defined in terms of any natural object.” (Merleau-Ponty 1962, 229) To me, this conception of “*speaking language*” (what he elsewhere refers to as “primary speech” is best understood as an intersubjective carrier of intentionality “polarized,” as he says, by a subject’s unique position. (For more on “primary speech” versus “secondary speech,” see, for example, Morris 2004, 83.) Given the weight he gives to the expression of freedom and agency, and, perhaps more importantly, to the possibility of communication itself, this also begins, I think, to answer the question of why literary texts can productively be considered in terms of this kind of creative gesturing.

its own. There is mutual confirmation between myself and others. Here we must rehabilitate the experience of others which has been distorted by intellectualist analyses, as we shall have to rehabilitate the perceptual experience of the thing. (Merleau-Ponty 1962, 215)

Through this dynamic, the gestures we witness outline an intentional world of objects that becomes shared through embodied mutuality. Further, the invitation to “concurrence” and the identification of one’s own path with that gestured by the other (or present in the aesthetic object) are both markers of freedom and choice. Yet there is also an automatic, embodied reaction implied in the experience of the other’s intentions “inhabiting my body.” The recognition involved is thus simultaneously a recognition of the freedom and the responsibility implicit in the other being separate from oneself: freedom because it marks oneself as being equally autonomous, responsibility because the other’s being “like me” brings (at least for these thinkers) an ethical entailment.

Importantly, this emphasis—on the role of intersubjectivity and communication as Merleau-Ponty develops it here—also suggests that the gestural dynamic could extend into literature more readily. And, of course, Merleau-Ponty is not alone in proposing this. We find a similar suggestion reiterated by Derrida (who could, on many occasions, be considered an existential phenomenologist). He writes: “One says that a gesture is pertinent when it touches, touches rightly, concerns, refers to, or stands as it has to in relation to what has to be, and thus adjusts itself to the contact. It is said, figuratively, about discourse itself or thought that inscribes itself in it.” (Derrida 2005, 269) Although Derrida (in line with his emphatic commitment to textualism) views this as a figurative gesture rather than a literal one, his language reveals a temptation to understand the gesture within literary texts as an embodied form of communication. This reading, too, challenges the boundaries between intellectual and bodily communication, suggesting that gestures and textual engagement can intersect in the reader’s embodied response.

A similar perspective appears in Sartre, who portrays literary writing as a “pact of generosity” between author and reader, promoting a mutual understanding that resists objectification. Sartre (1948, 51) explains that “what the writer requires of the reader is not the application of an abstract freedom but the gift of his whole person, with his passions, his prepossessions, his sympathies, his sexual

temperament, and his scale of values.”⁴ Sartre’s notion aligns with Merleau-Ponty’s emphasis on subjective freedom in communication, where the reader’s embodied engagement mirrors the intentionality of the writer.

By interpreting Sartre’s concept of the literary “pact” as a call for full engagement, we see a striking parallel with Merleau-Ponty’s idea of gesture as an invitation to inhabit shared meanings. Both thinkers suggest that this engagement is not merely cognitive but involves an ethical openness that fosters intersubjective understanding. As a consequence, we might say that literary or, in Barthean parlance, specifically “writerly”⁵ texts foster a meeting of intentionalities. In engaging with a text, readers bring their agency into alignment with the text’s gestures, allowing for a shared experiential world. Using Merleau-Ponty’s terms, the text’s affordances (albeit imaginary ones) invite the reader’s intentionality, thereby enacting a kind of gesture. This effectuation of the “literary pact” both demands and nurtures free engagement, affirming the gesture-like quality of literary representations and the “poeticity” of language. In using natural rather than logical language and by embracing its “fictionality” within this space, the literary text affirms its subjective and personal nature. This relationship between the text and reader is also why a purely functional or pragmatic text—like an instruction manual—can be argued to lack the literary “gesture.” Arguably, a manual does not extend an invitation to subjectivity, as its affordances are strictly prescriptive. Unlike a “writerly” text, it does not carry the same embodied intentionality and thus does not engage the reader’s agency in the same relational way. Conversely, the creative intentionality present in “writerly” texts appears to align with Merleau-Ponty’s view of gesture as an invitation to engagement. So, would it now be far-fetched to apply Merleau-Ponty’s statement about gesture (which I quoted above), instead, to *the text*? A text that, to re-cite Merleau-Ponty on gesture, similarly “presents itself to me as a question, bringing certain perceptible bits of the world to my notice, and inviting my concurrence in them. Communication is achieved when my conduct identifies this path with its own.” (Merleau-Ponty 1962, 215) In recognizing this dynamic, we are called to understand literature as

⁴ Furthermore, he notoriously claims, “the work of art, from whichever side you approach it, is an act of confidence in the freedom of men” (Sartre 1949, 63).

⁵ This seems a more productive term for the kinds of texts I discuss here. For Barthes: “The writerly text is a perpetual present, upon which no *consequent* language (which would inevitably make it past) can be superimposed; the writerly text is *ourselves writing*, before the infinite play of the world (the world as function) is traversed, intersected, stopped, plasticized by some singular system (Ideology, Genus, Criticism) which reduces the plurality of entrances, the opening of networks, the infinity of languages.” (Barthes 1974, 5)

more than a cognitive exercise, inviting a bodily, experiential reading that mirrors Merleau-Ponty's idea of gesture.

4. Embodied Literary Agency

When agency is discussed in relation to literary texts, the conversation typically revolves around how acts by agents are portrayed within the text. At times, theoretical attention also shifts to how literature uniquely enables the portrayal of agents' experiences—an ability rarely shared by other representational forms. Literature, in particular, allows and even encourages the representation of internal thoughts and intentional lives, aspects of subjectivity that are often considered inaccessible in other genres. But these are primarily questions of representation, while my concern here is with action—with literature as a performative mode that enacts agency rather than solely depicting it.⁶ And this is where the comparison to gesture as presented by Merleau-Ponty really comes into its own. Remember, as he puts it: “The gesture which I witness outlines an intentional object” (1962, 215). Similarly, recognizing a text as “writerly” involves perceiving the intentionality it presents, thereby extending an invitation to agency within the communicative act itself. This kind of intentionality is an aspect of texts that has been largely ignored, however, due to the prevailing textualist orientation in literary theory and philosophy, which have focused mainly on the text's intensionality and its linguistic meaning-making processes.

This oversight in recognizing textual intentionality reflects more than just a disregard for the author's role in meaning-making—the most common move. More significantly, it represents a denial of the text's function as an agentive force with an intentional orientation toward the world and an active role in its communication with the reader. In effect, this constitutes an aestheticization of the text, reducing it to a passive site of subjective interpretive play, often dominated by the literary critic. Yet this kind of intertextual/literary deconstructive playfulness disempowers the text, stripping it of its place in a world of meaningful relations, substituting this with a passive role wherein meanings can only be “read

⁶ David Bleich's study of agency, particularly through what he terms the “economies of everyday life,” provides an excellent exploration of how embodied feelings and passions shape our engagement with the world. For a discussion of agency and symbolic forms as gestures informed by these “economies,” see especially Bleich's remarks on language and gesture and his observations on how feelings and passions underlie the coordination of symbolic forms in daily life. (See Bleich 2006, esp. 610–618.)

into” it. To regard a “writerly” text in this way is, in my view, a sadly limited approach—a misinterpretation influenced by a narrow textualist reading of deconstruction.⁷

Merleau-Ponty, however, offers a counterpoint that lends itself to a deeper understanding of textual agency, even if his perspective sometimes enters a more mystical mode. He writes:

The writer is said [...] to dwell in already elaborated signs and in an already speaking world, and to require nothing more of us than the power to reorganize our significations according to the indications of the signs which he proposes to us. But what if language expresses as much by what is between words as by the words themselves? By that which it does not “say” as by what it “says”? And what if, hidden in empirical language, there is a second-order language in which signs once again lead the vague life of colors, and in which significations never free themselves completely from the intercourse of signs? (Merleau-Ponty 1964, 45; also cited by McCleary 1986, 46–47)

We may well ask what precisely he means by this. At first glance, his statement could easily—and perhaps too easily—be interpreted through the lens of textualist theory as it has been widely adopted in cultural analysis. While he seems unopposed to the textualist notion of deferred meaning, I would argue that the “between words” he speaks of indicates something beyond textual reference and structure, even in the most involved of senses. Textualism, by necessity, overlooks the tacit understandings and knowledge that Merleau-Ponty consistently foregrounds. Indeed, the temptation to interpret this as an embodied “knowing” is strong here as well. Given the literary examples he provides, we might indeed argue this case either way. To fully grasp his position, however, we should return to his concept of gesture and explore how the gesture-like qualities of a “speaking text” (a term I adapt here) generate experientiality by invoking agency.

Experientiality is obviously not a straightforward issue, and I am working here with a distinction between two levels of experientiality that textual representations might provide. First, there is experientiality as it is perhaps more

⁷ While a more moderate understanding of textuality (and thus also intertextuality) can prevent such a full divestiture of significance from texts, the issue of textual intentionality which I see as crucial for understanding the particularity of literary representation remains problematic.

commonly intended when speaking of literary texts. This is where reading a particular text has an emotional impact on us, creating lasting memories associated with the reading experience and the particular text. Re-enactment, empathy, the production of catharsis and the creation of closure and fulfilment all work toward this kind of *emotional experience*. Second, it seems useful to distinguish an experientiality in which we are viscerally *impressed* by the text, not necessarily through the experiences it describes, but through the act of reading itself as a lived experience—the phenomenologists' *vécu*. While the latter understanding may be seen as supplementary, distinguishing between the two permits pinpointing the role of agency in what can be broadly termed “the literary experience.”

It is this second, lived aspect of experientiality that generates understanding in an embodied, internalized way, which I aim to emphasize. This lived experience extends beyond emotional impact to something closer to what Merleau-Ponty calls the “I can” of embodied competence—a feeling of accomplishment or capability akin to skillful action.⁸ Thus, the experience in question is not merely one of heightened awareness or identification, which, after all, can be reached by various quite differing means. While readers may indeed encounter ecstatic intellectual recognition—a joyful “I know!” in response to the familiarity of the contents of a description, the recognition of a sound, taste, or fragrance, for example—the kind of experience I mean is a “now” and an “I feel!” that circumvents linguistic appropriation, operating instead as a direct, embodied understanding. To elucidate this ambiguous point, Merleau-Ponty's (1962, 184) example of gesture provides clarity here again: “The gesture does not make me think of anger, it is anger itself.” Here, the gesture is not simply a sign or a symbol but a component of meaning in itself. As Sue Cataldi (1993, 95) interprets Merleau-Ponty's view, “to ‘blindly apprehend’ an emotional significance is to ‘understand’ it by way of its opening onto the conformatively expressive possibilities of one's own bodily powers.” There exists, then—“between words”—a realm of embodied meanings that enrich the sense of the text, allowing it to communicate on a level beyond the linguistic.

Similarly, drawing on Husserl and Merleau-Ponty, Richard McCleary presents what he terms “carnal self-awareness” as a mode of corporeal reflection, explaining that this self-awareness serves as a medium—if not the ground itself—“through which the lived body acquires practical understanding of itself”

⁸ For his part, Todes (2001) insightfully refers to this embodied confidence as “poise,” which describes the skilled body's capacity for effective action.

(McCleary 1986, 90). This embodied understanding not only frames perception but also enables the abstraction of experience:

Carnal self-awareness and the practical understanding acquired with it are also the original form and continuing basis of imagining. For imagining is a matter of experiencing the here-and-now as expressing, through analogy, the not-here-and-now. And carnal self-awareness of one hand touching the other is a lived bodily experiencing and practical sense of the here-and-now touching of hands as expressing, through carnal analogy, absent, possible, and merely imaginary touchings which are not-here-and-now. (McCleary 1986, 90–91)

Conversely, a “writerly” description of touch can imaginatively *impress* an instance or experience of touching, and a literary description of a physical attitude or a gesture might, through the same mechanism, invoke or (re)create an experience of that attitude or gesture, up to and including the affective and emotional states attached to it—regardless of whether these states are textually made explicit. In this way, the text’s embodied meanings can speak to readers’ corporeal awareness, contributing to a rich, non-verbal understanding that surpasses what linguistic content alone can capture.

5. Excess and Disturbance in Embodied Reading

Although it does not involve the direct manifestation of action, a textual experience can impress upon readers a sense of agency at a fundamental level. Most often, texts seem to accomplish this in terms of confusion or contradiction, in the creation of which not only our cognitive abilities and emotions but also our bodily capacities and sensorimotor experiences are invoked. And this, to me, marks the crucial bridge from phenomenology to poststructuralism.

In this context, I intend the concept of “excess” in two senses: first, to illustrate the nature of literary works as overdetermined (a trait frequently noted in textualist analysis), which renders them incapable of providing room for the articulation of and understanding or response on an embodied or physiological level; and second, suggest their largely untapped potential for containing “too much” in terms of complexity. These two senses intersect in my argument that agentic literary representations, or “writerly” texts, are enriched by a

phenomenological notion of embodied “knowing” that extends beyond intellectualized understanding.

Textual strategies of confusion, contradiction, and ambiguity undeniably invoke agency in the reader by creating moments of tension within the reading experience. But they can only do so because literary reading positions us within a safe, nonjudgmental space, afforded by the literary compact, where we are free to explore disorientation without immediate risk. And, of course, because our attention is fully engaged. Michael Riffaterre articulates this process well: “Reality and the author are either verbally present [...] or deduced from the utterance and reconstituted by the reader [...]. *Contact is assured* not by the kind of passive reception involved in normal communication, but by the active performance (in the musical sense of the term) of the score represented by the text” (1983, 4, emphasis added; compare this with Merleau-Ponty above on communication). In this sense, the reader’s engagement becomes an active “performance,” whereby meaning is created not passively received, embodying an interaction with the text.

True to his (post)structuralist commitments, Riffaterre extends this concept further:

The text is a limiting and prescriptive code. [...] In nonliterary communication, because of the grammatical probabilities that every segment of the utterance permits us to evaluate [...], considerable latitude in decoding is left to the receiver of the message. If the situation were the same in literary communication, the text would not be a monument, for it would be incapable of enduring. Nor could it function as a program or score. It could not force the reader to experience a feeling of disorientation or strangeness. (Riffaterre 1983, 6)

Forcing the reader to experience “disorientation or strangeness,” to me, involves the engagement of bodily attitudes and “knowings.” Certainly, as meaning must be constructed by the reader, the type of agency involved here (along with the kind of experientiality I’ve argued for) is actively summoned by the act of reading. This quality of “writerly” and “poetic” texts—by which I try to include prose that employs modernist literary techniques as well as other avant-garde forms—acts as a cultural counterforce to any form of totalizing interpretation.⁹ Such forms

⁹ As Fludernik reminds: “A great number of 20th-century texts in the first-person format are written in the reflector mode, as Stanzel calls it; they concentrate on the I’s experiences without evaluating

emphasize sensory experiences and fragmentation, rejecting closure and any attempt to fix meaning.

A central factor in this resistance to closure is the recognition that all efforts at understanding are inherently subject to linguistic figuration. On this view, the notion of cognitive distance becomes, more than anything, a myth—useful, perhaps, yet still a myth. Whether one interprets this myth pragmatically, moving to a denial of representation *in toto*, or follows a phenomenological path, these strategies share a common focus: they call for a shift from thinking and conceptualizing to “doing” (to, that is, an interaction with the world) and to experiencing (to the sensual or sensational). And it would seem this is the direction in which avant-garde artistic expression has been taking us since the 1960s at least, with more entertaining, main-stream representations, especially popular fiction and cinema, slowly adopting some of the same means, and even academic cultural analysis making the occasional attempt, particularly evident in the spike in investigations of “embodiment” in the late 1980s and early 1990s.

In line with this broader argument, and channelling yet another phenomenologist, Paul Ricoeur, Monika Fludernik aptly observes that “narrative lives from configuration—what matters is the shape of events and their emotional meaning to the protagonists and the teller. Or, in the frame of my own theses, it is experientiality that counts” (Fludernik 2003, 130). This focus on experientiality becomes especially obvious when literary texts disrupt habitual patterns of reading, provoking readers to engage in a more embodied manner. Offering another illuminating example from her reading of Proust, Elizabeth Deeds Ermarth argues that “[i]t is precisely Proust’s ability to push his language to the point of catastrophe, of coherence lost, that teaches his readers its powers and limits. These threats are the very things that make Proust invigorating to read.” (Ermarth 2003, 406)¹⁰

A quick literary example of this kind of threat and disturbance—a favourite of mine from Nadine Gordimer’s *A Guest of Honour* (1970)—may help here. In a

them from the perspective of a teller who retrospectively views his or her former life and ensures narrative closure.” (Fludernik 2003, 123)

¹⁰ Ermarth (2003, 406) gives another very illuminating example of this: “Something similar can be said of Henry James, the first author writing in English actually to situate subjectivity at the perilous margins where syntax fails. James, especially in his magnificent late work, forces his readers to cruise that boundary. In such texts, it can be argued, subjectivity is most liberated precisely at the point of breakdown. This is not at all a psychoanalytic point; psychoanalysis is not required for a reader to recognize the necessary bond between risk and freedom.”

café scene, protagonist James Bray requests a cup of Turkish coffee, and Gordimer presents us with the following:

The pregnant wife, very young, with wisps of black hair showing in the white arches of her underarms, fetched a cup, unsmiling. In the cupola between the body and the arm he tasted always the sweetly-scentedness of something smeared there to disguise sweat, and the slight gall, as of an orange pip bitten into, of the sweat itself, and his tongue, moving one way smoothly, felt the nap of shaven hair-roots when it moved the other. The coffee was boiling hot, thick and delicious. Rebecca had not been in his mind at all; only this one part of her, suddenly, claimed him with an overwhelming sense of reality of its own. (370–371)

Now, Gordimer here makes use of the sense of taste, along with touch, to convey an experience that transcends visual perception., but the point is the same regarding the way the interruption circumvents our habitual reading responses by causing confusion, by moving from a visually perceived “there-now” to a visceral “here-then-and-now” with no explanation. Interruptions like this derive their power from bypassing, indeed violently disengaging, rationalizing processes and provoking a bodily response. Again, however, this works because the text is situated in a zone of relative comfort and security, against a backdrop of agreed-upon rules and conventions, as well as some more elusive “already known.”

Such disruptions of conventional meaning-making serve, as noted by Riffaterre, as a useful reminder of the limitations of interpretations that attempt to reduce ambiguity or explain away complexity. In Riffaterre’s estimation, “any explanation that, when faced with a group of obscure or ambiguous words, would attempt to reduce the ambiguity or explain away the obscurity must be rejected. This, of course, means that we must give up the acrobatic performances of traditional commentaries.” (Riffaterre 1983, 9–10) While Riffaterre is a textualist through-and-through, he reminds us that this does not mean that he is a relativist. Texts have meaning, and their integrity must be respected. To this end, the existence of ambiguity must be accepted and that is not the forte of a reductive, reflective, and hence effectively disembodied, imagination.

In effect, then, engaging with texts that operate at the boundaries of coherence allows readers to experience language as an embodied encounter rather than a purely cognitive exercise. Literary modernist and experimental techniques,

through their emphasis on sensation, defamiliarization, and ambiguity, prompt an active, participatory form of reading where language functions less as a medium of representation and more as an agent of experience. By prioritizing sensory engagement, these texts highlight the “excess” of literary language, drawing readers into an encounter that is felt as much as it is understood, and that challenges the confines of habitual interpretation. On this view, literary texts are not passive repositories of meaning but dynamic participants in a reciprocal exchange that foregrounds both reader agency and the text’s intentionality. Such texts ultimately transcend the limitations of standard language use, cultivating a relationship that embraces the reader’s active role in constructing meaning, and affirming the potential to evoke an unmediated, bodily dimension of understanding.

University of Turku

References

- Ball, K. (2005), “Toward an Anatomy of Visceral Reason”, *Parallax* 11(3), 1–7.
- Barthes, R. (1974), *S/Z*, trans. R. Miller, Hill and Wang, New York.
- Cataldi, S.L. (1993), *Emotion, Depth, and Flesh: A Study of Sensitive Space: Reflections on Merleau-Ponty’s Philosophy of Embodiment*, SUNY Press, Albany.
- Bleich, D. (2006), “The Materiality of Reading”, *New Literary History* 37(3), 607–629.
- Csordas, T.J. (1999), “Embodiment and Cultural Phenomenology”, in G. Weiss and H.F. Haber (eds.), *Perspectives on Embodiment: The Intersections of Nature and Culture*, Routledge, New York, 143–162.
- Derrida, J. (2005), *On Touching – Jean-Luc Nancy*, trans. C. Irizarry, Stanford University Press, Stanford.
- Ermarth, E.D. (2003), “Beyond ‘The Subject’: Individuality in the Discursive Condition”, *New Literary History* 31, 405–419.
- Fludernik, M. (2003), “Chronology, Time, Tense, and Experientiality in Narrative”, *Language and Literature* 12(2), 117–134.
- Fredriksson, A. (2014), *Vision, Image, Record: A Cultivation of the Visual Field*, Åbo Akademi University Press, Turku.
- Fredriksson, A. (2022), *A Phenomenology of Attention and the Unfamiliar: Encounters with the Unknown*, Palgrave Macmillan, Cham.
- Gallagher, S. (2005), *How the Body Shapes the Mind*, Oxford University Press, Oxford.
- Gordimer, N. (1970), *A Guest of Honour*, Penguin Books, Middlesex.
- Johnson, M.L. (1999), “Embodied Reason”, in G. Weiss and H.F. Haber (eds.), *Perspectives on Embodiment: The Intersections of Nature and Culture*, Routledge, New York, 81–102.
- Lakoff, G., & Johnson, M. (1980). *Metaphors We Live By*. University of Chicago Press, Chicago.

- Lakoff, G., & Johnson, M. (1999). *Philosophy in the Flesh: The Embodied Mind and Its Challenge to Western Thought*. Basic Books, New York.
- Marks, L.U. (2002), *Touch: Sensuous Theory and Multisensory Media*, University of Minnesota Press, Minneapolis.
- McCleary, R.C. (1986), *Imagination's Body*, University Press of America, Lanham.
- Merleau-Ponty, M. (1962), *The Phenomenology of Perception*, trans. C. Smith, Routledge, London.
- Merleau-Ponty, M. (1964), *Signs*, trans. R.C. McCleary, Northwestern University Press, Evanston.
- Morris, D. (2004), *The Sense of Space*, SUNY Press, Albany.
- Riffaterre, M. (1983), *Text Production*, trans. T. Lyons, Columbia University Press, New York.
- Ruthrof, H. (1997), *Semantics and the Body: Meaning from Frege to the Postmodern*, University of Toronto Press, Toronto.
- Sartre, J.-P. (1949), *What is Literature?*, trans. B. Frechtman, Philosophical Library, New York.
- Tambornino, J. (2002), *The Corporeal Turn: Passion, Necessity, Politics*, Rowman & Littlefield Publishers, Lanham.
- Todes, S. (2001), *Body and World*, MIT Press, Cambridge, Massachusetts, and London, England.
- Zahavi, D. (1999), *Self-Awareness and Alterity: A Phenomenological Investigation*, Northwestern University Press, Evanston.
- Zlatev, J. (2023), "The Intertwining of Bodily Experience and Language: The Continued Relevance of Merleau-Ponty", *Histoire Épistémologie Langage* 45(1), 41–63.

Chalmers, Scrutability, and Phenomenal Truths

PANU RAATIKAINEN

I first met Valtteri Arstila at the beginning of the 2000s when I acted as an external examiner for his licentiate thesis at the University of Turku. A bit later, we met and had fun together in New York. Ever since, I have followed the development of his career with respect and delight. Much of Valtteri's research focuses on the experience of time and colors. I don't feel I would have much to offer concerning those topics. However, more generally, Valtteri and I have always shared an interest in phenomenal consciousness. I want to congratulate Valtteri on his special day with this little note related to the latter theme.

+ + +

David Chalmers's *scrutability thesis* plays an increasingly central role in his thought (see esp. Chalmers 2012). This thesis is intended to provide a foundation for his epistemic two-dimensional semantics (see Chalmers 2006, 2012), which in turn plays an important role in Chalmers's argument against physicalism (see esp. Chalmers 2009). His favored version of the scrutability thesis says that *all truths* are *a priori* knowable from certain basic truths.

Chalmers contends that all truths (of, say, everyday knowledge and various special sciences) are scrutable *a priori*, roughly, from physical and phenomenal truths. More exactly, Chalmers adds that the base still needs to be complemented, first with (at least two) indexical truths (e.g., "I am such-and-such," "Now is such-and-such"), and second with a totality or "that's-all" truth (which entails various negative truths, such as "there are no ghosts"). Chalmers thus argues that all ordinary truths are *a priori* entailed by this class, which he abbreviates as *PQTI*

(physical, qualia, that's-all, indexical). *PQTI* is Chalmers's standard or default base (see Chalmers 2012; cf. Chalmers & Jackson 2001).¹

As to *P*, Chalmers writes:

Let *P* be the class of physical truths, including microphysical truths (truths about fundamental physical entities in the *language* of a completed *fundamental physics*) and macrophysical truths (truths about any entities, including macroscopic entities, in the *language* of *classical physics*). (Chalmers 2012, 110; my emphasis)²

Chalmers characterizes *Q* as follows:

Let *Q* be the class of phenomenal (or experiential) truths: truths about *what it is like* to be a given entity. Phenomenal truths will take the form 'There exists an entity with such-and-such phenomenal properties at such-and-such time', where phenomenal properties are properties specifying what it is like to be an entity. Phenomenal properties will be picked out using expressions for pure phenomenal concepts (Chalmers 2003), concepts that pick out phenomenal properties according to their *phenomenal character*: intuitively, a pure phenomenal concept of a phenomenal property is the sort of concept that would be deployed in knowledge of *what it is like to have that property*. (Chalmers 2012, 110–111; my emphasis)

However, in his defense of the scrutability thesis, Chalmers seems to be sliding, in the case of phenomenal truths *Q*, from one thing to another without noticing: On the one hand, according to his "official" definition, they only contain truths about pure phenomenal ("what is it like") experiences, or *qualia*. On the other hand, Chalmers often seems to assume in practice that the phenomenal truths can do all the epistemological work that observational statements do in the philosophy of science; that they cover all possible observational evidence about the world. This seems to give more plausibility to his *a priori* scrutability thesis (see, e.g., Chalmers 2012, 121–123, 126; cf. Chalmers & Jackson 2001, 329).

¹ I have scrutinized Chalmers's thesis critically, from different angles, in an earlier work (Raatikainen 2014; cf. Raatikainen 2021a, 2021b). This short note can be seen as complementing that critical discussion.

² In Chalmers and Jackson 2001, it is assumed that mere microphysical truths suffice.

However, the latter is arguably a much broader class of truths (or perhaps even a distinct class; see below) and is not limited to pure phenomenal experiences (even if some early logical positivists may have assumed that purely phenomenal truths would suffice).

Consider a textbook example of observation in science: determining the melting point of lead. The observational data consist of a series of measurements, namely thermometer readings. This is quite different from pure *qualia*. Or think of Mendel's observations over many years of the properties of peas, such as height, flower color, seed color, and seed shape; Darwin's systematic observations about the size and shape of the beaks of finches on the Galápagos Islands; Morgan's thousands of observations of the various traits (such as the color of the eyes, or the size of the wings) of fruit flies. In economics, observations may consist, for example, of certain numerical values of the gross domestic product or consumer price index. Unlike pure *qualia*, such scientific observations require considerable conceptualization (such observations are "theory-laden," to use the popular slogan).

Furthermore, the phenomenal aspect, the subjective "feel" of such observations, appears to be quite irrelevant. It seems plausible that a philosophical Zombie or a (with respect to phenomenal consciousness) unconscious intelligent robot could well make such observations and record them. If one takes Jackson's famous Knowledge Argument (Jackson 1982)³ seriously, as Chalmers seemingly does, pure phenomenal truths look like a bad candidate for evidence in science: if it is impossible to infer first-person phenomenal ("what is it like") knowledge from objective scientific knowledge, then, conversely, phenomenal knowledge cannot falsify objective scientific theories. There does not seem to be the right sort of inferential relation between these two realms. If this is so, pure phenomenal experiences cannot play the relevant epistemic role in the inference of any conventional intersubjectively communicable objective knowledge.

Some observations in physics can perhaps be classified as macrophysical truths, and hence be covered by Chalmers's *P*. However, it is quite clear that a great many crucial observations in, say, biology, psychology, or economics, which are conceptualized to be evidentially relevant for such special sciences, just cannot.

³ In Jackson's original thought experiment, Mary is locked in a black and white room. Her knowledge before her release focuses on neurophysiology. But quite obviously she can be given any amount of objective scientific knowledge from any discipline without altering the conclusions (this became clear to me in discussion with Tim Crane).

It seems the phenomenal truths and observational truths relevant for various special sciences are quite different classes. The former at least constitute too narrow a class. If my final reflections above are on the right track, they are not even necessary or useful at all for science. Be that as it may, Chalmers's scrutability thesis consequently looks much less plausible.⁴

Tampere University

References

- Chalmers, David (2006), "The Foundations of Two-Dimensional Semantics", in M. Garcia-Carpintero & J. Macia (eds.), *Two-Dimensional Semantics: Foundations and Applications*, Oxford University Press, New York, 55–140.
- Chalmers, David (2009), "The Two-Dimensional Argument Against Materialism", in B. McLaughlin, A. Beckermann & S. Walter (eds.), *The Oxford Handbook of Philosophy of Mind*, Oxford University Press, Oxford: 313–335.
- Chalmers, David (2012), *Constructing the World*, Oxford University Press, Oxford.
- Chalmers, David & Frank Jackson (2001), "Conceptual Analysis and Reductive Explanation", *Philosophical Review* 110, 315–61.
- Jackson, Frank (1982), "Epiphenomenal qualia", *Philosophical Quarterly* 32, 127–136.
- Raatikainen, Panu (2014). "Chalmers' Blueprint of the World", *International Journal of Philosophical Studies* 22, 113–128.
- . (2021a). "Chalmers and Semantics", *Theoria* 87, 1193–1221.
- . (2021b). "Troubles with the Canberra Plan", *Synthese* 199, 4039–4060.

⁴ I would like to thank Jaakko Kuorikoski for the valuable discussions about observation in science, and Carl Hoefer and David Papineau for their useful comments on an earlier draft of this paper.

From Predicting Brains to Normative Authority

JANI SINOKKI

1. Introduction

Predictive processing (PP) is a theory of neural function in cognitive and computational neuroscience. According to it, our brains are prediction machines that seek to match incoming sensory inputs against internally generated, model-based probabilistic predictions of such inputs (Clark 2013; Friston 2010; Hohwy 2013). PP is often viewed as a promising overarching theory of general brain function ‘about how brains come to approximate Bayesian inference in cortical information-processing through *hierarchical predictive coding* and *precision weighted predictions error minimization*’ (D. Williams 2018, 843; Clark 2013; Hohwy 2013). While the theory originally sprung from concerns related to sensory processing and the ‘problem of perception’¹, it is now viewed as offering promising hypotheses about many other mental phenomena as well, including cognition (Fabry 2018), emotions (Wilkinson et al. 2019), infant learning (Köster et al. 2020), depression (Kube et al. 2020), and many more.

One crucial issue in human mental life is the *normativity* that is essential to it. The workings and emergence of social, institutional, linguistic, and moral norms are an essential part of our humanity and constitute much debated topics of research for philosophers, social psychologists, and animal researchers.²

¹ I.e., given that the organism has access only to information stemming from its own sensory transducers, how does the brain infer the objective structure of its environment?

² For influential philosophical views related to norms, see (Wittgenstein 2009; Searle 1969; 2008; Kripke 1980; Putnam 1970; Brandom 1994; 2019). For a useful collection of papers combining philosophy with empirical studies of normativity, see (Bayertz and Roughley 2019). For philosophically interesting discussions of social and moral normativity by a leading primatologist, see (De Waal 2001; 2006). For a philosophically illuminating discussion of issues in empirical studies of social norms in nonhuman animals, see (Kappeler, Fichtel, and van Schaik 2019). For a useful review of how social norms are defined and operationalized in behavioral studies and economics, see (Farrow, Grolleau, and Ibanez 2017). For a sociological perspective on social norms and their emergence, see (Horne 2001).

Normativity, in a minimal sense, is a matter of *evaluability relative to some criterion* (i.e., a norm). However, mere evaluation, or the possibility thereof, is not sufficient for normativity. In living organisms, structures that function as proxies for other (external) states are ubiquitous (D. Williams 2018, 849; Bechtel 2007, chap. 5; D. Williams and Colling 2018). It is also a central tenet of the PP framework that ‘*all optimal systems exploit structures that are isomorphic to the systems they regulate*’ (D. Williams 2018, 849; Friston 2010). Such states or structures thus represent some other state for the organism and act as proxies for those other states and therefore are responsive to evaluation with respect to how well they perform this function. Nevertheless, such states are not *normative states*. They are at best functionally representational, as the level of a thermometer represents the temperature of its environment. Normativity, in contrast, requires that the state is responsive to the criteria of evaluation in the sense of *oughtness* or *obligation*. Failing to conform to a proxy function is just a malfunction, but failing to conform to the rules of chess, grammar, social codes, or demands of morality are violations of a norm and something one *should* not do. As Stephen Darwall puts it, evaluability at stake in normativity “entails that some action, attitude or mental state of some other kind is justified, an action one ought to do or a state one ought to be in” (Darwall 2001). Another way to put it, which is the one on which I focus in this paper, is to say that normativity has *authority* over the agents in question (e.g., Korsgaard 1996).

In this paper, I argue that once it is complemented with an understanding of relevant psychological processes, PP is well suited to explain how this authority of normativity comes about. Other scholars have already made similar suggestions. However, such proposals either confuse responsiveness to criteria with normativity, thus ignoring the crucial dimension of the authoritativeness of normativity altogether (e.g., Piekarski 2019), or leave the details undiscussed, admitting that the task is fraught with difficulties (D. Williams 2018). I won’t discuss these other proposals in this paper but focus on sketching out the details of my solution. Although my exposition showcases the explanatory powers of PP, the importance of such a possibility I am sketching out reaches beyond the PP framework.

Arguably, any naturalistic account of authoritativeness involved in normativity is doomed to fail (e.g., Korsgaard 1996). Therefore, any progress, even modest, in understanding normativity in a naturalistic context marks an important step forward for the naturalistically minded. But the issue is also timely.

Current theories of normativity have surprisingly little to offer in the face of current social crises, such as political polarization, hate crimes, transphobia, mainstreaming of misogyny or climate denialism, or normalization of racism in Western societies (Ebner 2023; M. L. Williams 2021; Saul 2024). Although many are eager to brush off such issues as explicable by social factors, such as lack of education, ignorance, and low socioeconomic status, it has become clear that the issues have less to do with such traits and all to do with the normative attitudes that individuals hold. This is attested by the finding that one of the largest groups of people harboring the most intolerant attitudes in Finland is found among well-off, educated, and healthy young people (Mannerström et al. 2020). While I do not pursue this topic further in this paper, I think that a philosophical account of normativity must be able to address and enable us to explain such phenomena to have *any* plausibility.

Because of my focus on naturalistic, evolution-based explanations, I use the term ‘normative’ in a looser sense than some philosophers do, which serves my purposes. I explain the topic and terminology further in the next section. In section 3, I describe the basics of the PP framework. I omit most of the technical details and focus on the mode of explanation of human behavior provided by PP that is in my interest and then connect PP with the topic of normativity (section 4). To determine how to connect the issue of normativity and PP, we need to understand normative attitudes, as discussed in section 5. In this task, I rely on Edward Westermarck’s psychological, evolutionary-based account of normativity. Once all this is in place, I will explicate the emotional and cognitive mechanism that results in normative attitudes and show how the PP framework provides the final push to see how certain attitudes and social regularities come to look as if they have intrinsic authority over us to obligate some actions and to forbid others.

While my paper is hardly a comprehensive discussion about the matter, it shows that a naturalistic explanation of normativity is indeed possible and plausibly a worthwhile endeavor. In addition, it serves to showcase how we might acquire a holistic understanding of our human nature by combining evolutionary biology, the PP account of the brain, and (social) psychology of emotions and actions and by adding a touch of philosophical analysis to all of this.

2. Authority of Norms

In the ordinary, non-philosophical sense, normativity denotes relation to a *norm*. For such a characterization to be informative, we need to understand what *norms* themselves are: how they exist, how they direct individuals over whom they have authority, and how they come to have such authority. Many norms, however they exist or operate, are clearly contingent in the sense that their function in producing social cohesion or homogeneity is crucial, yet their actual prescriptions (i.e., their contents) are haphazard or at least liable to change. Dress codes and linguistic norms fall into this category. However, linguistic norms serve other purposes in addition, for they are necessary for successful communication (Grice 1975).

Some norms are a matter of life and death. Some sacrifice their lives to oppose what they consider wrong, unjust, or evil. Yet others impose suffering, pain, and death on others for not conforming to norms. Here, the *authority* of normativity becomes salient (e.g., Korsgaard 1996). Sometimes, like not following rules of grammar, breaching a norm is an error (which one should not do because of the norm in place). However, norms sometimes demand much more and can even be against our own personal egocentric interests (such as not stealing even when the risk of being caught is nonexistent). Often, in such situations, we abide by the norms, which shows their efficacy in controlling our actions.

In philosophical discussions, especially in relation to practical reason (e.g., Audi 2001; Broome 2013) and metaethics (e.g., Scanlon 1998; Korsgaard 1996), normativity is used in a more restricted sense, denoting *intrinsic reasons for action*. Whatever is normative gives us reasons to act in certain ways and for not acting in some others. I will not be following this usage of ‘normative’, for such a use implies too rationalistic a view of human agency that connects the authority of norms too closely with the use of reason for my purposes. After all, most norms that have serious authority over us or our fellows arguably are not *rational* at all. In many past and current societies, those not conforming to traditional gender or sexual norms, for example, face a threat of violence or death for their perceived nonconformity to what is taken to be the norm. Rationalist philosophers, thus, would not admit such social norms to be *moral* norms (however, see Westermarck 1906; 1932; Mackie 1977). However, for agents abiding by the norms, there is no difference. The zealots and bigots are motivated to act because they are committed to a norm, just as the person does not steal even when it is against her financial interests. Although I cannot discuss the evidence here, there is no reason to think that there would be significant differences in the

level of brain functions or psychological mechanisms (which are connected, of course) that are involved in the actions of bigots and the virtuous. Both moral and immoral normativity must be based on the same biological mechanisms moving us to action.

In a weak sense, any activity involving rule-following is normative: the rules of chess, say, determine whether one is playing chess. Even here, normativity requires some kind of *understanding* and *intentionality* to act in accordance with a rule: merely pushing the pieces on a chessboard in a fashion that, by coincidence, conforms to actual chess-moves is not yet playing chess. An interesting question even this kind of weak normativity poses is about chess-playing computers.

Arguably, the computer has no intention of following the rules of chess, nor does it understand them; it follows rules simply because it is programmed to move the pieces only in some way. Because of its programming, it does not even have the possibility of breaking those rules. This point is important for the following reasons. In many cases, ordinary normative human action resembles the operation of a computer to a great degree. In any linguistic community conforming to a set of linguistic norms setting standards of ‘correct pronunciation, word-order or conjugation, address, reference or predication, and relevance’, it is the case that ‘the overwhelming majority of the speakers who meet the standards would be unable to formulate most of them’ (Roughley 2019, 30). Most moral judgments ordinary people make in a moral community seem to be like this. Often, the closest a person can come to articulating a moral norm when witnessing, or thinking about, an immoral act of, say, violence or betrayal, is saying that the act *is wrong*. In both cases, of violating rules of grammar or moral code, they seem to state only something to this extent: *you shouldn’t do that*. Thus, it is not clear what kind of understanding is involved here. Habituation to the linguistic or moral norms of the society into which one is born might resemble programming more than many would admit. In addition, while sometimes the agents seem to differ from the computer, at least in that they have the freedom to ignore the rules they are supposed to follow, it is not clear that this is generally true. The authority of norms is sometimes such that people defend their actions by saying that they *have no option*. In addition, returning to bigots and zealots, it is not clear how much real freedom utterly hateful individuals have. The psychological states that lead to hateful violence present the action not only as justified but also as demanded of them. The authority of the norm, the violation of which causes indignation, is witnessed precisely in its emotional force that can, and often does, override reason

and deliberation. If this is true, it is not clear how much ‘free choice’ there is to such hateful enforcement and policing of norms.

Explaining how normative authority comes about in a natural world that itself lacks normativity is a perennial philosophical question in meta-ethics. The traditional answer is that there must be something intrinsically normative, existing independently of humans and their motivations, to account for the commands, demands, obligations, and recommendations that normative things, such as morals, impose on us. The answers range from Plato’s *eternal forms* (Rist 2012) to Kant’s view of *objective rational laws* that *necessitate* rational actions (Ginther 2022) to more recent varieties of objectivism that treat things such as *interests* as *intrinsically normative* and *objective* (e.g., Nagel 1986). The basic assumption shared by such views is that normativity can stem only from something that already is itself normative. If this is so, then all attempts to ground normativity in something natural will either be circular or start an infinite regress (Korsgaard 1996, 30). However, it is not clear that this assumption is true.

In more recent discussions in analytic philosophy, the *collective recognition* of norms has been considered the source of normativity. One influential line of thought since Wittgenstein (2009) has been that what institute norms are our *practices* (e.g., Sellars 1954; Harman 1977; Peregrin 2021). For Robert Brandom, what specifically in such practices institutes their normative force – authority – are ‘acts of *assessment*, our attributions of normative significance, or simply, our *normative attitudes*’ (Giromini 2023, 1280; Brandom 1994, 32–34). For John Searle, the distinctive feature of social entities is what he calls their *status functions* (Searle 2010; 2019). The socially given status of certain physical objects, such as seashells or pieces of paper, as *currency* stems from our collective representations of those objects as currency. These representations, in turn, count as status function declarations “x counts as y”, which endow the relevant objects (x: e.g., seashells) with deontological traits (y: acceptance as currency) (Searle 2010, 13). Although the views of both aforementioned thinkers are very different, they both accept that, in the end, normativity stems from *collective recognition* (Searle 2010, 58) or *reciprocal recognition* (Brandom 2014; 2019) of norms. Only when a collective comes to attribute collectively normative status to something, does it come to have normative authority. Here, the weight of the argument is on the necessity of collective action. Mere individual attitudes, even if they, too, impose demands on

others, are not alone sufficient to explain how full-blown normativity and its authority arises (Giromini 2023, 1281), or so the thought goes.

However, there are several reasons for thinking that large-scale reciprocity is not required for normativity. First, the claim to which philosophers from Wittgenstein onward have been committed is that social interaction and cooperation are necessary for the emergence of social norms.³ However, the recognition of, and responsiveness to, certain criteria, be they linguistic conventions or thought-independent makers of truth and falsity, should not be confused with the question about authority. The existence of social norms obviously depends on the social—for instance, rules of correct use of grammar can emerge only in a community of language-users. The genealogy of such behavioral patterns or conventions, thus, is necessarily tied with the evolution of human sociality. But the question is not about the existence of certain behavioral patterns, conventions, or complex forms of reflective thinking about those patterns and conventions, which plausibly emerge only in sufficiently complex social interactions. The question is about the *deonticity* that those conventions and patterns exhibit. How and why do people view themselves and others as bound by such rules and norms and obey them, even to their deaths? These two issues, though related, are distinct. An explanation to the first does not necessarily amount to an explanation of the latter.

Therefore, we should not abandon the project of explaining normativity too quickly on the basis of individuals' states even if the behavior we are interested in exhibited only in collectives. There are some considerations that help us see why this is important.

According to empirical studies, people have many descriptive beliefs about how others behave (i.e., they *expect* certain behaviors) and about the attitudes of others toward certain behaviors (such beliefs are explicitly appealed to by Searle in his exposition of status functions). Interestingly, even if people are very much motivated by such beliefs about others, the beliefs need not accurately reflect the actual states of affairs (Farrow, Grolleau, and Ibanez 2017). This point comes to the fore in studies in behavioral economics, where it has become evident that influencing the behavior of people, say, increasing their recycling of waste, is surprisingly easy when they come to accept new

³ The stronger claim involved in the views is that also the *intentionality* of thought can emerge *only* through social interactions. While also relevant for my topic, I will have to ignore this issue due to limitations of space.

information challenging their prior beliefs about the behavior of others (e.g., Czajkowski, Zagórska, and Hanley 2019). As truth is not necessary for producing norm conformity (i.e., people can conform even to norms they are mistaken about), it is reasonable to think that what individuals believe individually is crucial to the authority of norms as well.

In addition, in animal studies, it is common to emphasize that many complex social patterns can emerge merely because of simpler, dyadic relations. For example, dominance hierarchies in primates are considered automatic byproducts of dyadic relations, which require only basic capacities of recognizing individuals and remembering previous interactions with an individual (Kappeler, Fichtel, and van Schaik 2019, 76). While the dominance hierarchy of a primate group has the appearance of normativity (e.g., those violating the social order are punished, even by uninvolved third parties; *ibid.*, 75), the complex social structure can result from simpler actions and attitudes that govern individual-to-individual relations. To be sure, the degree to which such noncollective behaviors and beliefs can also explain human collective behavior remains an open question. However, together, these two considerations show that searching for the authority of normativity from individuals and their personal attitudes toward others, instead of a platonic realm of intrinsically normative beings or even of collective sociality, might not be as hopeless as many have thought.

3. Predictive Processing

Predictive processing is a view in theoretical neuroscience about the biological neural functioning of the brain (Friston 2009; Clark 2013; Hohwy 2013; D. Williams 2018). It aims to offer an empirically plausible answer to the question of how brains implement computational processes that result in highly complex behavior and conscious, conceptual thought and has done so with remarkable success (Parr, Pezzulo, and Friston 2022). The main problem all neuroscientific theories face is the astonishing efficiency of our brains. In relation to the modest amount and speed of signals passing through our neurons, the output—our thinking and social behavior—is astonishing. Predictive processing solves this conundrum neatly.

According to PP, the neural signals forwarded from one brain area to another do not encode all the information seemingly at play but merely consist of a signal carrying information about ‘prediction error’. PP views the brain as a

prediction engine that aims to minimize its own prediction errors. Thus, the states of the brain are essentially predictive models of the body's internal and external environment. The signals traveling through neurons from the cortical area to another are related to the accuracy of that predictive model, which is updated according to the arriving signals. The higher areas of the cortex send information about their current predictions back to lower areas (those closer to the sensory input, e.g., sensory organs). As a result, those lower areas need only to inform the higher areas about the accuracy of their prediction concerning that lower state, and the neuronal signal need not carry other kinds of informational content or data (Pendl et al. 2017; Hohwy 2017). In this process, called *predictive coding*, all the information passed in the stages of information processing is relevant only for calculating the related prediction errors.⁴ This mode of operation explains how the relatively small amount of neural activity can produce astonishingly complex human behavior.

The aim of brain function is to reduce prediction error in the service of *homeostasis* (regulation of the stability of the internal processes of an organism) and *allostasis* (regulation of an organism's energy use according to environmental demands). The rationale for this description is called the 'free-energy principle' (Friston 2009), which relates the idea of prediction error reduction to a more general aspect of organisms. The principle states that 'any self-organizing system that is at equilibrium with its environment must minimize its free energy' (Friston 2010, 1). As Friston, the original formulator of the principle (see Friston, Kilner, and Harrison 2006) explains, 'the defining characteristic of biological systems is that they maintain their states and form in the face of a constantly changing environment', which entails that 'biological systems somehow manage to violate the fluctuation theorem, which generalizes the second law of thermodynamics' (Friston 2010, 1). While many other views view the regulation of the internal environment of an organism as the main business of the brain as well, the free-energy principle provides an 'information-theoretic interpretation both of what

⁴ The basic view of the brain's activity in producing perception and action, according to PP, is that brains implement probabilistic models to anticipate perceptions and the results of actions. In the more recent versions of PP, the brain implements weighted Bayesian calculations, which weightings explain differences in reactions to prediction errors. If the uncertainty involved in a prediction is great or the prediction is noisy, the significance of a prediction error signal is lowered accordingly. Much of the detail that makes PP an attractive theory in neuroscience is highly technical and fall beyond the scope of this paper (for details, consult, e.g., Parr, Pezzulo, and Friston 2022; for a review of current state of the art and the historical origins of PP, see Pezzulo, Parr, and Friston 2024).

this homeostatic process amounts to and how it can be achieved in biological agents via the minimization of a quantity to which they have internal access' (D. Williams 2018, 846). In the process of minimizing free energy, the brain uses Bayesian updating to change its predictions and uses actions, such as moving, to change its environment to match the prediction (D. Williams 2018, 842–44). Here, what is called *active inference* comes to the fore: as the brain constantly updates its generative model of anticipated sensory input, it has two ways of reacting to prediction errors. The brain can either update the prediction model so that it corresponds to the unexpected input (usually said to amount to updating beliefs⁵) or it can change the incoming signal so that it matches the predictions (i.e., by moving the body, amounting to initiating action) (Pezzulo, Parr, and Friston 2024, 3).

Although there is an analogy between more traditional accounts of the mind and PP's in terms of the *cognitive* vs. *conative* distinction, the resemblance is superficial. The traditional views usually treat cognition and conation as logically distinct, which results in several serious conceptual and empirical problems (see Helm 2001; Barrett 2017). In contrast, PP is not committed to such a distinction but in name, as the process of updating the predictions or initiating actions are two sides of the same process. In fact, some argue that one of the main strengths of PP is that it can explain how our perceptions and mental states comprise cognitive and conative, as well as emotional, aspects simultaneously in the same neuronal pathways, instead of requiring distinct processes for coding the different kinds of information, which matches the current scientific understanding of the brain (Wilkinson et al. 2019, 104). The insight here is that the inferences the brain makes are multilevel and incorporate interoceptive, proprioceptive, and exteroceptive sensory signals with top-down predictions related to an organism's ability to cope in its environment. This amalgamation explains why ordinary human experience constitutes something more like a mesh of information about the environment (what is going on), emotional evaluation (felt significance or value to the agent), and cues about possible actions (motivation and intention) than a set of neatly compartmentalized processes that traditional philosophical views often postulate.

⁵ It should be noted that in the context of PP 'belief' usually, but not always, is used in the technical sense denoting probability distributions over hidden variables, which should not be confused with the epistemic, philosophical sense of 'belief'.

4. Predictive Processing and Normativity

It has been argued that there is an intrinsic tension between PP and the theories of normativity that emphasize sociality (D. Williams 2018). One issue here is that PP is a thoroughly internalist view of human mentality, whereas social accounts are patently externalist modes of explanation. As we observed above, this tension might be based on flawed assumptions about the role of the social. To the best of our knowledge, social or collective actions might be the result of individualistic actions and states. Therefore, the distinction between ‘internalist’ and ‘externalist’ explanations should not be exaggerated, and both sides must be taken into account.

According to PP, the main function of the brain is to enable the organism to maintain its stability in a changing environment, which requires ‘successfully representing’⁶ the external environment in internal processes. Therefore, the processes described by PP are bound by success conditions relative to this aim. Here, we can see a problem that should be recognized: that the predictions of the brain are responsive to certain success conditions do not make them normative in the sense that they involve the authority that characterizes normativity (*pace* Piekarski 2019). As has been noted, the kinds of epistemic states in which philosophers are often interested are very different in that they are bound by distinct further conditions in addition to those related to their success in performing their function (D. Williams 2018). For example, the state of belief is responsive to rational reasons and evidence, and it has conditions of truth related to the things it is about. In contrast, the capacity of tracking one’s environment successfully is a matter of pragmatic success, which is not liable to evaluation by such additional criteria. Whether the states involved are true to the facts does not matter if they get the job done.

The liability to such additional evaluative constraints pertaining to many of our mental states is what motivates viewing the social domain as explanatorily relevant. On this point, they seem to be in the right: for the emergence of such additional evaluative criteria, such as rational warrant or truth, collective action seems indispensable. It is likely that at least human considerations related to such criteria emerged only in a social setting. However, as noted above, the emergence of a capacity to use such additional evaluative criteria, be it due to social

⁶ I use scare quotes to underline that the representation is used here in the sense of ‘proxy’, not in the sense of truth-evaluable state.

interaction or whatever, should not be confused with the question about their authority.

In its basics, PP offers a very intuitive explanation of human social behavior. As the function of the predictions the brain makes is to ensure the survival of the organism and its reproductive success, the view is highly consonant with current evolutionary explanations of our sociality. The evolutionary advantages of sociality for gregarious animals are well known and include ‘reduced vulnerability to predators, greater access to food resources, and protection from harassment’ (Silk 2014, 137). Moreover, ‘social ties seem to help animals cope with various sources of social stress, perhaps because their relationships provide a greater sense of predictability and control, and this in turn affects health outcomes’ (Ibid.). Furthermore, social bonds provide security, which consists of ‘the predictability of repeated interactions between partners’ (Kappeler, Fichtel, and van Schaik 2019, 73). As is clear on the basis of such statements referring explicitly to *predictability*, sociality itself can also be seen as providing evolutionary advantage in terms of enabling brains to predict the environment in a more stable fashion. Sociality itself provides predictability and safety so that evolutionary explanations and PP clearly converge on this point. Here, PP also offers us a way forward in seeking an explanation of the authority of normativity on the basis of individual brains.

5. Normative Attitudes

Arguably, social norms presuppose that individuals involved in their emergence, upholding, and policing have *normative attitudes*—attitudes that themselves have ‘ought-like’ content (Giromini 2023). Merely collectively held descriptive beliefs won’t do, for at best these could explain certain kinds of collective behavior, but not the behavioral or emotional responses to perceived or anticipated violations of the supposed norms. Normative attitudes, in contrast, are viewed as ‘attributions of normative statuses or qualities such as correctness, incorrectness, entitlements, commitments, etc.’ (Giromini 2023, 1271). That is, they attribute certain additional criteria of evaluation.

Many norms are special in that their violations are often met with general disapproval and hostile emotions, punishments, and third-party interventions (policing) by uninvolved individuals. Enforcing or policing the norm is not usually done out of personal interests, and some argue that this feature distinguishes social norms from mere conventions (Bicchieri, Muldoon, and Sontuoso 2023). In ‘the

case of conventions, there is a strong overlap between the individual's self-interest and the interests of the community that support the convention, whereas compliance with social norms is almost never in the immediate interest of the individual who has to conform' (Kappeler, Fichtel, and van Schaik 2019, 70). Indeed, the authority of the norms is exhibited in the fact that the norms are also upheld in situations where no personal gain is involved, and usually even where upholding the norm is directly against the personal, egocentric interests of the upholder (e.g., not stealing even when the risk of getting caught is low).

Giromini has analyzed the form that normative attitudes of individuals must have for them to function as grounds of collective normative attitudes: they must be *impersonal* and *projective* (Giromini 2023). 'Impersonality' refers to the feature that normative attitudes held by individuals are not merely personal—upholding the norm is not done out of personal interest; rather, the norm applies to everyone regardless of their personal traits and interests. 'Projection' relates to the domain of normative attitudes: normative attitudes are never merely about the act directly addressed, but they 'seek to promote or obstruct the continuity of *actual* acts, the repetition of *past* acts, and the actualization of *possible* acts' (Giromini 2023, 1278).

Finnish philosopher, sociologist, and social anthropologist Edward Westermarck (1862–1939) proposed a hundred years ago an account of human psychology and motivation that explains how our attitudes come to have the features of impersonality and projectivity. I have elsewhere (Sinokki 2024) discussed the details of his naturalistic, evolutionist account of morality and argued that while Westermarck's focus is on morality, his view successfully delineates the more general domain of normativity (in its different senses). I refer the reader back to that discussion for details and focus here only on the issues directly relevant to the issue at hand.

Westermarck bases his account of normativity on the idea of *objectivization*. This idea is not novel with Westermarck, although his use of it in explaining normativity is very original.⁷ According to him, the human mind has a definite tendency to confuse our experiential, subjective states with the physical, objective properties of the objects causing our experiences. We attribute colors, sounds, warmth, and pains to objects, even if those precise qualities are rather our

⁷ Already Hume stated it to be a 'common observation, that the mind has a great propensity to spread itself on external objects, and to conjoin with them any internal impressions, which they occasion' (Hume 2007, 112; 1.3.14).

human ways of experiencing the physical objects. It is our ingrained tendency to confuse our experiences with the dispositions in objects to cause such experiences, which Westermarck calls objectivization. Objectivization is thus a name for the inherent tendency of the human mind to project its features onto objects, as if they were objective, mind-independent properties of those objects.

Westermarck's important insight is that we do not perform this projection only in the case of sensation but also in the case of our emotions. When we disapprove of something, a similar objectivization occurs as with experienced colors, and we project the experience onto the objects as if it was an objective and mind-independent property:

By calling an object warm or pleasant, a person asserts that it is apt to produce in him a sensation of heat or a feeling of pleasure. Similarly, to name an act good or bad, ultimately implies that it is apt to give rise to an emotion of approval or disapproval in him who pronounces the judgment. Whilst not affirming the actual existence of any specific emotion in the mind of the person judging or of anybody else, the predicate of a moral judgment attributes to the subject a tendency to arouse an emotion. (Westermarck 1906, 4.)

The final sentence is crucial because it makes clear that Westermarck's view is not simple emotivism (contrary to what many have mistakenly thought; see Stroup 1982, 199ff; 1984).⁸ Westermarck is in fact very clear that our normative terms, such as 'good' and 'bad', do not refer to our emotions but rather precisely to the features of the objects that dispose them to cause certain emotions. That is, our language use follows mistaken objectivizations. This, as a sidenote, is the mistake Westermarck accuses normative theories of ethics of making: they take our normative language use at face value and mistakenly see normative terms as referring to objective normative properties. However, because that language reflects our erroneous objectivizations, all normative ethics is one enormous mistake. While this verdict might seem harsh to contemporary readers, it should be remembered that Westermarck worked in times before the discussions about rule-following by Wittgenstein (1963/2009), the reinvigoration of normative

⁸ As it happens, Westermarck is likely the first to present an *Error Theory* of morality, later developed and made famous by J. L. Mackie (Pietarinen 2015, 158–59; Mackie 1977). Mackie, of course, was an eager student of Westermarck's views (see, e.g., Mackie 1982).

political theorizing by Rawls (1971/2006), or anti-realist theories and interpretations in normative ethics (see Launis 2018). Nevertheless, Westermarck's point about objectivization seems accurate both in the case of sensory experiences and emotions.

The objectivization and predictive nature of the brain states that PP postulates go well together. In fact, they receive mutual support from each other. If we wonder why we so peculiarly objectivize our states onto external things when it is so clearly an error on our part, understanding that the business of the brain is not to truthfully represent our environments but to help us manage in those environments is helpful. The many additional criteria of evaluation, such as truth, have not been important for human evolution except insofar as they promote the pragmatic aim of surviving and getting by. However, there is also a deeper, more important connection: with PP, we can see that we project our states onto objects because those states are essentially predictions about the objects. Dealing with our physical environment is made simpler by using proxies such as colors, which represent only some, but not all, differences in physical surfaces. The differences they represent, of course, are such that they are important to our allostasis, but maybe not so much for understanding the physical or metaphysical nature of the surfaces.⁹

As noted above, according to PP, the function of our sensory system is *allostasis*, the physiological mechanism of internal regulation in which an organism anticipates and adjusts its energy use according to environmental demands. As many have emphasized, this purpose leads to the separation of truth-functionality and the evolutionary purpose of brain functions:

A brain did not evolve for rationality... or accurate perception. All brains accomplish the same core task: to efficiently ensure resources for physiological systems within an animal's body... so that an animal can grow, survive and reproduce. (Barrett 2017, 3; cited in D. Williams 2018.)

All functions of the brain are 'answerable to this pragmatic end' (D. Williams 2018, 847). Objectivization can be seen to stem from the brain's predictive processes because our perceptions are similar to lossy compression of information about our environments, restricted to only what matters for allostasis. In this sense,

⁹ As Descartes already pointed out, discovering the nature of things is a task to which our embodied minds are not well-suited (Descartes 1991, 226–29 To Princess Elizabeth, 28 June 1643).

the term ‘projection’ might be slightly off the mark. We do not, strictly speaking, project anything onto the objects, but we just come to see those things as colored and painful. Why? Because that is the way our brains deal with those things, and it does so highly efficiently.¹⁰ Objectivization, on the other hand, names the mistake of taking experiential qualities at face value as if they were objective, mind-independent properties, and seems accurate.

The objectivization of sensory states fits well with the explanatory framework of PP, but what about emotions? It has been argued that PP is well suited to account for emotions precisely because it does not view them as distinct from cognitions or conative states; these are all part of the predictive model of the brain incorporating information from inside and outside the body (Wilkinson et al. 2019).

Westermarck believes that our emotions have evolved from basic stimulus reflexes such that most animals, even the simplest ones, share pain- and pleasure-related mechanisms that direct their behavior and promote their survival (Westermarck 1906, 22–23). This is in line with current knowledge, according to which the basic emotional brain structures are of ancient origin and thus present in many animal species, and not only mammals (Nummenmaa 2010). Of course, human emotions have evolved in distinctive ways along with the rest of the human sociobiological nature. Westermarck thinks that as our social brain has evolved, basic emotions have received new functions (plausibly through *exaptation*, though he is not clear on this point) to meet the adaptive challenges brought about by a new, complex social environment (Westermarck 1906, 93–99).

This is an interesting idea worth highlighting: many basic emotions have a primitive function, which is similar in distinct species and promotes survival. In this sense, Barrett is correct in opposing many traditional views in defense of her PP view of emotions: ‘Emotions are constructions of the world, not reactions to it’ (Barrett 2017, 16). In relation to objectification, basic emotions are similar to sensations in that they are our way of dealing with the environment; they signify danger, utility and much more, and are an essential part of how we manage our way with things. However, in the naturalistic stance, our meals lack qualities such as deliciousness, the barking dog lacks the quality of fearsomeness, and the bodies we lust for lack properties of ‘hotness’. The objectivization seems to occur even

¹⁰ This is not intended as a metaphysical explanation about the ‘queerness’ of phenomenal qualities, but only as a functional explanation of colors in the process of objectivization I am discussing.

in the case of such emotions, and we go on to describe the experiential qualities as if they were external properties of things.

Crucially, in the human social sphere, old basic emotions come to take on new roles. For example, basic disgust is caused mainly by contaminants, such as rotten food, secreta, and excreta. However, in the social domain, disgust is elicited by very different things, including ‘treachery and betrayal; laziness, idleness and obsessions of different kinds; weaklings, traitors and devious criminals’ (Heinämaa 2020, 381). Also, deviations from traditional sexual norms can arouse disgust (Nussbaum 2019, 112, 128–30). It seems useful to distinguish basic, non-species-specific emotions from social emotions that we encounter only in humans and possibly in some other gregarious animals (see De Waal 2006). In the case of social emotions, such as approval and disapproval, the environment they predict is essentially the *social environment*. The mechanisms we above observed as characterizing the objectivization of sensations and basic emotions also seem to apply here. The function of social emotions is to represent our social environment, and the emotional agitations we feel are related to the prediction error signals involved in the same way as with basic emotions (Barrett 2017; Wilkinson et al. 2019). However, to see how all this connects to normativity and its authority, we need to take yet a closer look at the social emotions themselves. Westermarck is also useful here.

6. From Objectivization to Authority

Westermarck’s view of social emotions is highly original in that he postulates that, owing to evolution, humans and other social animals exhibit special emotions that he calls *retributive emotions* (Westermarck 1906, chap. 2; 1932, chap. 3; Sinokki 2024). These emotions are intrinsically action prompts to reciprocate actions directed at us. More precisely, retributive emotions prompt us to respond to those actions by paying them back in kind, both positively and negatively. This idea of emotions as action-prompts aligns well with the basic tenet of PP, according to which emotions are not distinct from cognitions or conations but are parts of the brain’s general constructions of the world in service of allostasis. Furthermore, this aspect of Westermarck’s view is still useful, for example, for primatologists: ‘When I watch primates, measuring how they share food in return for grooming, comfort victims of aggression, or wait for the right opportunity to get even with a rival, I see very much the same emotional impulses that Westermarck analysed’ (De Waal 2001, 350).

Westermarck's account contains several details about how social emotions come about, and I can only briefly explain them here. The bottom of all social emotions in all gregarious animals, according to Westermarck, is *affection*. This emotion is the glue that bonds parents to their offspring and family members to each other and operates in all social settings (Westermarck 1906, 109–11). Affection is the basis of *experienced value*: things matter to us because we feel affection for them. Westermarck is clear that he does not view affection in deterministic terms. Affection is not merely something that happens to us but “we take an active part in its production” (Westermarck 1906, 110), which explains the differences between subjects, say, in empathy toward others. Empathy, and affection in general, is shaped by our cognitions—beliefs, worldviews, ideologies—as well. It is not merely reactive.

Affection is also the evolutionary drive of sociality and related emotions: ‘Among social animals whose gregarious instinct has developed into social affection, sympathetic resentment is felt towards the enemy of any member of the group’ (Westermarck 1906, 112). By ‘sympathetic’, Westermarck means emotions that are not self-focused; affection to others makes us care for their well-being so that their pains come to pain us, and their happiness is a source of joy to us. Here, we arrive at what Westermarck calls *disinterestedness*. Some of our emotions and attitudes are disinterested, meaning that their center of significance is not directly our own wellbeing and interests but rather that of someone else.¹¹ Such feelings mark the emergence of proper social emotions. We form social bonds because of affection, but proper social interactions require the ability to make others' interests one's own. With the help of disinterested emotions, we can navigate complex social environments because we can, as it were, share feelings of others—to be pained by blows directed to our comrades, to partake in the joys of our loved ones. These social bonds are precisely those that biologists describe as providing the adaptive advantage mentioned above, which is related to the enhanced predictability of our environments. Moreover, we come to experience *disinterested retributive emotions*: when witnessing that our kin or friends are

¹¹ Here Westermarck explicitly challenges the sentimentalist tradition, which sees all altruism as egoistic at bottom: the good Samaritan helps the bleeding because he wants to remove his own pain, caused by witnessing the suffering. Westermarck, in contrast, thinks that affection must be the primitive ingredient here. We are pained by the suffering of others because their well-being already matters to us, not the other way around, as sentimentalists suggest (Westermarck 1906, 111).

treated badly, we feel indignation just as if the violation would have been directed against ourselves and are prompted to act to pay back the harm to the perpetrator.

When we combine the ideas of objectivization and disinterested emotions (and our cognitive abilities), we can see how something that Westermarck labels *impartiality* arises. When we objectivize some actions that arouse indignation, we treat them as having those qualities independently of our emotions. This objectivization is further amplified because, in many cases, those acts that cause indignation in us are not directed toward us but rather at our fellows. When the objectivization is also generalized (a cognitive process) properly as if a feature in the act, we come to apply normative terminology to it: the act that is disposed to cause indignation in such ways comes to be called ‘bad’ or ‘wrong’. We come to view the ‘badness’ in the act as an independent property of the act (which mistake, according to Westermarck, fuels all normative ethics).

In connection with impartiality, it should be noted that impartiality need not be actual or real. The mere appearance of impartiality is sufficient: “it is not necessary that the moral emotion should be really impartial. It is sufficient that it is tacitly assumed to be so, nay, even that it is not knowingly partial” (Westermarck 1906, 104). The role that objectivization plays in our experience of such emotions explains why our partiality and self-interest, which can be quite significant, often go unnoticed by us. Notably, in connection with sympathetic retributive emotions, we come to view those emotions as indicators of objective matters. They are felt impartially as if there was a natural law that poised us to respond to certain kinds of actions in a certain fashion emotionally and behaviorally. This is objectivization in action. Westermarck’s metaethical insight is that when we make generalizations over such impartially felt emotions, we arrive at normative concepts (Westermarck 1906, 4–5). The actions that cause us to feel impartially felt indignation become conceptualized as ‘bad’ and ‘wrong’, and those doing the actions are conceptualized as acting contrary to ‘obligation’ or ‘duty’. The sense of oughtness we attribute to behavior stems from the complex process involving the objectivization of retributive emotions, leading to our experience of them disinterestedly and impartially and from our cognitive generalizing capacities. This is Westermarck’s account of normative attitudes and normative concepts.

However, social interactions are necessary for *social norms* to emerge, according to Westermarck: ‘Society is the birthplace of the moral consciousness; ... the first moral judgments expressed, not the private emotions of isolated

individuals, but emotions which were felt by the society at large; ... tribal custom was the earliest rule of duty' (Westermarck 1906, 117–18). The objectivized, impartial emotions and concepts need to be shared for social norms to arise. Here, Westermarck is with the tradition of emphasizing collective action in the creation of social norms. The widely held social patterns and regularities we usually call social norms, which make some behaviors obligatory, some forbidden, stem from the collective sharing of attitudes. As we all learn the correct use of normative concepts as well as the norms of accepted behaviors of our societies from others, the patterns can become stable and enduring. To an observer, it might even appear as if there was something over and above the individuals acting in concert that determines their actions or imposes its authority over them—like a *moral law*. Nevertheless, for Westermarck, the proper mechanisms that explain how the norms in a society come to have authority—how they are emotionally experienced as authoritative—boil down to the above-described psychological mechanisms shared by (nearly) all members of human species because of evolution. While society gives rise to social rules and is necessary in understanding them, it is not what ultimately explains their emergence or their authority.¹²

Now, we can finally see how the authority of social norms, or the appearance thereof, comes about.¹³ First, objectivization creates the appearance that the norms are something external to us. True enough, conventions and regularities, which we internalize or copy from others, are real and external. However, power that obliges conformity (ignoring external powers, such as coercion and the threat of punishment, of course) stems from within. It is the emotions that we objectivize as marks of the 'obligatoriness' of norms. However, here, Westermarck needs some help from PP.

According to some defenders of PP, the *emotional valence* (painfulness vs. pleasurableness of the state) is related to the level of our prediction errors and the level of reducing those prediction errors (Van de Cruys and Wagemans 2011). Pleasure occurs when our actual rate of error reduction is higher than we predicted it would be, and a painful state occurs when the rate is lower than expected (Wilkinson et al. 2019, 109). Now, the conventions and regularities are part of the

¹² The explanatory role of the social was a point of contention between Westermarck and Emile Durkheim and his followers. The latter believed that social phenomena should be explained by reference only to other social phenomena, while Westermarck defended a multidisciplinary, multilevel approach relying on biology, psychology and sociology (Pipatti 2024, 22).

¹³ It would not be off the mark to characterize this account as an error theory of normative authority. See footnote 8 above.

environment in which we operate and predict. Observing them being violated produces a prediction error signal.

Consider a simple case of a (repeated) violation of some simple linguistic rule. When we perceive such a violation, our brains immediately seek an explanation. The violation might turn out to be a pun or joke, in which case we feel amused—the prediction error vanishes because of the realization that the deviation was intentional and meant as humor. However, if we find no clear reason for the deviation, the brain continues to resolve the error, and felt irritation about the matter lingers on. We might, for example, become excessively conscious of the error so that we cannot even properly focus on anything else that is said. The prediction error is not reduced as fast as expected, which results in an unpleasant emotion.

It could well be that one of the slowest rates of error reduction occurs when the violation is indeed intentional, but the reason for the violation is something we cannot understand or something we find unacceptable. The irritation or indignation a judgmental (elderly) person might feel when listening to youth slang comes to mind as a linguistic example. In accordance with active inference, the options are changing the prediction or trying to change the ways the youth speak. However, the prediction error does not go away easily because the prediction is deeply rooted in the form of internalized linguistic regularity, and the prospects of getting the youth to conform are likely dim. (Of course, those with more agile minds *can* change the prediction, e.g., by simply adjusting it relative to different age groups of speakers.) Importantly, the same considerations seem to apply to all social regularities, be they rules, conventions, or norms.

When our predictions go amiss in the case of conventions or norms, the predictions in question stem from a regularity we have internalized. Such regularities are, in a way, central building blocks of our whole predictive model—fundamental assumptions about the way the world is and in which it operates. The option of correcting the prediction is thus always quite costly. If the prediction error cannot be corrected easily by actions, for example, when cooperation of unwilling others who disagree with us over the norm is needed, the reduction rate of that error signal is bound to be very low. It causes acute emotional experiences with negative valence—pain. If it is true, as I just speculated, that prediction errors related to norms are especially prone to exhibit very low reduction rates of error prediction, then violations of internalized norms are among those that are likely to produce the most violent negative emotions and to last for long periods. Although

I cannot discuss the matter here, I think that there might be a highly important connection between norms and *hate* which must be studied further.

If the above discussion is on the right track, then we can explain why violations of norms can cause highly negative, acute emotions, such as irritation, indignation, and outbursts of anger. Here, we might consider those getting quite upset about violations of grammar, about the toilet roll being put ‘in the wrong way’, or about other people driving in the ways they are not supposed to drive (i.e., road rage). This PP-derived explanation of the emotional responses to violations of regularities, be they conventions, customs, norms, or whatever, fits nicely with the above-described psychological mechanism of objectivization. Because of the objectivization, we come to see the force of the strong emotions that impel us to action—from correcting the grammar of the youth to lynching and stoning the norm-violators—as something outside us, making demands on us. We come to see it as *authority* possessed by the norm itself.

7. Concluding Remarks

If the line of reasoning above is roughly correct, then we need not postulate an objective realm of normative beings to understand how normativity comes to have its authority. Additionally, contrary to what some have suggested, the suggested explanation for the power norms have does not stem from collective attitudes. Rather, collective attitudes can merely aggregate and amplify what already exists at the level of individuals. While my discussion is more a sketch than a full story, I think that what I have said suffices to show that there are prospects for explaining normativity and its authority in naturalistic terms and reducing the force of normativity to the level of individual minds. I think that the sketch also speaks favorably of the theories here discussed—Predictive Processing and Westermarckian account of normativity—but also shows that overarching grand explanations uniting biological, psychological, and normative explanations are possible, even in a naturalistic framework. Finally, the account of the authority of normativity presented here is patently *subjectivist*. This, however, does not mean that the force of such authority would be merely imaginary. We cannot escape it anymore than we can make colors cease to exist with the realization that they are parts of our experiences rather than physical qualities. The authoritativeness of norms and our normative attitudes are very real and have real effects in our societies and to others (including minorities), especially when aggregated into collective attitudes. With the above account, we at least can start to see why

normativity has the power to move us, and why in so many cases that power has very little to do with rational reasons.

University of Turku

References

- Audi, R. (2001), *The Architecture of Reason: The Structure and Substance of Rationality*, Oxford University Press.
- Barrett, L. F. (2017), “The theory of constructed emotion: An active inference account of interoception and categorization”, *Social Cognitive and Affective Neuroscience*, 12(1), 1–23. <https://doi.org/10.1093/scan/nsw154>
- Bayertz, K., and Roughley, N. (2019), *The Normative Animal?: On the Anthropological Significance of Social, Moral and Linguistic Norms*, Oxford University Press.
- Bechtel, W. (2007), *Mental Mechanisms: Philosophical Perspectives on Cognitive Neuroscience*, Psychology Press. <https://doi.org/10.4324/9780203810095>
- Bicchieri, C., Muldoon, R., Sontuoso, A. (2023), “Social Norms”, in E. N. Zalta and U. Nodelman (Eds.), *Stanford Encyclopedia of Philosophy*. <<https://plato.stanford.edu/archives/win2023/entries/social-norms/>>
- Brandom, R. (1994), *Making it explicit: Reasoning, Representing and discursive commitment*, Harvard University Press.
- Brandom, R. (2014), “A Hegelian Model of Legal Concept Determination: The Normative Fine Structure of the Judges’ Chain Novel”, in G. Hubbs and D. Lind (Eds.), *Pragmatism, Law, and Language*, 19–39, Routledge.
- Brandom, R. (2019), *A Spirit of Trust: A Reading of Hegel’s Phenomenology*, Harvard University Press.
- Broome, J. (2013), *Rationality Through Reasoning*, John Wiley and Sons.
- Clark, A. (2013), “Whatever next? Predictive brains, situated agents, and the future of cognitive science”, *Behavioral and Brain Sciences*, 36(3), 181–204. <https://doi.org/10.1017/S0140525X12000477>
- Czajkowski, M., Zagórska, K., and Hanley, N. (2019), “Social norm nudging and preferences for household recycling”, *Resource and Energy Economics*, 58, 101–110. <https://doi.org/10.1016/j.reseneeco.2019.07.004>
- Darwall, S. (2001), “Normativity”, In *Routledge Encyclopedia of Philosophy*, Taylor and Francis; *Routledge Encyclopedia of Philosophy*. <https://www.rep.routledge.com/articles/thematic/normativity/v-1>
- De Waal, F. (2001), *The ape and the sushi master: Cultural reflections by a primatologist*, Basic Books.
- De Waal, F. (2006), *Primates and philosophers. How morality evolved*, Princeton University Press.
- Descartes, R. (1991), *The Philosophical Writings of Descartes: Volume 3* (J. Cottingham, R. Stoothoff, D. Murdoch, and A. Kenny, Eds.).
- Ebner, J. (2023), *Going Mainstream*, Ithaka.

- Fabry, R. E. (2018), “Betwixt and between: The enculturated predictive processing approach to cognition”, *Synthese*, 195(6), 2483–2518. <https://doi.org/10.1007/s11229-017-1334-y>
- Farrow, K., Grolleau, G., and Ibanez, L. (2017), “Social Norms and Pro-environmental Behavior: A Review of the Evidence”, *Ecological Economics*, 140, 1–13. <https://doi.org/10.1016/j.ecolecon.2017.04.017>
- Friston, K. (2009), “The free-energy principle: A rough guide to the brain?” *Trends in Cognitive Sciences*, 13(7), 293–301. <https://doi.org/10.1016/j.tics.2009.04.005>
- Friston, K. (2010), “The free-energy principle: A unified brain theory?” *Nature Reviews Neuroscience*, 11(2), 127–138. <https://doi.org/10.1038/nrn2787>
- Friston, K., Kilner, J., and Harrison, L. (2006), “A free energy principle for the brain”, *Theoretical and Computational Neuroscience: Understanding Brain Functions*, 100(1), 70–87. <https://doi.org/10.1016/j.jphysparis.2006.10.001>
- Ginther, L. (2022), “Kantian Objectivism and Subject-Relative Well-Being”, *Dialogue*, 61(3), 407–419. <https://doi.org/10.1017/S0012217322000269>
- Giromini, J. (2023), “The Logical Structure of Normative Attitudes”, *Philosophia*, 51(3), 1271–1291. <https://doi.org/10.1007/s11406-022-00587-9>
- Grice, H. P. (1975), “Logic and Conversation”, in *Syntax and Semantics* 3, 41–58, Elsevier.
- Harman, G. (1977), *The Nature of Morality: An Introduction to Ethics*, Oxford University Press.
- Heinämaa, S. (2020), “Disgust”, in T. Szanto and H. Landweer (Eds.), *The Routledge Handbook of Phenomenology of Emotion*, 380–391, Routledge.
- Helm, B. W. (2001), *Emotional Reason: Deliberation, Motivation, and the Nature of Value*, Cambridge University Press. <https://doi.org/10.1017/CBO9780511520044>
- Hohwy, J. (2013), *The predictive mind*, Oxford University Press.
- Hohwy, J. (2017), “Priors in perception: Top-down modulation, Bayesian perceptual learning rate, and prediction error minimization”, *Cognitive Penetration and Predictive Coding*, 47, 75–85. <https://doi.org/10.1016/j.concog.2016.09.004>
- Horne, C. (2001), “Sociological perspectives on the emergence of social norms”, *Social Norms*, 3–34.
- Hume, D. (2007), *A Treatise of Human Nature* (D. F. Norton and M. J. Norton, Eds.; Clarendon Edition of the Works of David Hume), Clarendon Press.
- Kappeler, P. M., Fichtel, C., and van Schaik, C. P. (2019), “There Ought to Be Roots: Evolutionary Precursors of Social Norms and Conventions in Non-Human Primates”, In N. Roughley and K. Bayertz (Eds), *The Normative Animal?: On the Anthropological Significance of Social, Moral, and Linguistic Norms*, 65–82, Oxford University Press. <https://doi.org/10.1093/oso/9780190846466.003.0003>
- Korsgaard, C. (1996), *The Sources of Normativity*, Cambridge University Press.
- Köster, M., Kayhan, E., Langeloh, M., and Hoehl, S. (2020), “Making Sense of the World: Infant Learning From a Predictive Processing Perspective”, *Perspectives on Psychological Science*, 15(3), 562–571. <https://doi.org/10.1177/1745691619895071>
- Kripke, S. (1980), *Naming and Necessity*, Harvard University Press.
- Kube, T., Schwarting, R., Rozenkrantz, L., Glombiewski, J. A., and Rief, W. (2020), “Distorted Cognitive Processes in Major Depression: A Predictive Processing

- Perspective”, *Mechanisms of Major Depression*, 87(5), 388–398.
<https://doi.org/10.1016/j.biopsycho.2019.07.017>
- Launis, V. (2018), *Ihmisarvo, Vastapaino*.
- Mackie, J. L. (1977), *Ethics: Inventing Right and Wrong*, Penguin Books.
- Mackie, J. L. (1982), “Morality and the retributive emotions”, *Criminal Justice Ethics*, 1(1), 3–10.
- Mannerström, R., Muotka, J., Leikas, S., and Lönnqvist, J.-E. (2020), ”Hyvin toimeentuleva, terve ja suvaitsematon – Suomalaisten nuorten poliittiset ideologiat asenneprofiileina tarkasteltuna”, *Politiikka*, 62(1), Article 1.
<https://doi.org/10.37452/politiikka.83247>
- Nagel, T. (1986), *The View from Nowhere*, Oxford University Press.
- Nummenmaa, L. (2010), *Tunteiden psykologia*, Tammi.
- Nussbaum, M. (2019), *The Monarchy of Fear*, Simon and Schuster.
- Parr, T., Pezzulo, G., and Friston, K. J. (2022), *Active inference: The free energy principle in mind, brain, and behavior*, MIT Press.
- Pendl, S. L., Salzwedel, A. P., Goldman, B. D., Barrett, L. F., Lin, W., Gilmore, J. H., and Gao, W. (2017), “Emergence of a hierarchical brain during infancy reflected by stepwise functional connectivity”, *Human Brain Mapping*, 38(5), 2666–2682.
<https://doi.org/10.1002/hbm.23552>
- Peregrin, J. (2021), “Normative Attitudes”, In L. Townsend, P. Stovall, and H. B. Schmid (Eds.), *The Social Institution of Discursive Norms: Historical, Naturalistic, and Pragmatic Perspectives*, 121–137, Routledge.
- Pezzulo, G., Parr, T., and Friston, K. (2024), “Active inference as a theory of sentient behavior”, *Biological Psychology*, 186, 108741
<https://doi.org/10.1016/j.biopsycho.2023.108741>
- Piekarski, M. (2019), “Normativity of Predictions: A New Research Perspective”, *Frontiers in Psychology*, 10, 1–4. <https://doi.org/10.3389/fpsyg.2019.01710>
- Pietarinen, J. (2015), *Suomalaisen filosofian vaiheita*, Areopagus.
- Pipatti, O. (2024), *The Origins Of Human Social Nature: Westermarckian Sociology and Social Anthropology*, Palgrave Macmillan.
- Putnam, H. (1970), “Is Semantics Possible?”, *Metaphilosophy*, 1(3), 187–201.
<https://doi.org/10.1111/j.1467-9973.1970.tb00602.x>
- Rawls, J. (2006), *A Theory of Justice: Revised Edition (Vol. 5)*. Belknap Press: An Imprint of Harvard University Press.
- Rist, J. M. (2012), *Plato’s Moral Realism: The Discovery of the Presuppositions of Ethics*, The Catholic University of America Press.
- Roughley, N. (2019), “Might We Be Essentially Normative Animals?” In K. Bayertz and N. Roughley (Eds.), *The Normative Animal?: On the Anthropological Significance of Social, Moral and Linguistic Norms*, 3–37, Oxford University Press.
- Saul, J. (2024), *Dogwhistles and Figleaves: Devices for Mainstreaming Racism and Falsehood*, Oxford University Press.
- Scanlon, T. (1998), *What we owe to each other*, Belknap Press of Harvard University Press.
- Searle, J. R. (1969), *Speech acts. An Essay in the Philosophy of Language*, Cambridge University Press.

- Searle, J. R. (2008), “Language and social ontology”, *Theory and Society*, 37(5), 443–459. <https://doi.org/10.1007/S11186-008-9068-Y>
- Searle, J. R. (2010), *Making the Social World: The Structure of Human Civilization*, Oxford University Press.
- Searle, J. R. (2019), “Money: Ontology and Deception”, In *Money, Social Ontology, and Law*, 7–24, Routledge.
- Sellars, W. (1954), “Some Reflections on Language Games”, *Philosophy of Science*, 21(3), 204–228.
- Silk, J. B. (2014), “Evolutionary Perspectives on the Links Between Close Social Bonds, Health, and Fitness”, In M. Weinstein and M. A. Lane (Eds.), *Sociality, Hierarchy, Health: Comparative Biodemography: A Collection of Papers*, 121–144, National Academies Press. <https://www.ncbi.nlm.nih.gov/books/NBK242452/>
- Sinokki, J. (2024), “Westermarck on Demarcating the Moral from Nonmoral” in E. Kaila, I. Niiniluoto, and S. Pihlström, (Eds.), *Acta Philosophica Fennica*, Vol. 101).
- Stroup, T. (1982), *Westermarck’s ethics*, Tidningsbokhandeln.
- Stroup, T. (1984), “Edward Westermarck: A Reappraisal”, *Man*, 19(4), 575–592. <https://doi.org/10.2307/2802327>
- Van de Cruys, S., and Wagemans, J. (2011), “Putting Reward in Art: A Tentative Prediction Error Account of Visual Art”, *I-Perception*, 2(9), 1035–1062. <https://doi.org/10.1068/i0466aap>
- Westermarck, E. (1906), *The Origin and Development of the Moral Ideas*, Books for Libraries Press.
- Westermarck, E. (1932), *Ethical Relativity*, Kegan Paul, Trench, Trubner and Co. Ltd.
- Wilkinson, S., Deane, G., Nave, K., and Clark, A. (2019), “Getting Warmer: Predictive Processing and the Nature of Emotion”, in L. Candiottto (Ed.), *The Value of Emotions for Knowledge*, 101–119, Springer International Publishing. https://doi.org/10.1007/978-3-030-15667-1_5
- Williams, D. (2018), “Pragmatism and the predictive mind”, *Phenomenology and the Cognitive Sciences*, 17(5), 835–859. <https://doi.org/10.1007/s11097-017-9556-5>
- Williams, D., and Colling, L. (2018), “From symbols to icons: The return of resemblance in the cognitive neuroscience revolution”, *Synthese*, 195(5), 1941–1967. <https://doi.org/10.1007/s11229-017-1578-6>
- Williams, M. L. (2021), *The Science of Hate: How prejudice becomes hate and what we can do to stop it*, Faber and Faber.
- Wittgenstein, L. (2009), *Philosophical Investigations* (G. E. M. Anscombe, P. M. S. Hacker, and J. Schulte, Eds.), Blackwell.

Why Should I Care about My Choices?

SUSANNE UUSITALO

Human agency and its various aspects are of interest to many disciplines. We manage to live our lives everyday navigating through duties and tasks, throwing ourselves into leisurely activities or omissions. Some prefer running marathons, others... do not. Some become absorbed in writing a manuscript, losing track of time and realizing in the middle of the night that they have forgotten to eat and meet other commitments, whilst others... do not. Nonetheless, these agents who differ in their ways of expressing their agency still share the mechanisms by which they, we, act in the world. We make choices and decisions based on what we think is possible and worthwhile in those circumstances.

Choices and decisions made by the human agent have also become more and more of an issue, especially in relation to discussions concerning autonomous artificial intelligence, AI. This discussion refers to technologies that involve some form of AI that is self-governing. From an ethical perspective this is of course relevant: who holds control over the actions (for considerations of different kinds of agential control over substance use, see Uusitalo 2023)? Control is not, however, the only relevant issue in making autonomous and meaningful decisions. Epistemic considerations play a role too. How much information suffices in making informed decisions? Another aspect is questions of authenticity: what choices and decisions are truly ours, representative of our values and commitments? In the framework of these aspects and questions, autonomy in human agency and action seems to be a prominent value that demarcates agency and action in various ways in different contexts.

In what follows, I will showcase certain kinds of human behaviour and hope to illustrate the salience of autonomy in understanding human agency and action. In the end, I will apply the discussion of autonomy in human agency and action to a new context and ask why I should care about my choices.

1. Gambling: A challenge to agency and action

An exemplar of the kind of behavior that challenges the traditional ways of understanding human agency and action is gambling. In philosophy, in the traditional, Anglo-American style theory of action, Donald Davidson advises us to see actions as causal chains of pro attitudes and beliefs that bring about intention and, voilà, we have action (unless of course there is something that causes deviation) (Davidson 1980). Applying this to gambling does not seem to do justice to the phenomenon. In fact, the theory of action seems to fall short of contributing to understanding what takes place when agents engage in gambling. This has to do with several aspects which are more or less characteristics of gambling. Roughly, gambling can be understood as behaviour, as action, in which the agents wager something of value, e.g. money, in order to gain a reward in a situation of high uncertainty. The circumstances of high uncertainty call into question the idea that it is actually reasonable, or worthwhile, to engage in the action. Of course the reward may in itself be something entirely else than “winning”. People report all kinds of reasons for their gambling and under those descriptions gambling could be seen as worthwhile regardless of the unlikelihood of their winning in the means of gambling (see for instance Grönroos et al. 2024). In any case, these call into question the rationality in the action, especially when the rationality is understood in an instrumental way – as means to gain the reward for which the agent has wagered something of value. Second, some of the most common ways of gambling actually involve aspects that challenge the implicit control conditions of the standard way of understanding and explaining action.

In disciplines interested in the mind, different kinds of cognitive models of (problematic) gambling have been proposed. Also, Arstila and many of his colleagues (2022, 2023) have put their brains to work to solve the enigma of gambling. They have been interested, in particular, in understanding the action of gambling and its relation to the kind of gambling that is considered somehow problematic. In other words, understanding gambling first as the kind of action that makes sense without determining every individual who gambles as irrational or out of control in their behaviour just by virtue of engagement in the act(s) of gambling.

2. The value in agency and action

Arstila, Uusitalo and Palomäki (2023) argue that Jacobs' general theory of addiction, for instance, fails to characterize problematic gambling appropriately with its dissociative framework because it disregards the fact that also non-problematic gambling has dissociative characteristics, e.g., experiences of flow. Flow is an optimal state of mind, an experience in which the mind is focused on a task that matches the abilities of the agent and is a good fit for the demandingness of the task. Flow experiences has been studied in various contexts, including in poker (see Palomäki & Laakasuo 2016). The concept as a meaningful tool to understand a certain psychological state can be traced back to the work of Mihaly Csikszentmihalyi (e.g., Csikszentmihalyi 1990) and the discussion of its role and relevance in action has been studied a great deal since (see for instance Palomäki et al. 2021). Even if this paper does not go into the subtleties of a flow account of gambling, flow and how it is understood need to be highlighted here: this kind of discussion about the states of mind in gambling is highly relevant from an ethical point of view. In discussing different kinds of states of mind, we can always consider what kind of agential possibilities and limitations the states of mind allow and contribute to. In issues related to autonomous choices and issues of responsibility, we are typically interested in the kind of control the agent can have and actually has when acting. Also understanding and knowledge play a role in those considerations of autonomous choices and responsibility attributions. In gambling understood either in a dissociative framework or in a flow framework, the traditional way of understanding "control over one's actions" is already challenged. Yet, it would be quite drastic to say that we do not or cannot exercise our autonomy in any cases when we engage in gambling that involves these kinds of states of mind. (Playing lotto is a form of gambling, but does not typically involve these kinds of states whilst playing poker or slots machines do.)

Gambling seems a phenomenon that invites researchers to approach it with different cognitive models. As it turns out, Arstila, Tuominen and Uusitalo (2023) have also provided an account of gambling that looks at gambling from the perspective of predictive processing. Predictive processing is a model that explains people's actions on the basis of generative models of the world that the agents hold and prediction error that can be minimized in either changing the model or acting upon the world. Arstila and his colleagues' move is not the first that has been made in the attempt to understand gambling from the perspective of predictive

processing (see Schwartenbeck et al. 2015), but it takes gamblers' prior beliefs to be different and varied in a way that can explain suboptimal behavior in a way that considers the agent's action rational, at least from the agent's own perspective. In gambling research the often cited "pathways model of problem and pathological gambling" could be seen to support this as well; Blaszczynski and Nower (2002) identified in a clinical population different subgroups that hold different characteristics, background and experiences regarding gambling.

The way in which the predictive processing model views gambling in general allows us to have the kind of control that seems to be required for autonomous choices and still be sensitive to forms of behaviour that require help from health care, for instance. Regardless of the scientific value of the model of predictive processing, this sensitivity is an ethical reason for accepting the account. It is not, of course, a sufficient reason for accepting the account, but it is from a research ethical view, a necessary one. The relevance of conceptualization and methodological choices that the scientific community makes in studying human behaviour, especially questions that concern behaviour and action that is not optimal, has strong ethical implications (see Uusitalo & Egerer 2024): by making us understand and view individuals' agency with the help of these models, we hold the power to regard the individuals who are conceptualized in these ways as being able to make autonomous, rational and worthwhile choices regarding their own actions or as not being able to make such choices.

3. An application of the value of autonomous choices in the AI-augmented world

So far, I hope to have convinced the reader of the importance of understanding nuances in human agency in theories and models of human mind and action with the help of examples of making sense of gambling with cognitive models. This is important for human agency and, as I have tried to show, it is also consequently important for keeping theoretical room and sensitivity for the ability to make autonomous choices and decisions in the models. This discussion has ethical implications in everyday lives too when the structures of the society function on the basis of sufficiently autonomous decision-makers' decisions and choices regarding their own lives.

Questions of autonomous decision-making surface in a topical discussion on persuasive technologies, in particular applications of AI that make suggestions to

us concerning what music to listen to, what route to take, what to read, eat, watch, buy... and so on (or arguably, make our decisions and choices for us). The applications are designed to make people's lives easier, to provide options that – on the basis of their previous choices or behaviour – they are likely to choose and behave in accordance with. This is another way of modelling the individuals.

Arstila and his colleagues have come to the conclusion on some late summer night that there needs to be caution with the use of AI in mental health (Uusitalo, Tuominen & Arstila 2021) and the same could be highlighted here. If there is value in making autonomous, meaningful choices in our everyday lives, engaging in the procedures of autonomous choices, reflecting on our values and preferences, is there a risk in these applications that they actually reduce the exercise of individual autonomy by offering us the most convenient, easiest, options? Is something lost with the use of these applications? If so, is it something that should not be lost? What is the value of individual autonomy in these times of the digitalization of many aspects of human life? She asks and clicks on the assisted editing button in her word processing programme.

*University of Turku*¹

References

- Arstila, V, Uusitalo, S and J. Palomäki (2023), "Miksi pelaamme rahapelejä?", in S. Uusitalo and V. Arstila (2023), 133–155.
- Arstila, V, Tuominen, J. and S. Uusitalo (2023), "Heads, I Win – Tails, You Lose: Gambling from a Predictive Processing Perspective", in M. Curado and S. S. Gouveia (2023), 181–199.
- Blaszczynski, A. and L. Nower (2002), "A pathways model of problem and pathological gambling", *Addiction* 97, 487–499.
- Curado M. and S. S. Gouveia (eds.) (2023), *Predictive Minds: Old Problems and New Challenges*, Wilmington, Vernon Press.
- Csikszentmihalyi, M. (1990), *Flow - The Psychology of Optimal Experience*, New York, Harper Perennial.
- Davidson, D. (1963), "Actions, reasons, and causes", in D. Davidson (1980), 3–19.
- Davidson D. (1980). *Essays on Actions and Events*, Oxford, Clarendon Press.
- Grönroos, T., Salonen, A., Kontto, J. and H. Hagfors (2024), Suomalaisten rahapelaaminen 2023: Rahapelaaminen vähentynyt, peliongelma yleistynyt ja suhtautuminen pelaamiseen muuttunut, Tilastoraportti 15/2024, Terveyden ja hyvinvoinnin laitos.

¹ Supported by the University of Oulu and the Research Council of Finland Profi 7 352788.

- Nuotio, K., Simojoki K. and E. Yli-Hemminki (eds.) (2023), *Parempaa päihdepolitiikkaa: Rangaistuksista hoidon poluille*, Helsinki, Gaudeamus.
- Palomäki, J and M. Laakasuo (2016), “Zone out and your money flows away; go with the flow and your money might stay”, *Nordic Studies on Alcohol and Drugs* 33(1), 111–113.
- Palomäki, J., Tammi, T., Lehtonen, N., Seittenranta, N., Laakasuo, M., Abuhamedh, S., Lappi, O. and B. Ultan Cowley (2021), “The link between flow and performance is moderated by task experience”, *Computers in Human Behavior* 124: 106891.
- Schwartenbeck, P., FitzGerald, T. H. B., Mathys, C., Dolan, R., Wurst, F., Kronbichler, M. and K. J. Friston (2015), “Optimal inference with suboptimal models: Addiction and active Bayesian inference”, *Medical Hypotheses* 84 (2), 109–117.
- Takala T. and M. Häyry (eds.) (2024), *Concise Encyclopedia of Applied Ethics in the Social Sciences*, Cheltenham and Camberley, Edward Elgar Publishing.
- Uusitalo, S. (2023) ”Halun viemää? Filosofiseettinen näkökulma päihteidenkäyttöön”, in K. Nuotio, K. Simojoki and E. Yli-Hemminki (2023), 169–184.
- Uusitalo S. and V. Arstila (eds.) (2023), *Ongelmallinen rahapelaaminen*, Turku: Eetos ry.
- Uusitalo, S. and M. Egerer (2024), “Addiction studies”, in T. Takala and M. Häyry (2024), 1–9.
- Uusitalo S., Tuominen J. and V. Arstila (2021), “Mapping out the philosophical questions of AI and clinical practice in diagnosing and treating mental disorders”, *Journal of Evaluation in Clinical Practice* 27(3), 478–484.

Jenann Ismael on Causation and Free Will¹

SAARA WUOKKO

1. Introduction

Causal determinism is often contrasted with free will. In a simple form the claim is that if every event has causal antecedents that determine the event, it is not possible for an agent to act freely. Especially classical physics is in many cases taken to prove that free will is impossible. Physics seems to imply that from the so-called ‘god’s eye-view’, if all the initial conditions and all the laws were known, it would be possible to calculate what happens at any given time, even in the future. Jenann Ismael argues that both free will and determinism are true, and wants to show, as the name of her book (2016) suggests, “how physics makes us free”.

Ismael’s view is that the physical laws of the universe are deterministic (Ismael 2023), but human beings still have free will as agents.² John Maier writes in a book review: “Ismael is a compatibilist—although this is a term that she herself seldom uses, and about which I’ve indicated some unease—but her view is unlike many of the views that go under that label, and is best appreciated on its own terms.” (Maier 2017, 196.)

¹ This paper has undergone peer review.

² It is usually said that there is indeterminism in quantum mechanics: some quantum events seem to be totally random and there are genuine chances (that is to say, chances that would still be chances even if we had all the relevant knowledge). Why then concentrate on deterministic classical physics? In the preface of her book *How physics makes us free*, Ismael notes that it is not yet known how quantum mechanics relates to human actions and that she thus focuses on classical physics (Ismael 2016). Elsewhere she has written: “Since we are likely too big, too warm, and too slow for quantum effects to show up in our behavior, classical physics is likely to be the effective physics for understanding the human being.” (Ismael 2023, 148.) She also follows many others claiming that quantum indeterminism does not help if we want to defend free will. Indeterminism means that we should rely on chances that we cannot control, but we want to be able to make the decisions, to control our own actions. (Ismael 2016, 98.)

My main interest is in the causal aspects of the free will debate. In causal determinism, causation together with the thought that everything is already determined, leads to the idea that the future is laid out and it cannot be changed. Causal chains start from the initial conditions, from the beginning of the universe, and continue all the way to this moment, causing what will happen next. However, contrary to that thought, Ismael uses the notion of causation in order to explain how human beings can be free in a deterministic universe. This understanding of freedom would be compatible with classical physics.

In this text I will focus on the role of causation in Ismael's arguments for free will. My claim is that Ismael's way of reconciling determinism and free will leaves some open questions and that Ismael's causation-related arguments for free will are insufficient. First, I will briefly explain what causation is and introduce some history of causation from Ismael's viewpoint. Then I will go to Ismael's arguments for the importance of causation, how we need causal talk to notice possibilities, what could or would happen, and to separate effective strategies from ineffective ones by making the choice of exogenous and endogenous variables. After that, I will turn to Ismael's argument that causation emerges in the subsystems of the universe and free will arises from that through deliberation. Lastly, I will summarize how asymmetry of causation is presented in her texts and briefly discuss perspectivalism in relation to causation and free will.

2. What is causation?

Developmentally, we become familiar with causation from a very early age. When a young child learns how to control their limbs, the child is not far away from figuring out how to make an effect on the surrounding world. (Ismael 2012, 210.) All the pushing and pulling results in changes in the environment and later on, it is natural to think that it is oneself, it is a 'me', doing all the pushing and pulling, voluntarily. We have the feeling of freedom because we seem to be able to causally affect things.

Causation, or cause-effect relation, can be defined as a relationship in which a thing gives rise to another thing, for example, an event. According to the manipulability theories of causation, cause and effect are in question, if by manipulating A, one can make a manipulation or change in B. In interventionist theories these manipulations are called interventions. The notion of intervention specifies that the change in A needs to be 'surgical', surgical meaning that the

causal route has gone via A and resulted in changes in B, and changes in B did not happen in any other way. (Woodward 2023.) Interventionists think that manipulability theories are anthropocentric because manipulation needs an agent. Interventions, on the other hand, do not need human action. (Ismael 2016b, 247.)

It is usually thought that causes always precede their effects in time, because this is what causation seems like in our everyday life. David Hume put it in his definition of causation that cause is followed by its effect (Morris & Brown 2022). Nowadays, some philosophers and physicists argue that backwards causation could be possible. In backwards causation or ‘retrocausality’ an effect would happen before its cause in time. For example, Huw Price argues that retrocausality could explain some of the oddities in quantum mechanics (see, for example, Price 1997).

As Ismael points out, the Newtonian dynamical laws do not involve causation. They are “regularities of motion. They describe how the values of quantities that characterize the state of a system vary over time” (Ismael 2016, 117). The dynamical laws do not tell us what would have happened in a slightly different situation if things were in a different way. For that we need contrafactual thinking and the idea of causation. Some people have thought that the absence of causation in the dynamical laws is a reason to abandon the notion of causation.

3. From refuting causation to acknowledging the importance of it

Ismael mentions Russell many times in her works (see, for example, Ismael 2007; 2016; 2018), and especially his text *On the notion of cause* (1913). In this text Russell famously announces that causation is “a relic of a bygone era”. He argues that there is no place for causation in physics because causes are not mentioned in the physical laws. Another writer Ismael refers (for example, Ismael 2012; 2016; 2018) a lot to is Nancy Cartwright, who (1979) writes about how causation is needed for us to know which strategies are effective and which ineffective. The example Ismael borrows from Cartwright is smoking: bad breath and cancer are positively correlated, but treating bad breath does not cure cancer. This is because the cause of both is smoking, and changing people’s smoking habits makes the difference. Not smoking is an effective strategy for avoiding bad breath and cancer that are both caused by smoking. (Ismael 2016, 116.)

Ismael suggests that the dichotomy between causation being either necessary or a naïve illusion is outdated (Ismael 2016c, 434). Causal reasoning is needed in

everyday life, but according to Ismael, it is also a “truism” in sciences, for example, in psychology (Ismael 2016, 135). By being able to separate effective strategies from ineffective ones, we can decide what is a good way to act. With the help of causation, we can make decisions and affect things around us. Causal thinking tells us what could happen or would happen in a certain situation. We can ponder if we should do this or that, and how would doing this or that change the outcome. We are not only interested in what actually happens, but what could happen. In life in general, but also in sciences, we want to know how things could turn out, if the conditions were slightly different. We need to find the ways to alter the situation in a way that we reach the desired outcome. When we talk about what is possible or necessary, we narrow our scope and choose the variables according to the context.

4. Endogenous and exogenous variables

Different perspectives give different responses to the question ‘what was the cause of this?’. To be able to talk meaningfully about causes, we need to choose which conditions we call causes in specific situations. Every event is affected by so many aspects that it is impossible to consider all the causes of a particular situation. That is why we need to limit our discussion and make the choice of variables. According to computer scientist Judea Pearl, that choice has to do with sufficiency and necessity. In his 2018 book Pearl uses “a classic example”: a match is struck, and a fire happens. We might ask³: “What caused the fire, striking the match or the presence of oxygen in the room?” Both striking the match and presence of oxygen are necessary causes. What makes them different is the aspect of sufficiency, but Pearl adds that there seem to be psychological factors, too. (Pearl & Mackenzie 2018, 289–290.)

Ismael writes about Pearl’s notions of exogenous and endogenous variables. When treating a system causally, we choose which variables to treat as endogenous and which as exogenous. Endogenous variables are the ones in the system and exogenous outside the system. (Ismael 2016b, 258.) In different situations different things can be regarded as endogenous or exogenous variables and some things are held fixed. For example, there might be many things that can lead to a car accident, but we only consider some of them as causes. Some things

³ We probably would not ask this kind of question in a normal situation, since we would quite automatically say that striking the match is the cause.

we might call background conditions instead of causes and some of the causes or background conditions we are not even aware of.

All that is said here before seems to only address how causation and free will work in general in our daily life. To me, the problem lies elsewhere. There seems to be bigger questions about the reality of causation and free will that have nothing to do with the purposeful pushing and pulling that we know how to do already as little children, or the capacity that we humans have for causal thinking. Where does causation really come from, is it ‘out there in the world’, or is it something subjective? What is the metaphysical status of causation in Ismael’s texts?

5. Open and closed systems, and deliberation

Ismael emphasizes that causation does not work in the level of the whole universe, but in smaller subsystems of it. The universe as a whole is a closed system, which means that there is nothing affecting it from the outside. Because there are no outside powers, there is no causation. In smaller systems, like human beings, or engines (Ismael uses a car engine as an example) there are a lot of effects from the outside world. The smaller, open, subsystems are where free will arises. (Ismael 2012.)

Ismael writes: “Although many of the systems that we study in the laboratory approximate closed systems, the only truly closed system is the universe as a whole, since it has no environment” (Ismael 2016, 90). Without environment there are no outside effects that could affect the system. Contrary to the whole universe, the open subsystems of the universe have outside effects and therefore also causality.

It is important to know what would happen to an engine in certain situations that are not part of the engine’s normal working mechanism. It is not enough to know how the engine would work if nothing special happened. We want to be able to act in an abnormal situation as well, and that is why causal knowledge is needed. (Ismael, 2016.) Causation arises in the local systems, subsystems of the universe. From causation emerges free will, through deliberation.

According to Ismael, we encounter the world as agents (2018, 175). We are doing things, but first deciding to do them after some time of deliberation. Ismael claims that deliberation is our means for operating free will. In deliberation, we decide how to act. In difficult situations, where we must make a choice, we can

feel the weight of our decisions. This is a sign that we practice free will to come up with a plan of action. Ismael writes that “causal relations are handmaids to choice” (Ismael 2016, 134). This means that causation is our way of affecting the world. Already as little children we feel the power of our own decisions resulting in actions. When we manipulate causes or make interventions, we use our free will to make a change in the world. But is this enough to cut the ties of causal determinism?

6. Asymmetry of causation

As mentioned before, causation is often thought as a temporally asymmetric relation: cause comes first, then its effect. One thing that could ground this asymmetry is the thermodynamical asymmetry together with the low entropy initial condition of the universe. According to that theory, entropy is low now, but was even lower in the beginning of the universe. Increasing entropy and laws of thermodynamics explain why everything seems to happen temporally to one direction. (See, for example, Price 1997.)

According to Ismael, the thermodynamic gradient with the low entropy boundary condition, assuming the interventionist causal picture, generate the asymmetry of causation. Ismael writes that there are two asymmetries in causation: one practical and one epistemic. Human beings cannot affect the past but can affect the future, by means of making actions that generate records. Ismael claims that causal asymmetry is at least partly objective because the boundary condition is an objective fact about the world. Although the asymmetry is seen only in the macrostructures (there is symmetry in the micro level), it is not entirely subjective or perspectival, but it reflects the way the world is. (Ismael 2023b.)

7. Fixed past and open future

Usually, it is thought that we cannot affect the past, but can make an effect on the future. Price claims that our temporal orientation is fixed because of the laws of thermodynamics, but only from our perspective (Price 1997, chapter 7). Ismael argues against the view that determinism seems to entail, that the fixity of the past would mean the fixity of the future. According to Ismael, relativity prevents the fixity of the future: “In Minkowski space-time, notions of past and future become relativized to points, replacing the pre-relativistic division of events into past,

present, and future with the division into causal past, absolute elsewhere.” (Ismael 2023, 166.)

When it comes to perspectival or objective characters of causation, Price and Ismael seem to be mostly talking about the same things but stressing different viewpoints. Price highlights the more subjective aspects and Ismael emphasizes objectivity. What differs in their views is that Price is ready to go further in the pragmatic direction (see Price 2023 on neopragmatism) and Ismael comes across as more realist about causation. For example, in her book (2016) Ismael uses the term perspectival only a few times but defines her view in some other works (see, for example, Ismael 2023b).

8. Discussion

One argument Ismael uses as evidence for human freedom is common sense and the feeling of freeness. Her example is small children, who gradually learn to affect their environment. But as I mentioned before, the idea of mere feeling of freedom being important feels empty. We can feel all sorts of things, but it does not necessarily tell us how things really are.

What comes to deliberation, we might acknowledge that there are reasons for a person to act in a certain way, but we still think that usually humans have the capacity to make their own, deliberate decisions. For instance, bad background conditions do not necessarily have to lead to bad behavior. This is why we think that people can be morally blamed for their actions. We often think that they had the possibility choose otherwise. But, in my opinion, it is difficult to say, what is the degree of freedom of a human being, to whom every single event from the past is making an effect upon. Tuomas Pernu (2017) writes in a review article of *How physics makes us free*:

The problem is that when you scratch the surface, it starts to look like the fundamental issues of the free will debate are left largely untouched. Most of the ammunition in the discussion is spent on giving a naturalistic account of decision making and action production. However, whether our decisions are truly effective, and whether their effectiveness can be explained in naturalistic terms, is not the core question of free will—the incompatibilists have no quarrel with the idea that we act according to conscious decisions,

even if determinism holds. Rather, the core question is this: are we free to make the decisions according to which we act?

I am not sure that Ismael gives an answer to that question. I agree with Pernu, that most of the arguments are arguments that explain our decision making, not free will. Ismael also argues that we should concentrate on the subsystems, not the world as a unit, because we never encounter the universe as a whole. We can never access the world from the god's eye-view. She claims that "Philosophers tend to be uninterested in partial views of bits of the world. They make a lunge for the most encompassing view. Most day-to-day science, however, is not concerned with the world as a unit, but is focused on local subsystems." (Ismael 2018, 176.) For me, the fact that philosophers concentrate on the big picture is not a problem. While it might be true that science has to do with the local subsystems, I think that one of the tasks of philosophers is to think outside of the scope of everyday life.

One argument against causation is that physical laws are time-symmetric, and causation seems to be asymmetric. In classical Newtonian physics, laws are time reversal invariant, which means that the laws do not have temporal direction. Cause and effect depend on the choice of variables that are held fixed and those that are held free. Causation, on the other hand, seems directed. Ismael argues that the directionality of causation has to do with the asymmetry of the thermodynamic gradient. She claims that this asymmetry is objective, as in not only perspectival, and calls her view thoughtful perspectivalism in contrast to glib perspectivalism, which "about some feature of the world says that it is merely an artefact of the lenses through which we see it" (Ismael 2023b, 7).

This split between thoughtful and glib perspectivalism is interesting because it indicates the problem that perspectivalism might seem too subjective. If we want to say that causation and free will are not only inside our heads, but have objective elements, too, it is useful to make the division. Glib perspectivalism does not necessarily bother me, since I would be ready to believe that there are no such things as causation or free will. But to be more moderate, I could say that I am leaning toward the perspectival view on causation that, for example, Price (2023) writes about under the name of (neo)pragmatism, and I think it could fit with Ismael's view of causation that only appears in the local subsystems.

9. Closing words

I agree with Ismael and many others that we need the notions of causation and free will in the daily life, but I am not convinced about their ontological status. My problem with Ismael's viewpoint is that I do not think she sufficiently explains what freedom is metaphysically. To me, her notion of freedom feels different from the idea of freedom that we normally have, which is more fundamental. I think we mean two different things when we speak about free will as fighting against determinism, and free will as our deliberate choices in day-to-day life. Freedom feels like something big and encompassing, and not like something that we have just because of our limited perspective. Claiming that free will happens in a smaller scale does feel, in a way, begging the question.

Ismael writes in her 2016 book in the "suggestions for further reading" part: "There is a turning tide in the literature, to some degree, towards regarding causation as non-fundamental. – – These ideas are beginning to be applied to discussions of free will." (Ismael 2016, 255.) I do not know what Ismael's thoughts are on this "turning tide". To me, it seems like some people have regarded causation as non-fundamental for a long time. But maybe the difference is that causation is now taken at the same time seriously, as a necessary part of our everyday life, and as not fundamental. I wonder also what Ismael thinks about free will being non-fundamental. In a way she seems to be advocating free will as a fundamental thing, saying that we have real freedom, but also that the kind of freedom we have depends on us:

One natural way of understanding freedom is as the clear apprehension of our position in the choice situation: the awareness of being "unbound" by anything that is built into the fabric of the universe, in charge of our own fates, able to decide our own responses to incoming information, and to thumb our noses at anyone that thinks they know what we will do. (Ismael 2016, 191–192.)

Maybe this kind of freedom is motivational in that it satisfies the intuition of having free will and does not contradict with physics. It also lets us hold other people ethically responsible. However, I think that this psychological freedom is not enough. Ismael continues later in the same paragraph:

In the remainder of the book, I want to look at how robust and real the kind of freedom that we have (the capacity to possess) in virtue of being choosers is. You will have to judge for yourself, of course, but in my view, it is not some second-rate or illusory freedom. It is the real deal. And the whole deal. (Ismael 2016, 192.)

But what does ‘real’ mean here? If causation and free will are perspectival, but not illusory, where goes the line between perspectival and real aspects in the case of causation and free will? Which parts are perspectival and which ‘real’? I do not quite see how causation, if taken to be itself a perspectival notion, could help us explain free will, that is also perspectival. What is there to choose? ⁴

University of Helsinki

References

- Cartwright, N. (1979), “Causal Laws and Effective Strategies”, *Noûs* 13/4, 419–437, <https://doi.org/10.2307/2215337>.
- Ismael, J. (2003), “Closed causal loops and the bilking argument”, *Synthese* 136/3, 305–320, <https://doi.org/10.1023/A:1025170026539>.
- Ismael, J. (2007), “Freedom, Compulsion, and Causation”, *Psyche* 13/1, April, <https://journalpsyche.org/files/0xab0e.pdf>.
- Ismael, J. (2012), “Causation, Free Will, and Naturalism”, *Scientific Metaphysics*, edited by Harold Kincaid, James Ladyman, and Don Ross, Oxford University Press, pp. 208–236.
- Ismael, J. (2016), *How Physics Makes Us Free*, Oxford University Press.
- Ismael, J. (2016b), “How Do Causes Depend on Us? The Many Faces of Perspectivalism”, *Synthese* 193/1, 245–67.
- Ismael, J. (2016c), “Review of Causal Reasoning in Physics, by M. Frisch”, *The Philosophical Review* 125/3, 431–435, <https://www.jstor.org/stable/26783725>.
- Ismael, J. (2018), “Causal Content and Global Laws”, *The Experimental Side of Modeling*, edited by Isabelle F. Peschard, and Fraassen, Bas C. van, University of Minnesota Press.
- Ismael, J. (2023), “Rethinking time and determinism: what happens to determinism when you take relativity seriously”, *Time and Science*, edited by Remy Lestienne and Paul Harris, World Scientific Publishing.
- Ismael, J. (2023b), “Reflections on the asymmetry of causation”, *Interface Focus* 13: 20220081, <https://doi.org/10.1098/rsfs.2022.0081>.

⁴ Acknowledgements. I would like to thank the referees for their very useful suggestions. Funding: University of Helsinki, Appearance and Reality in Physics -project; Finnish Concordia Fund.

- Maier, J. (2017), “How Physics Makes Us Free, by J. T. Ismael: New York: Oxford University Press, 2016, pp. xiv + 273, £19.99 (hardback)”, *Australasian Journal of Philosophy* 96(1), 196–199, <https://doi.org/10.1080/00048402.2016.1276945>.
- Morris, William Edward and Charlotte R. Brown. 2022. “David Hume”, *The Stanford Encyclopedia of Philosophy* (Summer 2022 Edition), Edward N. Zalta (ed.), <https://plato.stanford.edu/archives/sum2022/entries/hume/>.
- Pearl, J. and D. Mackenzie (2018), *The Book of Why: The New Science of Cause and Effect*, Basic Books, New York.
- Pernu, T. (2017), “Can Physics Make Us Free?”, *Frontiers in Physics* 5, <https://www.frontiersin.org/articles/10.3389/fphy.2017.00064>.
- Price, H. (1997), *Time's Arrow and Archimedes' Point: New Directions for the Physics of Time*, Oxford University Press, New York.
- Price, H. (2023), “Time for Pragmatism”, in Joshua Gert (ed.), *Neopragmatism: Interventions in First-order Philosophy* (Oxford, 2023; online edn, Oxford Academic, 24 Aug. 2023), <https://doi.org/10.1093/oso/9780192894809.003.0002>.
- Russell, B. (1912), “On the Notion of Cause”, *Proceedings of the Aristotelian Society* 13, 1–26, <http://www.jstor.org/stable/4543833>.
- Woodward, J. (2023), “Causation and Manipulability”, *The Stanford Encyclopedia of Philosophy* (Summer 2023 Edition), Edward N. Zalta & Uri Nodelman (eds.), <https://plato.stanford.edu/archives/sum2023/entries/causation-mani/> .

II

TIME AND TEMPORAL EXPERIENCE

Explaining Temporal Phenomenology: Hume's Extensionalism and Kant's Apriorism¹

ADRIAN BARDON

The empiricist needs to explain the origin, in perception, of the idea of time. Kant believed the only answer was a kind of idealism about time. This essay examines Hume's extensionalism as a possible answer to Kant. Extensionalism allegedly accounts for the experience of time via the manner of presentation of experiences, rather than the content of experience.

Empiricism has a time problem. The classical empiricist, by definition, is committed to the proposition that all ideas or concepts are ultimately derived from some combination of sensation, reflection upon sensation, and abstraction. So where do concepts like duration or succession come from? Locke's answer is that the idea of succession comes from reflection on the train of ideas passing through our minds:

It is evident to any one who will but observe what passes in his own mind, that there is a train of ideas which constantly succeed one another in his understanding, as long as he is awake. Reflection on these appearances of several ideas one after another in our minds, is that which furnishes us with the idea of SUCCESSION: and the distance between any parts of that succession, or between the appearance of any two ideas in our minds, is that we call DURATION. (*Essay* I.xiv.3)

¹ This paper was originally published in *Kant-Studien* (2019, 110(3), 463-476, <https://doi.org/10.1515/kant-2019-3006>) and is reprinted here with the permission of publisher De Gruyter.

The idea of duration comes from reflection on the distance between any two such ideas, as measured by the number of ideas that may occur between them. The problem (if there is one) comes from the fact that Locke proposes no access to past ideas except via memory. Reflection upon one's ideas, without any retention or reference to past ideas, cannot constitute reflection upon a succession of ideas. If we are relying on the revival of past ideas via recollection, then "reflection on those appearances of several ideas one after another" can only mean the juxtaposition, in the mind, of an idea represented as present with some other ideas represented as past. But a memory is a presently occurring idea that we think of as representing or otherwise referencing a past idea. In order for us to be conscious of a present idea *as* a representation of a past impression, we must already be applying the concepts of succession and duration. Only then can a memory have the meaning for us that it does, such that it can be taken to stand for a segment of a "train of ideas". This would just bring us back to the question of the origin of temporal concepts.

Thus, we have the (alleged) problem of time for empiricism, insofar as the empiricist's assumption is that our awareness of change over time is dependent on memory:

- All our ideas derive from original impressions of the senses, reflection on those impressions, and acts of conceptual abstraction.
- If the past and future are not directly perceived, then neither is temporal succession (or duration).
- Awareness of any current state as a complex state – the parts of which represent temporally distinct experiences – depends on a grasp of temporal succession.
- So, our grasp of temporal succession cannot originate either in a simple impression of the senses or in simple reflection on experience.

A successive occurrence of ideas is not the same as an awareness of succession. This observation has been made more than once:

Successive ideas are not yet the idea of succession, because succession *in* thought is not the thought *of* succession. (Volkman, cited in James 1910, 629)

The succession of sensations and the sensation of succession are not the same (Husserl 1964, 12)

[A] succession of ideas is quite different from an idea of succession. (Paton 1929, 318)

Et cetera.² The juxtaposition of a recollected idea with a present impression seems (at best) a presentation of co-presence, not succession.

The basic problem is that the raw content of experience does not in itself include the spatiotemporal framework within which experienced external (and internal) events are all thought to occur. Thus, Kant's focus on spatiotemporal unification as the mind's primary contribution to conscious experience – in particular the unification of all actual and possible experience in one time: As he points out in the *Aesthetic*, all actually – and potentially – experienced events occur in time, while not all need be represented as spatially related. In contemporary cognitive science, the problem of how the brain organizes its various perceptual inputs into one coherent whole is known as the “binding problem”.³ This problem is sometimes expressed as two different problems: The segregation problem (or “BP1”), or how discrete inputs are allocated to discrete objects; and the combination problem (“BP2”), or how discrete elements of experience at any level are transformed into a single, unified experience. Whether these are really two different problems is controversial (and Kant most certainly views them as part of the same problem, his solution to which takes up much of the *Critique of Pure Reason*); yet Kant's Analogies are most explicitly aimed at the combination problem. His solution is a set of innate, a priori guidelines for the unification of experience in time.

As he repeatedly states in the Analogies, the combination problem arises because “time is not perceived”. The content of experience itself does not tell us how (or even whether) different events are related to each other in time. His solution, of course, involves a priori rules for the combination of experiences in time that in turn are partly constitutive of objective experience.

² For examples and discussion, see Hoerl 2013.

³ See, for example, Roskies 1999.

Neither Kant's apriorism nor his subsequent idealist turn sit well with the later Western empiricist consensus. And it all hinges on the problem of the unification of experience in time. The empiricist must account for this unification consistently with empiricist principles. For the most part, empiricist answers to this problem have historically gone in one of two directions: either looking for temporal guidance in aspects of the inner presentation of experience ("retentionalism"), or by redefining experience itself to incorporate the temporal structure of experience ("extensionalism").

In his *Principles of Psychology*, William James argued that one's 'current' state encompasses a stretch of time – a "specious present". The specious present forms part of my experience because the 'pastness' of some of the contents of my current state is somehow a felt aspect of my complex current state.

To remember a thing as past, it is necessary that the notion of 'past' should be one of our 'ideas.' [...] But how do these things get *their* pastness? What is the *original* of our experience of pastness, from whence we get the meaning of the term? [...] [A]ll our concrete states of mind are representations of objects with some amount of complexity. Part of the complexity is the echo of the objects just past, and, in a less degree, perhaps, the foretaste of those just to arrive. Objects fade out of consciousness slowly. If the present thought is of A B C D E F G, the next one will be of B C D E F G H, and the one after that of C D E F G H I – the lingerings of the past dropping successively away, and the incomings of the future making up the loss. These lingerings of old objects, these incomings of new, are the germs of memory and expectation, the retrospective and the prospective sense of time. (James 1910, 606)

For the retentionalist, experience encompasses a stretch of time in the sense that some just-past experiences are retained as new information accrues; and the retained experiences (as distinguished from memory) have some character of 'pastness' as a part of their fundamental presentation. This is how the concept 'past' arises, and is subsequently applied to memory and past events. Other well-known retentionalists are Edmund Husserl and Bertrand Russell.

There are two standard objections to retentionalism. First is that of J.D. Mabbott (1951), who (in response to the Jamesian school) pointed out that a moving specious present as described by James, incorporating an extended sequence ABCD (with ABC presented as past), and followed by BCDE (with BCD presented as just-past), would seem to suggest a phenomenology of multiple, overlapping experiences: On the face of it, this account would suggest that the experience of, say, event B would echo several times in phenomenal experience.

Russell (1995, 157) argued that a sense-organ, once stimulated, “goes on vibrating, like a piano-string, for a while after the stimulation.” While the sensation fades, we continue to literally perceive what happened a short while ago.⁴ This vibration subsequently fades away; but in the meantime we have a kind of direct perceptual access to past inputs. Kant scholar H. J. Paton’s (1951, 105-107) response to Russell:

[I]f in a moment I can sense several different positions of a second hand, then these different positions would be sensed as being all at the same moment. That is to say what I should sense would not be a movement, but a stationary fan covering a certain area, and perhaps getting gradually brighter towards one end. Anything else would surely be a miracle. You can’t see a sensum that isn’t there. If you see it, it is there at the time you see it.

Similarly, a series of notes all heard at once would produce the experience of a chord, not a melody. Paton insists instead that the only way we could be aware of past perceptions is through memory. Regardless of whether you prefer the retentionalist’s or the memory theorist’s account of the structure of temporal experience, the route to the temporal phenomenology – the seemingly pure experience of motion/change – remains mysterious. Retentionalism does not make the perception of change any *more* mysterious than it was before; the problem is that it does not illuminate a way out of the temporal binding problem either. It just re-describes the binding problem to be addressed.

Empiricists are trying to avoid the Kantian idealist’s solution to all this: Namely, that temporal experience means discriminating earlier-later

⁴ See Bardon 2013, 39–40.

within one momentary state, which in turn is a *cognitive* achievement requiring a priori rules. Conscious awareness of *subjective* temporal complexity is based on a priori operations regarding *objective* alteration and causation. In his Analogies, Kant is quite clear about this as regards the sort of temporal order discriminations necessary to making empirical judgments about change and causation. (That is, to bringing experiences under empirical concepts.) Sensory inputs and reflections thereupon become experience for us because we have a priori guidance in relating the diverse elements of sense-data together under one timeline, and in distinguishing between a mere subjective sequence of inputs and an objective order incorporating all events.

There is, however, an empiricist alternative to retentionalism: extensionalism, according to which we perceive succession directly via temporally extended acts of awareness. Under extensionalism, there is no inference to be drawn (as there is under retentionalism) from the character or content of experience to its temporality; according to extensionalism, there is no distinction between an extended sequence of experiences and an experience of sequence.

Lived experience of movement/change/process suggests something more direct than a reconstructive judgment. The prima facie advantage for extensionalism is that “pace perceived”⁵ – the experience of change (such as the observed movement of a second hand) as opposed to the judgment that change has occurred – is allegedly accounted for simply by “a concurrence between the process of experiencing and the experienced process” (Hoerl 2009, 5). This concurrence sounds more like the experience of change simpliciter, as opposed to a representation based on recreation or reverberance. The extensionalist insists on what Barry Dainton (2017) calls the Immediacy Thesis: namely, that “Change, succession and persistence can feature in our experience with the same vivid immediacy as colour or sound, or any other phenomenal feature.” This would dissolve the combinatorial temporal binding problem. Prominent contemporary proponents include Dainton (2000), Ian Phillips (2014), and Christoph Hoerl (2013).

Oddly, it is almost impossible to find any references to David Hume in either recent or contemporary discussions of extensionalism. In the second section of his *A Treatise of Human Nature*, Hume articulates an

⁵ A term introduced by Kelly 2005.

extensionalist account of our perception of space and time — an almost entirely disregarded account that really ought to be the locus classicus for extensionalism.

Hume's *Treatise* Book I, Part 2 is entitled "Of the ideas of space and time". There Hume (*Treatise* 1.2.1, 27) describes our impressions of sequences in time and things in space as discriminable ultimately into (effectively) unextended "minima sensibilia":

[T]he imagination reaches a minimum [...] of which it cannot conceive any sub-division [...]. It is the same case with the impressions of the senses as with the ideas of the imagination. Put a spot of ink upon paper, fix your eye upon that spot, and retire to such a distance, that, at last you lose sight of it; it is plain, that the moment before it vanished the image or impression was perfectly indivisible.

Minima sensibilia are, for Hume, fundamental constituents of our sensory impressions; but such minima are not the *only* fundamental constituents. Under strict Lockean empiricism, particular ideas are all derived from impressions of sensation and reflection, and general ideas by abstraction of reason applied to these. But Hume disagrees with Locke's claim that the idea of time is derived from reflection upon a "train of ideas". Key to Hume's disagreement with Locke is Hume's 'copy principle' of representation: For Hume, mental representations represent by *copying* the represented: As he (1978, 1.4.5, 233) says, "How can an impression represent a substance otherwise than by resembling it?"⁶

Critically, this includes the "manner(s) of disposition" of the minimal elements of phenomenal experience. Manners of disposition are not simple impressions, but aspects of compound impressions.

The table before me is alone sufficient by its view to give me the idea of extension. [...] But my senses convey to me only the impressions of coloured points, *disposed in a certain manner*. If the eye is sensible of any thing farther, I desire it may be pointed out to me. But if it be impossible to show any thing farther, we may conclude with certainty,

⁶ For the essential discussion of the importance of Hume's copy principle in this context, see Falkenstein 1997, 179–201.

that the idea of extension is nothing but a copy of these coloured points, *and of the manner of their appearance*. (*Treatise* 1.2.3, 34 [my emphases])

The ‘copy principle’ of representation includes temporal dispositions:

Five notes play’d on a flute give us the impression and idea of time; tho’ time be not a sixth impression, which presents itself to the hearing or any other of the senses. Nor is it a sixth impression, which the mind by reflection finds in itself. [...] [H]ere [the mind] only takes notice of the *manner*, in which the different sounds make their appearance [...]. (*Treatise* I.2.3, 36)

For Hume, ideas of spatiotemporal relations (disposition, contiguity, and conjunction in space and time) are uniquely independent, in that they do not derive entirely from aspects of the ideas thereby related.⁷ As Lorne Falkenstein (1997, 193) puts it,

Hume does not take an idea to be an act whereby a mind thinks of or intends a particular object (in this case, a relation of adjacency in space or succession in time). Instead, he takes ideas to be objects in their own right, that represent something else by resembling it. [...]. An idea of time represents time by being itself extended in time.

In his *Stanford Encyclopedia of Philosophy* article on temporal experience, Dainton (2017) states incorrectly that “Hume’s position on the origin of our concept of succession is [...] broadly similar to that of Locke.” For Locke, each moment “perishes” as it passes (*Essay* 2.14.1 and 2.14.18); whereas for Hume, time is composed of non-co-existent moments (*Treatise* 1.2.2), and succession is *perceived*, not reconstructed.⁸ For Hume, our ideas are direct copies, not just of simple impressions, but of *simple impressions appearing in a certain manner in space and time*. Falkenstein (2017, 194) again:

⁷ See Falkenstein 2013, 102–119, 111.

⁸ See Falkenstein 2013.

When Hume's own account of what it means for an idea to represent an object is taken into account, the distinction between an ordered series of perceptions and the perception of an ordered series collapses. [...] [A]n idea of time represents time by enduring over time. [...] [T]o have an idea of succession is to have an idea that takes time to occur, and over which a number of parts occur successively.

So, for Hume, to experience a succession of perceptions is *ipso facto* to have a representation of succession. This would solve the problem of time-awareness for empiricism by accounting for our awareness of temporal succession and duration consistently with empiricist principles: *Pace* Kant, temporal succession and order would be *given* in perception; no a priori principles needed.

The extensionalist Christoph Hoerl (2017, 178) rejects the view that "Perceptual experience [...] is a matter of the perceiver being in a certain kind of representational state, where this is to be conceived of as a state with a content in the sense of a set of veridicality or accuracy conditions." He favors rather a Reidian relational account of perceptual experience, according to which experience is best characterized as a relation of acquaintance with its object (be it a mental or external object). In the context of the relational view, Hoerl sees nothing stopping us from saying we are acquainted with succession, in such a way as we neither must infer succession from memory, nor interpret retained experiential content as representative of the past. The evidence that we are acquainted with succession is, essentially, the fact that we are aware of succession – that is, phenomenal change itself:

On the relational view of experience, there is no representational vehicle the subject instantiates, the particular properties of which fix the contents of her experience. This of course does not mean that questions about the temporal properties of experience don't also arise for the relational view of experience, but they do not involve the idea of two things, the temporal properties of which can either match each other or not, as envisaged by the picture of experiences as involving both a vehicle and a content. The question for the relationalist on her view, rather, is whether the relation of acquaintance in which we stand to objects of experience is most fundamentally a relation in which we stand to objects at a time, or must most fundamentally be seen as something that itself unfolds over intervals

of time. As I have suggested, an extensionalist motivated by the relational view of perception will argue that we need to think the latter because we can perceive instances of succession. As the relation of acquaintance is a generic one in which we simply either do or do not stand to items, perceiving instances of succession must involve standing in such a relation to a succession of such items over time, as they succeed each other. Experiencing goes on over time, on this view, and this is how the items that can figure in experience include not just individual entities that succeed each other, but the very instance of succession in which they partake.⁹

When Hume tells us that we are immediately acquainted with the manners of disposition of events in time, this strongly suggests Hoerl's relationist extensionalism. (Whether a relationist account of perception would be consistent with what Hume intimates elsewhere about mental representation is another matter.¹⁰) Recall Barry Dainton's Immediacy Thesis: Hume's account makes the 'earlier-than' relation a feature of perception, just as phenomenal redness, or the 'to-the-left-of' relation, can be a feature of outer awareness.

A standard objection to the idea that we can be directly acquainted with succession (see, e. g., Meinong's (1978/1899) essay) involves, simply, the idea that, for any span of events supposedly directly 'given' in a single, extended act of perception, part of what is given will be given to one part of the act, and not to another. This general concern is characterized by Dainton (2017) as the Extensional Simultaneity Problem: "How is it possible for contents that are (i) experienced together, and (ii) experienced as present, to be experienced as anything other than simultaneous?" According to this objection, the content-parts of an act of awareness are either presented together (and thus simultaneously) or separately (in which case there are really multiple distinguishable acts of awareness). This basically amounts to insisting that an extended act is divisible into act-parts and their contents – the unification of which under one act is taken for granted by the extensionalist. Kant's approach in the Analogies strongly suggests that he would

⁹ Hoerl 2017, 179 f.

¹⁰ As noted earlier, Hume himself speaks of the way impressions "represent" a substance in Treatise 1.4.5. Also, for Hume, memories are copies of past impressions (Treatise 1.1.3). See, for example Allison (2008). Falkenstein (2017, 49f) suggests that Hume's account of habituation makes room for an experience of the 'past' that is pre-cognitive and does not require unity of consciousness.

agree. Thus, his claim that a *conceptual* grasp of change is a precondition of *experience* of change. (And unity of consciousness too.)

Kant's argument appears to implicitly assume the Extensional Simultaneity Problem. Information about how to represent temporal diversity in one's own states is not part of the raw content of sense-experience; recognition of temporal complexity requires the application of concepts that distinguish between a subjective succession of apprehension and an objective temporal order. (Subjective experience is only comprehensible in relation to a conceptualized objective order.) He is clear in the Aesthetic that the recognition of the temporal complexity of our own states at any given time depends on the status of time as a priori form of intuition:

Here I may add that the concept of alteration, and with it the concept of motion, as alteration of place, is possible only through and in the representation of time; and that if this representation were not an *a priori* (inner) intuition, no concept, no matter what it might be, could render comprehensible the possibility of an alteration, that is, of a combination of contradictorily opposed predicates in one and the same object, for instance, the being and the not-being of one and the same thing in one and the same place. Only in time can two contradictorily opposed predicates meet in one and the same object, namely, *successively*.¹¹ (KrV, B 48 f)

Kant famously notes in the Second Analogy that all inputs are successive in raw form, including the experience of static states of affairs. How would the contents of a given 'extended experience' get their temporal interpretation without a prior conceptual grasp of alteration/causation? That is to say, we need a conceptual grasp of temporal objectivity even to consciously experience the runs of events that we then organize under empirical concepts.

Hume seems to have addressed this problem with the notion of an inherently successive *manner of disposition* as a basic aspect of the (primordial) presentation of a process or sequence of events – primordial in the sense that our awareness of manners of disposition are not inferred,

¹¹ All translations of Kant are from the Guyer-Wood (1999) edition of the *Critique of Pure Reason*.

derived, or abstracted from our experiences of spatial arrays and temporal sequences. (Most scholars I have seen examining this issue believe Kant probably read the *Enquiries* but did not read the *Treatise*; thus, Kant's knowledge of Hume's work on the issue of spatial and temporal awareness may have been limited or nonexistent. Indeed, Kant's failure to in some way confront an extensionalist account of temporal awareness in the *Critique of Pure Reason* strongly suggests he was, indeed, unfamiliar with that chapter of the *Treatise*.) For Hume, we are acquainted with succession because we perceive it tout court (or, alternatively, we perceive it because we are acquainted with it). As Ian Phillips (2014, 149f) puts it, the whole advantage of extensionalism is that it allows us to think in terms of "durations of experience that are explanatorily [...] prior to their temporal subparts." So for the extensionalist there is, supposedly, no issue of needing a conceptual schema for understanding the relations between the subparts in order to understand the whole, as Kant claims. No need for Kant's "recognition in the concept". For extensionalists, perceptual experience of movement/change precedes the "recognition" that a current state represents both past and present inputs. Hoerl thinks Kant may have a point when it comes to "the separate issue as to what is involved in grasping temporal concepts" (i. e., in the developmental sense, not in the sense that temporal concepts are a precondition for coherent experience).¹² But extensionalists claim to be able to account for temporalized experience independent of temporal conceptualization. For extensionalists, perceptual experience of movement/change operates independently of the conscious "recognition" that a current state represents both past and present inputs. After all, non-human, lower-order animals are evidently able to process motion/change and temporal sequence without, presumably, self-aware unity of consciousness (the "I think" of the Transcendental Deduction).

Yet also in this context we should not neglect Kant's comments about alteration in the Refutation of Idealism, along with the clarificatory General Note on the System of the Principles. Commentators – most notably Guyer in his *Kant and the Claims of Knowledge* – have been puzzled as to how to reconcile the anti-skeptical strategy of the Refutation with the general approach to time-awareness in the Analogies. The claim in the Refutation is that self-consciousness as a relatively permanent subject of changing states depends

¹² Private correspondence.

on perception of something permanent (other than the self, which –as Hume famously noted– does not present itself to reflection). This claim is more helpfully elucidated in the General Note than in the Refutation itself; in the General Note Kant explains that the issue is the comprehension of temporal complexity itself:

[I]n order to exhibit *alteration* [...] we must take as our example motion, that is, alteration in space. Only in this way can we obtain the intuition of alterations, the possibility of which can never be comprehended through any pure understanding. For alteration is combination of contradictorily opposed determinations in the existence of one and the same thing. Now how it is possible that from a given state of a thing an opposite state should follow, not only cannot be conceived by reason without an example, but is actually incomprehensible to reason without intuition. The intuition required is the intuition of the movement of a point in space. The presence of the point in different locations (as a sequence of opposite determinations) is what alone first yields to us an intuition of alteration. For in order that we may afterwards make inner alterations likewise thinkable, we must represent time (the form of inner sense) figuratively as a line, and the inner alteration through the drawing of this line (motion), and so in this manner by means of outer intuition make comprehensible the successive existence of ourselves in different states. (KrV, B 291 f)

In the Analogies, Kant stresses cognition's a priori contribution to the possibility of representing a determinate objective (and thus also a subjective) time-order, but in the "Refutation" Kant adds that the a posteriori exhibition of an objective succession (in the form of objective alteration) is a necessary condition of this representation.

The precise nature of this alleged contact with objective succession is difficult to figure out in relation to the argument of the Analogies. In the Analogies, we are asked to imagine an a priori, schematized grasp of causation as a way of relating representations to each other in a determinate time-order. In the Refutation and General Note, objective succession of the states of an object must be experienced in order to be grasped. Experienced

how? In the Analogies we have every indication that Kant is (like Locke) a memory theorist who suggests that all we have ‘before us’ at a time is current input plus recollected content – content which, without interpretation in terms of time, can mean nothing to us.¹³ This, plus Kant’s claim that time-order is not mandated by the content of intuition (“time is not perceived”) is what creates the problem of the representation of a determinate time-order in the first place. The solution he offers is that we are equipped with an a priori conceptual schema that makes interpretation of that state in terms of temporal complexity possible. But if Kant also concedes the empiricist point that a concept like alteration is impossible unless underwritten by sense-experience, we seem compelled to go back to the original problem of time for empiricism – a problem that, in some ways, seems best addressed by Hume’s extensionalism.

In a 2018 paper, philosopher and cognitive scientist Valtteri Arstila (2008, 293) argues that Jamesian ‘specious present’ accounts of temporal phenomenology like retentionalism and extensionalism are not well-supported by the evidence, as empirical studies dispute the “perceptual availability” of just-past stimuli that would be predicted by any specious present theory. Arstila presents evidence for an alternative “dynamic snapshot” theory of time perception. Arstila (2018, 290) cites phenomena like the waterfall illusion and other stationary motion illusions, wherein “we can have an experience of motion without an object appearing to us as being in different places at different times.” (The idea is that stationary motion illusions represent cases in which unmoving images exhibiting some combination of priming and/or figure and visual texture trigger various automatic motion detectors.) Even more intriguing is evidence for non-memory-based experiences of “pure succession”, wherein

¹³ Hoerl (2013) and Dainton (2017), each citing KrV, A 102, join in calling Kant a retentionalist. Kant is not a retentionalist. A 102 explicitly refers to “reproduction” (also “reproduction”, in German) rather than retention; Kant also famously claims that “*as contained in one moment* no representation can be anything other than absolute unity” (KrV, A 99); and the problems of time-awareness throughout the *Critique* can be straightforwardly understood as problems that would arise for a memory theorist. Kant’s main argument for a priori contributions by cognition in the Analogies is based on his observation that “time is not perceived” – that is, time-order is not to be found in the raw content of experience. For the retentionalist, information about time-order is part of the content of one’s experience at a moment.

people can tell that two auditory stimuli are asynchronous when these are presented a mere 2 milliseconds apart. However, if the asynchrony is less than 30 milliseconds, they cannot tell which one of the stimuli was presented first. Given that it is generally agreed that we can have experiences of succession, one way to describe this situation is to say that the two stimuli appeared asynchronously for the subjects simply because the latter have the experience of succession. However, because they do not have the corresponding awareness of which stimulus succeeded the other, this amounts to having an experience of pure succession. (Arstila 2018, 296, citing Hirsh & Sherrick 1961)

Arstila (2018, 290) proposes that “all temporal phenomenology can be explained comparably, by appealing to the existence of automatic, encapsulated, and domain-specific mechanisms.” What Arstila is suggesting is that we are equipped with a variety of motion/change/succession sensors (akin to the basic line-and-orientation sensors Hubel and Weisel famously discovered in feline brains.) Kant’s Refutation of Idealism (in conjunction with his General Note) could be construed as an a priori, transcendental argument demanding the very sort of immediate motion/change detection described by this theory. The dynamic snapshot theory’s account of temporal phenomenology would seem to perfectly satisfy Kant’s requirement that change be directly exhibited in experience. Time-awareness based on motion/succession snapshots obviously avoids the standard Meinongian objection to extensionalism (=the Extensional Simultaneity Problem). The dynamic snapshot theory would not require a relational theory of perception (Locke, Hume and Kant all generally seem to favor a representational theory, with, of course, the possible exception of Hume’s extensionalist account of temporal phenomenology above). Unfortunately for Kant, this vehicle for the generation of the idea of alteration would not contribute to the goal of the Refutation of proving the existence of permanent items outside the mind undergoing alteration. The possibility of experiences of pure succession would also radically undermine the entire rationale for the argument of the Second Analogy, wherein coherent experience rests on the recognition of a determinate time-order in our own states, which in turn rests on the recognition of objective succession — something Kant insists is only possible via the application of a priori rules.

To sum up, Kant's transcendental argument addressing the binding problem doesn't consider the option of Humean extensionalism. The latter solves many problems, but demands we adopt a relationist account of perception. Kant's anti-skeptical argument in the Refutation insists that objective alteration is immediately exhibited in experience, but leaves the delivery system mysterious. The dynamic snapshot theory offers both an alternative to extensionalism and an intriguing way to fill the gap left by Kant; most helpfully, it points the way to further empirical research. A priori reasoning about time perception will be of little further use without a better understanding of how the brain actually processes motion and change.

Wake Forest University, USA

References

- Allison, H. (2008), *Custom and Reason in Hume: A Kantian Reading of the First Book of the Treatise*. Oxford University Press, Oxford.
- Arstila, V. (2018), "Temporal Experiences without the Specious Present", *Australasian Journal of Philosophy* 96, 287–302.
- Arstila, V, & D. Lloyd, D. (eds) (2014), *Subjective Time: The Philosophy, Psychology, and Neuroscience of Temporality*. The MIT Press
- Bardon, A. (2013), *A Brief History of the Philosophy of Time*. Oxford University Press , New York.
- Brook, A. & Akins, K. (eds) (2005), *Cognition and the Brain: The Philosophy and Neuroscience Movement*. Cambridge University Press, Cambridge 208–239.
- Dainton, B. (2000), *Stream of Consciousness: Unity and Continuity in Conscious Experience*. New York.
- Dainton, B. (2017) "Temporal Consciousness", *Stanford Encyclopedia of Philosophy* (last rev. 2017).
- Dyke, H. & Bardon, A. (eds) (2013), *Companion to the Philosophy of Time*. Malden, MA.
- Falkenstein, L. (1997), "Hume on Manners of Disposition and the Ideas of Space and Time", *Archiv für Geschichte der Philosophie* 79, 179–201.
- Falkenstein, L. (2013), "Classical Empiricism", in H. Dyke & A. Bardon (2013), 102–119.
- Falkenstein, L. (2017), "Hume on Temporal Experience", in I. Philips (2017).
- Hoerl, C. (2009) "Time and Tense in Perceptual Experience", *Philosophers' Imprint* 9, 1-18.
- Hoerl, C. (2013), "'A Succession of Feelings, in and of Itself, Is Not a Feeling of Succession'", *Mind* 122, 373–417.

- Hoerl, C. (2017), “Temporal Experience and the Philosophy of Perception”, in I. Phillips 2017: 171–183, 178
- Hume, D. (1978), *A Treatise of Human Nature*, L. A. Selby-Bigge (ed). Oxford.
- Husserl, E. (1964), *Phenomenology of Internal Time-Consciousness*. Bloomington.
- James, W. (1910), *Principles of Psychology, Vol. I*. New York.
- Kant, I. (1999), *Critique of Pure Reason*, P. Guyer & A. W. Wood (eds), London.
- Kelly, S. (2005), “The Puzzle of Temporal Experience”, in A. Brook & K. Akins. (2005), 208–239.
- Mabbott, J. D. (1951), “Our Direct Experience of Time”, *Mind* 60, 153–167.
- Meinong, A. (1978), “On Objects of Higher Order and the Relationship to Internal Perception”, in Schubert Kalsi, M-L. (1978), 137–208.
- Paton, H. J. (1951), *In Defence of Reason*. London.
- Paton, H. J. (1929), “Self-Identity”, *Mind* 38, 312–329.
- Phillips, I. (2014), “The Temporal Structure of Experience”, in V. Arstila & D. Lloyd (2014), 139–158.
- Phillips, I. (ed) (2017), *The Routledge Handbook of Philosophy of Temporal Experience*. Routledge, Oxford.
- Roskies, A.L. (1999), “The Binding Problem”, *Neuron* 24, 7–9.
- Russell, B. (1995), *An Outline of Philosophy*. Routledge, London.
- Schubert Kalsi, M-L. (ed) (1978) *Alexius Meinong on Objects of Higher Order and Husserl’s Phenomenology*. M. Nijhoff, The Hague.

From Here to Eternity: Two Absolutists (Newton, Clarke) and Two Relationalists (Leibniz, Kant of New Elucidation) on the Essence of Time with a Consideration of Their Similarities and Differences¹

VISA HELENIUS

A dispute about absolute and relative time in the eighteenth-century is famous and much discussed.² According to the *absolutist view*, endorsed by Newton and Samuel Clarke (1675–1729), time is an absolute entity, whose existence does not require other beings except God.³ In the *relationalist view*, which is endorsed by Leibniz and the pre-critical Kant, the absolutist position is denied since time is, the relationalists believe, a relational entity, which emerges from relations of beings. Later, Kant developed a groundbreaking theory of time and space in his transcendental idealism: They are the form of human cognition, that is, “a pure form of sensible intuition” and “the form of inner sense” (Kant 1996, 86 and 88). He also combines the aspects of the absolutist and the relationalist view.⁴ Andrew Janiak (2022, chapter 2.1) clarifies that Kant “attempts to engage the Leibnizian “relationalist” and the Newtonian “absolutist” conceptions of space with transcendental idealism”. Therefore, the dispute is relevant to the history of philosophy and especially concerning Newton’s and Leibniz’s philosophies and the background of Kant’s critical philosophy.

¹ This paper has undergone peer review.

² See e.g. Arthur 1994; Ballard 1960; DiSalle 2016; Futch 2008; Janiak 2008; Janiak 2014, xxix–xxxv; Janiak 2022; Melamed 2014; Meli 2014; Rynasiewicz 2022; Slavov 2022; Vailati 1997, 109–137; Yakira 2012.

³ See Janiak 2022, first chapter; Rynasiewicz 2022, third paragraph.

⁴ See Janiak 2022; Kant 1996, 71–75 and 85–104. Kant argues, *inter alia*, that time is not an empirical concept but “the formal a priori condition of all appearances generally[.]” and by it (with the concept of space) “we can draw a priori [...] synthetic cognitions[.]” (Kant 1996, 88 and 92.)

In this article, I examine the dispute by reviewing the views of the above-mentioned four philosophers. My aims are, first, to shed light on the philosophy of time by examining the views, and second, to argue that, in addition to the fundamental difference between the absolutist and the relationalist view concerning the ontological status of time, there are other differences, and that they share some features. The dispute happened within Newtonian classical mechanics and early modern *rationalism*, according to which reason is epistemologically or metaphysically primary in comparison with experience.⁵ Although the discussion about the essence of time in the eighteenth century was intertwined in modern physics and theories of force, I will focus here on ontological questions in the rationalistic framework. Regarding Kant, I discuss his writing *A New Elucidation of the First Principles of Metaphysical Cognition* (*Principiorum primorum cognitionis metaphysicae nova dilucidatio*, 1755)—hereafter “New Elucidation” for short. He revises two powerful metaphysical principles, the principles of contradiction and sufficient or determining reason, and offers noteworthy points about the philosophy of time.

First, since the subject matter of time is a complex whole,⁶ I propound a brief contextualization of the problematic nature of the philosophy of time. I consider this reasonable also because this article is a part of the *Festschrift* celebrating Professor Valtteri Arstila’s fiftieth birthday, and temporal experience is one of his central research interests.⁷ Furthermore, the contextualization also connects the dispute loosely to Kant’s critical philosophy. After this, I examine Newton’s, Clarke’s, Leibniz’s, and Kant’s views. Lastly, I propound critical remarks about their differences and similarities.⁸

1. What is time?

Augustine’s realization of time goes as follows: “[W]hat is time? If no one asks me, I know; if I want to explain it to someone who asks me, I do not know.” (Augustine 2019, 210.) This observation is insightful⁹ as it hits the nail on the

⁵ For more on rationalism, see Huenemann 2008.

⁶ See Emery, Markosian, and Sullivan 2020.

⁷ See e.g. Arstila 2018; Arstila 2022.

⁸ I am grateful to two anonymous referees for their comments and the editors of this *Festschrift*.

⁹ See e.g. Aristotle 1957, 373–427; Melamed 2014; Plato 1929, 75, 77, 79, 81 and 83. For an overview about philosophy of time, see Emery, Markosian, and Sullivan 2020. For more on Augustine’s conception of time, see Hausheer 1937.

head, so to speak, regarding the subject matter: “Time” is, I believe, by nature something, which is, on the one hand, familiar to us but, on the other hand, surprisingly difficult to explain.¹⁰ To be precise, time is experienced in normal circumstances¹¹ as the passage of time.¹² Accordingly, we can distinguish our past experiences from our present experiences and, because of this, we can effortlessly use time related indexicals, such as “yesterday” and “now”.¹³ Despite our temporal experiences, the theorizing of time has turned out to be difficult. One of the reasons is that a question about the essence of time leads to some of the most fundamental philosophical issues, such as what kind of ontological status time has, and how temporal experience, perception and consciousness are related to one another.¹⁴

The philosophy of time arises from experience of *change* and *permanence*. For example, I see a withered flower, which was blooming earlier. In that case, the flower as an object of perception has certain permanence over time (granting the flower is one and same flower at various times), whereas its perceptible states (blooming, withering) indicate its variable properties. Along with these kinds of observations, one could deduce that our experience of time emerges because of the permanent and variable things in the world. Nonetheless, this duality may not be true. For example, Parmenides has claimed that change is impossible as everything is immutable by essence,¹⁵ and Heraclitus has denied not only the Parmenidean view but all permanent features of the world since, for him,

¹⁰ See Melamed 2014.

¹¹ In certain psychological disorders, such as some forms of schizophrenia, (normal) temporal experience is disturbed. An important philosophical question is to what extent these experiences reflect reality.

¹² For more on the passage of time, see Maudlin 2007, 104–142; Styrman 2023.

¹³ Plato describes, in my view strikingly, time and tenses as follows: “For simultaneously with the construction of the Heaven [the Demiurge, i.e. the divine craftsman] contrived the production of days and nights and months and years, which existed not before the Heaven came into being. And these are all portions of Time; even as “Was” and “Shall be” are generated forms of Time, although we apply them wrongly, without noticing, to Eternal Being. For we say that it “is” or “was” or “will be,” whereas, in truth of speech, “is” alone is the appropriate term; “was” and “will be,” on the other hand, are terms properly applicable to the Becoming which proceeds in Time, since both of these are motions; but it belongs not to that which is ever changeless in its uniformity to become either older or younger through time, nor ever to have become so, nor to be so now, nor to be about to be so hereafter, nor in general to be subject to any of the conditions which Becoming has attached to the things which move in the world of Sense, these being generated forms of Time, which imitates Eternity and circles round according to number.” (Plato 1929, 77.)

¹⁴ See Arstila 2018; Arstila 2022.

¹⁵ For more on Parmenides’ arguments, see Laks and Most 2016b, 43, 45, 47, 49 and 51.

everything is fluid by essence.¹⁶ In that case, either our experiences of the passage of time and change are illusory, or everything is changing all the time, in which case our experiences of temporally persistent things are false.¹⁷ Nevertheless, both possibilities are problematic because there obviously are—I return now to the dualistic view—both variable (e.g. mental states, blossoming and withering of flowers) and lasting (my existence, ordinary concrete objects, such as stones and cats) objects in the world.¹⁸

Kant's problematization in *The Critique of Pure Reason* is as follows:

Time cannot be intuited [*anschauen*]¹⁹ outwardly, any more than space can be intuited as something within us. What, then, are space and time? Are they actual beings? Are they only determinations of things, or, for that matter, relations among them? If so, are they at least determinations or relations that would belong to things intrinsically also, i.e., even if these things were not intuited? Or are they determinations and relations that adhere only to the form of intuition and hence to the subjective character of our mind, so that apart from that character these predicates cannot be ascribed to any thing at all? [Kant 1996, 77.]

Kant highlights the problematic nature of time well: It is not obvious by experience whether time is a real being or an inner thing of human perception and cognition? Consequently, the problem requires a deeper analysis. Kant's own answer to the question in *Inaugural Dissertation* is, I believe, generally illustrative concerning the essence of time:

¹⁶ For more on Heraclitus' dynamic monism, see Laks and Most 2016a, 159, 169, 171, 177, and 179.

¹⁷ See Laks and Most 2016a, 153, 155 and 157.

¹⁸ Regarding the dualism, Plato (1929, 49) has famously combined variability and permanence of the world as follows: There is a primary permanence behind perceived change, so that these invariant things are "really existents", whereas variable things are secondary and objects "of opinion with the aid of unreasoning sensation". (See Plato 1929, 49, 51 and 53.) For him, variable things are temporal and secondary in the ontological hierarchy, whereas invariable things are primary in the hierarchy and atemporal as they exist always. The first also are subordinate to causality and experience and reflect the permanent things.

¹⁹ "[I]ntuition is that by which a cognition refers to objects directly, and at which all thought aims as a means. Intuition, however, takes place only insofar as the object is given to us; but that, in turn, is possible only [...] by the mind's being affected in a certain manner." (Kant 1996, 71 and 72.) See also Kant 1996, 366. According to Janiak, "intuition" is for Kant a conscious, objective representation, and it differs from sensation (Janiak 2022, chapter 2.2).

Time is not something objective and real, nor is it a substance, nor an accident, nor a relation. Time is rather the subjective condition which is necessary, in virtue of the nature of the human mind, for the coordinating of all sensible things in accordance with a fixed law. It is a pure intuition. [Kant 1992b, 393.]²⁰

According to the passage, time is a prerequisite for human experience. It is, therefore, not a real thing, that is, “objective and real, nor [...] a substance, nor an accident, nor a relation”, but a condition and foundation of the experience itself. Janiak (2022, chapter 1) explains that time is “causally inert” for Kant. I add that this feature can be seen as a distinctive feature of time in general since I acknowledge that time is a thing, which is clearly experienced, but which does not have causal power—it is not the case that time causes our ageing but that we are getting older in time *by various factors*. Accordingly, time is a mind-dependent entity in Kant’s view, as temporal experience requires conscious being. In that case, the analysis of time concerns the human mind and its operations essentially. Furthermore, time seems to be a *sui generis* entity, let it be real or not, which deserves special attention.²¹

Regarding the peculiarity of time, the distinction between *absoluteness* and *relationality* of time is important, which is a theme of this article. First, if time is an absolute entity, its existence does not require other things, except, possibly, God. However, absolute time does not have to be a substance, in which case it has ontological independence to a lesser extent. Second, (1) if time is a relational entity, then the existence of many things is required because relational things always depend on the existence of other things—the property of relationality requires many existing things, so that there can be relations altogether.²² Furthermore, in relational time (2) the passage of time is defined as *the change in the objects in space*. Regarding this, Leibniz’s description of relational time is as follows:

[Relational time] is the order of existence of things possible successively. As a physical body is to space, so the status or series of things is to time. The

²⁰ Cf. Kant 1996, 73.

²¹ See Futch 2008, 30.

²² I clarify that an absolute entity can have inner relations, so that it is an independently existing compositional thing.

body and the series of things add to space and time, motion or action and passion, and the principle of motion. [Leibniz 1989, 536.]²³

2. Newton on the Essence of Time

Newton was against Descartes' view of space.²⁴ Regarding space and time, he endorses the absolute view, according to which both are absolute entities. What is Newton's "absolute time"? He argues in *On Gravitation (De gravitatione*, date unknown) that

[1] [*a*]bsolute, true, and mathematical time, in and of itself and of its own nature, without reference to anything external, flows uniformly and by another name is called duration. *Relative, apparent, and common time* is any sensible and external measure (precise or imprecise) of duration by means of motion; such a measure – for example, an hour, a day, a month, a year – is commonly used instead of true time. [...] [2] In astronomy, absolute time is distinguished from relative time by the equation of common time. For natural days, which are commonly considered equal for the purpose of measuring time, are actually unequal. Astronomers correct this inequality in order to measure celestial motions on the basis of a truer time. It is possible that there is no uniform motion by which time may have an exact measure. All motions can be accelerated and retarded, but the flow of absolute time cannot be changed. The duration or perseverance of the existence of things is the same, whether their motions are rapid or slow or null; accordingly, duration is rightly distinguished from its sensible measures[.] [Newton 2014, 84 and 85. Italics added.]

Newton begins here by describing absolute and *relative time*. Most importantly, time is, for him, absolute in essence; it is that kind of entity, which does not depend on bodies and their effects. Absolute time is also a *dynamic* entity in that sense that it flows or passes uniformly everywhere, and a *quantitative* entity as it is "mathematical", which means that it is measurable. Furthermore, absolute time can be identified with "duration", which refers to the continuity of things, time

²³ See also Ballard 1960, 61.

²⁴ See DiSalle 2016; Janiak 2008, 130–162; Janiak 2014, xviii–xx; Rynasiewicz 2022, third chapter.

intervals, and the amount or quantity of absolute time.²⁵ However, Newton points out that absolute time is understood *relatively* in everyday life by observations and measurements of motion, movement, permanence, and change. For example, let us think about the function of the hourglass: The flowing of sand in various hourglasses depends on the volumes of their glass bulbs and the amount of sand. Therefore, there are various kinds of relative measurement methods of absolute time by these devices. Nevertheless, Newton admits that the measurement of motion may not ever be an exact measurement method of absolute time since movement is always variable but the flow of time constant since, as he argued, the flow of absolute time cannot ever be changed.

I sum up that Newton's absolute time is a dynamic and non-causal entity, which is manifested by permanence and change in the world, and a homogenous or uniform entity since the flow of time happens identically everywhere. However, absolute time is understood relatively and is measured by relative methods, from which follows the common conceptions of relative times.²⁶

Newton's view raises questions, especially philosophical, such as "What kind of entity is absolute time in essence?" Accordingly, many philosophers criticized his view.²⁷ One reason for this is that Newton was principally a mathematician, physicist, and natural philosopher and not a metaphysician who develops an exact metaphysical system, like Spinoza and Leibniz. Regarding the criticism, Robert DiSalle (2016, 41) points out that Newton's aim is only to describe *the structure of time* by his theory—the characteristic of the Newtonian absolute time is not "the distinct individuality of its moments" but its temporal equality and objectivity—and that he gives a definition of absolute time within his mechanics and, in fact, does not try to prove it. As for Newton's defence, Janiak (2008, 9 and 139) argues that Newton is, in the end, "systematic and philosophical without articulating a philosophical system[.]" and that his "philosophical critics did not meet Newton on his own ground – they did not determine what conception of space, time, and motion was presupposed by, or fitted most closely with, Newton's own laws of motion." Consequently, one question is how well Newton's

²⁵ Newton's view of absolute space and time has also been called *substantivalism* (in contrast to Leibnizian relationism) (Rynasiewicz, 2022, third paragraph; Slavov 2022, 13). Concerning *primitivist* accounts, according to which the passage of time is primitive, see Maudlin 2007, 104–142; Styman 2023.

²⁶ See Rynasiewicz, 2022, first paragraph and chapter 5.1.

²⁷ See DiSalle 2016, 34; Janiak 2008, 139.

views were understood by the critics.²⁸ Regarding this, Leibniz's and Clarke's correspondence is one significant source.

3. Leibniz, Clarke, and the Dispute About the Essence of Time

During Newton's lifetime, Clarke and Leibniz disputed about the essence of time in their famous correspondence (1715–1716). Clarke defends Newton's view, whereas Leibniz rejects it and endorses the view of relational time. Interestingly, both agree with *the principle of sufficient reason* and use it in their argumentation. According to the powerful principle, everything has a reason or cause.²⁹ The principle is associated with rationalism³⁰ since the intelligibility of the world, which is a premise of rationalism, requires that everything must have a reason or cause, which explains its existence.³¹ It follows that time must be an explicable entity. At this point, it should be noted that Leibniz's and Clarke's views represent *moderate rationalism* since both accept a posteriori knowledge,³² and that Newton's connection with the principle and rationalism is complex.³³

I begin with Leibniz's view since he offers his argument first in the correspondence. Leibniz's endorsement of relational time and main argument against absolute time are as follows:

[1] As for my own opinion, I have said more than once that I hold space to be something merely *relative* [*relatif*], as time is: that I hold it to be an order of co-existences as *time is an order of successions*. [...] [2] Supposing anyone should ask why God did not create everything a year sooner, and the same person should infer from thence that God has done something concerning which 'tis not possible there should be a reason why he did it so and not otherwise; the answer is that his inference would be right if time was

²⁸ See Janiak 2008, 1–10; Janiak 2014, vii–xiii and xii–xxvi.

²⁹ See Leibniz 1989, 677, 678, 680, and 683–685. It should be noted that the principle is somewhat ambiguous; for example, Leibniz uses the word “reason” (*ratio*) and Clarke “reason” and “cause” (see Vailati 1997, 81), but these may be not identical as “reason” may be more extensive regarding, in addition to events, also propositions.

³⁰ See Della Rocca 2010.

³¹ It is to be noted that the principle has variants (see Melamed and Lin 2023, first chapter). For more on the principle, see Della Rocca 2010; Melamed and Lin 2023.

³² Concerning Leibniz, see Huenemann 2008, 1–4; Leibniz 1989, 645 and 646. Concerning Clarke, see Vailati 1997, 81; Yenter and Vailati 2021, chapter 2.2.

³³ See DiSalle 2016, 55 and 56; Janiak 2008, 49, fn. 87; 138, and 165. For interpretations of Newton's views, see DiSalle 2016; Janiak 2008, 11–49.

anything distinct from things existing in time. For it would be impossible there should be any reason why things should be applied to such particular instants rather than to others, their succession continuing the same. But then the same argument proves that instants, considered without the things, are nothing at all and that they consist only in the successive order of things, which order remaining the same, one of the two states, viz., that of a supposed anticipation, would not at all differ, nor could be discerned from the other which now is. [Leibniz 1961, 363; Leibniz 1989, 682 and 683. Italics added.]

Leibniz claims, first, that time and space are relational entities.³⁴ In this view, time emerges by means of the relations of actual (temporal) beings. Therefore, I use the term “relationalism” in this article. However, according to Michael Futch (2008, 30), Leibniz’s view can also be called *reductionism* because, Futch argues, “[s]pace and time are not independently existing entities or structures in their own right, and spatial and temporal facts are not grounded on facts about such independent structures.” In relational time, existence of two or more beings are required, so that time can emerge because one existing thing could not have external relations at all—Kant deals with this idea more closely, as we will see below. Importantly, it is to be noted that time is not a real being for Leibniz: It is secondary to individual substances or *monads*, which are basically ideal and unextended beings.³⁵ Ballard (1960, 61) clarifies that “[s]pace and time are definitely not relations among individual monads[.]” Namely, it is true that the world consists of the ideal monads, which are real beings, external to God, and causally “windowless”,³⁶ but the visible world including space and time is a phenomenon, which emerges from the fundamental monadic level. In other words, relational time is manifested only in the phenomenal world, which differs from the fundamental monadic reality.

Second, Leibniz argues in the passage against absolute time that if time would be absolute, as Newton and Clarke believe, then God could not have a

³⁴ See Leibniz 1989, 703–705. See also Arthur 1994; Emery, Markosian and Sullivan 2020, chapter 2; Futch 2002, 1 and 2. For more on Leibniz’s view of time and its interpretation and problems, see Futch 2002; Futch 2008.

³⁵ See Leibniz 1989, 636, 643–645 and 647–649. For more on Leibniz’s *phenomenalism*, see Adams 1994, 217–261.

³⁶ See Leibniz 1989, 643. Monads operate in accordance with *the pre-established harmony* (see Leibniz 1989, 651).

sufficient reason for the moment of the creation concerning God’s chosen order of succession of contingent beings—these are possible and actual but not necessary—since absolute time is in essence a homogeneous and uniform entity, and, consequently, all different moments of time are equal. Here, Leibniz (1989, 687) uses *the principle of indiscernibles*, according to which “[t]here is no such thing as two individuals indiscernible from each other”. Therefore, Leibniz thinks that the moment of creation would be arbitrary for God in absolute time since every moment of time would be perfectly similar, but this is not compatible with God’s freedom, omniscience, and perfect rationality.³⁷

In my view, Leibniz’s argument is ingenious: It seems plausible that if contingent existence precedes absolute time, so that the homogenous and uniform time is prior to contingent (temporal) beings, then God cannot have a sufficient reason for the moment of creation—there cannot be, simply, a rational reason for God’s choice since every moment of time is basically indistinctive in absolute time. The second strength of Leibniz’s view is as follows: It is possible that the Newtonian absolute time, which is a *sui generis* ontological category, is an unnecessary metaphysical postulate. If so, it can be rejected by Occam’s razor, according to which unnecessary postulates must be rejected as a simpler theory is preferable.³⁸ Leibniz’s view does not have this problem since relational time is not a real being but a phenomenon.

Regarding Clarke, there is a consensus that he was Newton’s follower,³⁹ and he undoubtedly echoes Newton in the correspondence by arguing for the absoluteness of time. For example, he states that “the parts of time are as exactly alike to each other as those of space; yet two points of time are not the same point of time, nor are they two names of only the same point of time” (Leibniz 1989, 692) and claims explicitly that time is “*absolute* and unvaried [q]uantity” (Leibniz 1961, 430. Italics added).

To begin with, Clarke describes time—to be precise, *duration*—as follows: It is an immense, immutable, eternal, and necessary consequence of the existence of God (Leibniz 1989, 692 and 693), so that (absolute) space and time are not separate from God. They are namely, he explains, “caused by and are immediate and necessary consequences of his existence. Without them [God’s] eternity and ubiquity (or omnipresence) would be taken away” (Leibniz 1989, 693).

³⁷ See Leibniz 1989, 639, 640, and 646–648.

³⁸ For more on Occam’s razor in the philosophy of time, see Maudlin 2007, *passim*.

³⁹ See Yenter and Vailati 2021, chapter 2.1.

Regarding this view, the following point is important: God is not, for Clarke, a temporal being, but time follows from his existence. Clarke explains elsewhere in the correspondence that space is a *property*, and that it cannot be identified with God. I remark that the same applies, naturally, to time. Therefore, time and the existence of God are fundamentally intertwined in Clarke's (and Newton's) view: Time is not an essential property of God, like his omnipotence and eternity, not at least in the strict sense—apparently, God would exist even if there would be no contingent beings or time. Instead, time is that kind of independent and consequential property, perhaps a logical accident (i.e. contingent property) or a predicable (“*x* is a temporal being”),⁴⁰ which is a necessary consequence of God's existence. In other words, God is not a temporal being, but temporality and (absolute) time follows from his existence, in which case they, so to speak, go hand in hand.

Clarke's counter argument against Leibniz's relational time goes as follows:

[I]f time was nothing but the order of succession of created things [like Leibniz believes], it would follow that if God had created the world millions of ages sooner than he did, yet it would not have been created at all the sooner. Further, space and time are quantities, which situation and order are not [Leibniz 1989, 685].

First, Clarke argues that temporal determination and scheduling would not be possible for God without absolute time; if time would be a relational entity and merely the order of succession of contingent things in the Leibnizian manner, God would not have been able to create the world at his desired moment of time. It seems, therefore, that relational time is conceptually possible for Clarke, but then, as he obviously thinks, God cannot have knowledge of the moment of time of creation. In other words, there would not be a sufficient reason for it. To be precise, God would not have temporal knowledge and a conception of transcendental time—this refers to temporality before the creation—in relational time, despite the created beings themselves having temporal relations *ex hypothesi*. Consequently, Clarke concludes that time must be absolute. Second, he specifies that time is a quantity, like space, but situation and order are not. I interpret that time is a measurable entity for him, like for Newton, whereas situation and order are primarily qualitative entities: Although situation and order are compositional

⁴⁰ Cf. Baumgarten 2014, 109 and 112.

entities, they form qualitative unities. For example, it is not reasonable to ask which of two situations or orders is greater or lesser (in the quantitative sense).

To sum up, Newton and Clarke endorse the view of absolute time (or duration). According to the first, absolute time is a quantitative, dynamic, and homogenous or uniform entity but not a substance, and it is experienced relatively and can be measured by relative methods. The latter clarifies that time is absolute, immense, immutable, eternal, property, quantity, and a necessary consequence of God. In that case, time follows from the existence of God, but he is not a temporal being himself. In this view, time precedes bodies and has, thus, ontological independence. As for Leibniz, he disagrees with them and argues for relational time: Time emerges along with contingent beings (to be precise, phenomenal bodies) and depends completely on theirs and God's existences. Both Leibniz and Clarke accept the principle of sufficient reason, which entails that time must be an intelligible entity as everything is explicable by the principle. Accordingly, the dispute happens within the framework of rationalism. According to Leibniz's argument, the Newtonian absolute time cannot be true since then God could not have a sufficient reason for the moment of time of the creation as absolute time is indistinguishable in essence. According to Clarke's counterargument, the Leibnizian relative time cannot be true because then God could not perform temporal determination for the creation in absolute time—God would lack, respectively, a sufficient reason for the creation concerning its temporal determination. Hence, both use the principle in their argumentation. In my view, Leibniz's argument has the following strengths for the two above-mentioned reasons (i.e. Leibniz's argument against absolute time and the possibility that absolute time is an unnecessary postulate). However, although Newton's aims were more practical and scientific, his view is not straight away philosophically worse, as Janiak (2008, 9 and 139) and DiSalle (2016, *passim*) have remarked.

4. Kant on Time in *New Elucidation*

Regarding Kant's philosophical development, he prepares the ground for his critical philosophy in his pre-critical period, in which he also deals with the philosophy of time.⁴¹ *New Elucidation* is an early writing of this period, and he discusses Leibniz's, Christian Wolff's (1679–1754), Alexander Gottlieb Baumgarten's (1714–1762), and Christian August Crusius' (1715–1775) views

⁴¹ See Kant 1992b, *passim*.

and revises the principles of contradiction and sufficient or determining reason in the rationalistic framework. Regarding the philosophy of time, Kant offers some additions and clarifications to the Leibnizian view of relational time. Especially a derivative of the principle of determining reason, *the principle of succession*, is metaphysically truly relevant, as we will see below.

Kant's principle of determining reason is a hybrid of Leibniz's, Wolff's, and Crusius' views of the principle of sufficient reason; he aims to improve their views of the principle by revision. Kant's own view is based on the concepts of *determining* and *ground*, which he (1992a, 11) explains as follows: "[D]etermining" is "to posit a predicate while excluding its opposite", and "ground" is "[t]hat which determines a subject in respect of any of its predicates." Thus, Kant endorses the Leibnizian *subject-predicate structure*, according to which all truths are analytic, so that all possible predicates (e.g. wisdom) of the subject A (e.g. Socrates) belongs the concept of A (e.g. the complete individual concept of Socrates).⁴² The principle itself states that everything has a determining ground, which is either *antecedently* or *consequentially determining ground*: The first, "the ground of being or becoming", enables the intelligibility of the determined beings, and the latter, "the ground of knowing", enables knowledge about them. (Kant 1992a, 13.) Kant argues elsewhere in *New Elucidation* that every contingent being must have an antecedently determining ground.⁴³

Kant discusses several things in *New Elucidation*, which are related to the philosophy of time. First, he claims that the principle of determining ground holds true as "[n]othing is true without a determining ground" (Kant 1992a, 13. Italics removed). This implies, I note, that also time is an intelligible entity as temporal phenomena of the world must have determining grounds (i.e. the grounds of being or becoming and knowing for temporal experiences and change). Indeed, Kant explains the existence of time by the principle of succession, which is a derivative of the principle of determining ground.

Second, Kant argues that nothing has the ground or cause of its existence within itself since, then, a thing would be its own ground or cause. This is impossible because, Kant (1992a, 14) explains, "the concept of a cause is by nature prior to the concept of that which is caused, the latter being later than the former, it would follow that the same thing would be simultaneously both earlier and later

⁴² See e.g. Adams 1994, 57–63; Leibniz 1989, 267–269, 310 and 311. Cf. Kant 1996, 51. Kant's own description is as follows: "Every true proposition indicates that the subject is determinate in respect of a predicate" (Kant 1992a, 13).

⁴³ See Kant 1992a, 17–20.

than itself.” This does not require a temporal order between causal things necessarily, but Kant’s view has, however, a connection with temporality as follows: A thing, which causes itself, would have two categorically separate states, a state when it is a cause for itself (whatever this means, like “a being plans its own forthcoming existence”), and a state when it is a caused beings by its own causation. Concerning this subject, Kant (1992a, 14) deduces that God, who is a necessary being, does not have an antecedently determining ground since he exists due to necessity; in that case, God is eternal.

Third, Kant gives a short argument for the existence of one God, which has a connection to the philosophy of time. My shortened version of it is as follows: (1) Let us suppose that finite contingent beings would exist only in time, so that they would form together an entire contingent reality, and that there would be no absolute necessary being—the traditional view is that an absolute necessary being contains all possibilities and is, thus, the origin of all contingent *realities*.⁴⁴ In the above-mentioned case, the contingent existences, according to Kant (1992a, 16), “would be a repetition made a number of times[.]”(2) This is, nonetheless, untrue since the idea is against the concept of absolute necessity: Only one absolute necessary being can guarantee the reality of possible beings consistently. (3) It follows that only one absolute being exists necessarily, God. Now, let us pay attention to the description “a repetition made a number of times”. This refers to that kind of contingent existence, which would not exist due to an absolute necessary being. Regarding the subject matter of time, the contingent existences in question can consist of, I argue, repeatable temporal beings. Then, the same contingent beings and events could exist repeatedly, such as Plato writes *Republic* repeatedly in various times (I return to this idea below). Kant’s surely would deny this: Repeatable temporal beings are not possible since the existence of the absolute necessary being prevents their possibility for the above-mentioned reason, that is, only one absolute necessary being exists certainly but contingent beings do not, in which case they need a ground for their consistent existence, which is one and only one necessary being.

Fourth, Kant (1992a, 37) introduces the principle of succession in the last part of *New Elucidation*. According to it, “[n]o change can happen to substances except in so far as they are connected with other substances; their reciprocal dependency on each other determines their reciprocal changes of state.”⁴⁵

⁴⁴ See Leibniz 1989, 646 and 647.

⁴⁵ Cf. Baumgarten 2014, 144.

Therefore, change is possible only by the mutual interaction of substances. This implies of Kant's endorsements of *substance pluralism*, according to which many substances exist,⁴⁶ and a view that interaction between substances can only enable change and the passage of time. Indeed, Kant concludes that, because of the principle, "if the connection of substances were cancelled altogether, succession and time would likewise disappear" (Kant 1992a, 37). In that case, time is possible by means of the existence of many substances and their mutual interaction. If only one substance would exist, it would be a completely static and unchanged being (Kant 1992a, 37).

Kant also deduces two important things by the principle of succession, which are related to the philosophy of time: (1) There must be (external) bodies, which enable our inner changes, and (2) God is essentially immutable since he is free from dependency including external relations and an organic body. (Kant 1992a, 39.) In sum, by the principle of succession, the existence of time follows from interaction between substances, the external world and bodies are true, and God is an immutable and unextended being.⁴⁷

5. On Similarities and Differences Between the Absolutist and the Relationalist View of Time

Lastly, I propound comparative remarks about the views of Newton, Leibniz, Clarke, and Kant. I begin with *objectiveness* and *subjectiveness of time* and their connection with the dispute. I agree with the Augustinian point of view, according to which the passage of time is experienced subjectively as conscious temporal experiences, which presupposes perceiving conscious subjects.⁴⁸ However, since time is also a common and shared object, it has an objective and an intersubjective aspect. Consequently, there is scientific knowledge (e.g. different species operate within different time scales) and terminology (e.g. the standardized units of time) about time, in which physics (e.g. the concepts of space-time and arrow of time) and subjective (the phenomenology of time) and intersubjective (e.g. necessary conditions and the concept of time) analysis of time are related. Regarding the dispute, I conclude that the absolutists and the relationalists conceive of time as an

⁴⁶ See Kant 1992a, 40–45.

⁴⁷ Regarding space, see Kant 1992a, 40–45.

⁴⁸ Our temporal experience may be intuitive knowledge (like the proposition "I exist", which is, to be sure, the most foundational knowledge). See Kant 1996, 88; Kant 1992b, 393. However, cf. Arstila 2018.

objective entity regardless of its ontological status: Although time is experienced subjectively and has relative features, such as it can be measured by relative methods, like by hourglasses, it is, above all, a shared object in both standpoints; time is for Newton a scientific object, and for Leibniz a phenomenal object. However, I emphasise that Newton deals with time more objectively by his technical terms and laws of motion of his mechanics.⁴⁹

Second, I return to the duality of change and permanence. The absolutists and the relationalists would obviously accept that these features belong to the world. Nevertheless, the first consider time as a dynamic entity, which encompasses temporally enduring beings, and which manifests change indirectly and is “causally inert”. As for the relationalist view, time completely depends on the existence of God and the interaction between bodies, so that change and time result from the existence of substances and their internal activity.⁵⁰ Then, time is an epiphenomenon of monadic (i.e. substantial) primary existence and, likewise, causally inert. Now, the causality of the world is temporally oriented by experience (this is, however, not necessarily true, as Parmenides has argued), so that the order of events in the passage of time is irreversible. As far as I can tell, both the absolutists and the relationalists agree with this: The first because absolute time makes its way only in one direction, and the latter because God has created the dynamic substances, of which follow the dynamic and variable world.

Third, the absolutists and the relationalists accept the distinction between contingent and necessary beings.⁵¹ Regarding the philosophy of time, if a necessary thing is a temporal being, it exists at every moment of time, whereas a contingent temporal being exists only at some moments of time.⁵² Alternatively, a necessary thing can be atemporal and immutable, as we have seen. Both postulate the necessary God and the contingent world. However, time is necessary for the absolutists in that sense that absolute time follows from the existence of God; according to Clarke, the existence of God and time are intertwined since time is a

⁴⁹ See Janiak 2008, 139.

⁵⁰ See e.g. Leibniz 1989, 643 and 644; Kant 1992a, 32–34 and 37–45.

⁵¹ Both contingent beings (e.g. Plato) and necessary beings (e.g. arithmetic truths, God) are possible as they are not contradictory themselves, but only the denial of the latter leads to contradiction. In other words, contingent and necessary things are possible, but the first could have been merely a possibility, that is, a non-actual possible thing. Regarding time, if a necessary thing is a temporal being, it exists in every moment of time, whereas a contingent temporal being exists only at some moments of time.

⁵² One question is whether or in what way God, if he exists, is a temporal being. For more on this question, see Ganssle 2024; Spinoza 2002, 182–186.

“immediate and necessary” consequence of God (Leibniz 1989, 693), as I explained above. Regarding Leibniz’s view, time is, on the one hand, a contingent entity in that sense that the actual world, *the best possible world*, is not necessary.⁵³ On the other hand, time is necessary in that sense that monads are internally changing entities (Leibniz 1989, 643–645). The same applies to Kant’s view, albeit more distinctly: Every possible world with two or more monadic substances is necessarily mutable and temporal, as we saw above.

Fourth, I discuss the possibility of repeatability of beings in time. In metaphysics, entities are thought to be unique, that is, particulars, such as individual beings (e.g. Plato or Da Vinci’s Mona Lisa excluding all replicas of the painting, however identical they may be to the original one) or repeatable, that is, universals (e.g. redness, which is instantiated in many objects). Regarding the philosophy of time, I claim that it is possible, correspondingly, that there are *unique* and *repeatable temporal beings*: The first are temporally unique because they have unique existence in time. For example, if time has only one direction (“the arrow of time”⁵⁴) and dimension, individual (temporal) beings can only be temporally unique since every moment of time is a unique event or situation. Consequently, physical processes, *inter alia*, are irreversible. As for repeatable temporal beings, they can exist at various times. Regarding this possibility, one thing is clear in Newton’s case: If individual beings, such as Socrates, are repeatable in time granting their basic conditions are not violated,⁵⁵ such as internal consistency and individuality, these have different temporally determined existence inevitably due to the nature of absolute time, which is a dynamic and “flowing” entity in essence. The case is different in Leibniz’s case: It seems that Leibniz’s God can repeat, at least in principle, *compound individual beings* of the actual world, such as Socrates, by creating and destroying them again and again since time is merely a consequence of active substances;⁵⁶ the world is eternal for Leibniz at least in that sense that monads are ideal beings and, possibly, eternal in a sense—every monad has origin but not necessarily end—but monadic compounds, such as stones and cats, have temporary existence.⁵⁷ It seems,

⁵³ See Leibniz 1989, 648.

⁵⁴ Cf. Kant 1996, 259–276.

⁵⁵ Concerning identity, Kant denies Leibniz’s principle of indiscernibles, according to which every external and internal property of an individual is essential for it (see e.g. Leibniz 1989, 307 and 308; cf. Adams 1994, 106–109); accordingly, internally identical things can exist simultaneously (Kant 1992a, 35 and 36).

⁵⁶ For more on this subject matter, see Adams 1994; Koistinen and Repo 1999.

⁵⁷ See Leibniz 1989, 636. Cf. Leibniz 1989, 643.

therefore, that Leibniz's God can arrange substances in such a way that the same compounds are repeated because there is no absolute time, which would bind contingent beings in the flow of time, like in Newton's case.

I conclude that although the absolutist and the relationalist view of time are not compatible by their fundamental differences, they share many points of contact, such as that time is not a substance and that the principle of sufficient (or determining) reason is needed for the demonstration of its essence. Differences between the standpoints also are occasionally delicate, such as the relative features of time.

University of Turku

References

- Adams, R. (1994), *Leibniz – Determinist, Theist, Idealist*, Oxford University Press, Oxford.
- Aristotle (1957), *Physics*, translated by F. Macdonald Cornford and P. Wicksteed, Harvard University Press, Cambridge, Massachusetts, and London.
- Arstila, V. (2018), "Temporal experiences without the specious present", *Australasian Journal of Philosophy* 96(2), 287–302.
- Arstila, V. (2022), "Ajankulun ja kokemisen arvoituksellisuudesta", *Ajatus* 79, 201–223.
- Arthur, R. (1994), "Space and Relativity in Newton and Leibniz", *The British Journal for the Philosophy of Science* 45(1), 219–240.
- Augustine (2019), *Confessions*, translated by T. Williams, Hackett, Indianapolis, Cambridge.
- Ballard, K. (1960), "Leibniz's Theory of Space and Time", *Journal of the History of Ideas* 21(1), 49–65.
- Baumgarten, A. (2014), *Metaphysics*, translated by C. Fugate and J. Hymers, Bloomsbury, London, New Delhi, New York, and Sydney.
- Della Rocca, M. (2010), "PSR", *Philosophers' Imprint* 10(7), 1–13.
- DiSalle, R. (2016), "Newton's Philosophical Analysis of Space and Time", in Iliffe, R. and G. Smith (eds.) (2016), 34–60.
- Emery, N., N. Markosian, and M. Sullivan (2020), "Time", in E. Zalta (ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2020 Edition), URL: <<https://plato.stanford.edu/archives/win2020/entries/time/>> (21.5.2024).
- Futch, M. (2002), "Leibniz's non-tensed theory of time", *International Studies in the Philosophy of Science* 16(2), 125–139.
- Futch, M. (2008), *Leibniz's Metaphysics of Time and Space*, Springer, Netherlands, Dordrecht.
- Ganssle, G. (2024), "God and Time", *The Internet Encyclopedia of Philosophy*, ISSN 2161-0002, URL: <<https://iep.utm.edu/god-time/>> (8.6.2024).

- Garrett, A. (ed.) (2014), *The Routledge Handbook of Eighteenth Century Philosophy*, Routledge, London, and New York.
- Hausheer, H. (1937), “St. Augustine's Conception of Time”, *The Philosophical Review* 46(5), 503–512.
- Huenemann, C. (2008), *Understanding Rationalism*, Acumen, Slough.
- Iliffe, R. and G. Smith (eds.) (2016), *The Cambridge Companion to Newton*, Cambridge University Press, Cambridge.
- Janiak, A. (2008), *Newton as Philosopher*, Cambridge University Press, Cambridge.
- Janiak, A. (2014), “Introduction”, in Newton, I. (2014), vii–xxxv.
- Janiak, A. (2022), “Kant's Views on Space and Time”, in E. Zalta (ed.), *The Stanford Encyclopedia of Philosophy* (Summer 2022 Edition), URL: <<https://plato.stanford.edu/archives/sum2022/entries/kant-spacetime/>> (21.5.2024).
- Kant, I. (1992a), “A New Elucidation of the First Principles of Metaphysical Cognition”, in Kant, I. (1992b), 1–45.
- Kant, I. (1992b), *Theoretical philosophy, 1755–1770*, translated by D. Walford and R. Meerbote, Cambridge University Press, Cambridge.
- Kant, I. (1996), *Critique of Pure Reason*, translated by W. Pluhar, Hackett, Indianapolis.
- Koistinen, O. and A. Repo. (1999), “Compossibility and Being in the Same World in Leibniz's Metaphysics”, *Studia Leibnitiana* 31(2), 196–214.
- Laks, A. and G. Most (2016a), *Early Greek Philosophy: Western Greek Thinkers* (Volume 3, Part 2), translated by A. Laks and G. Most, Harvard University Press, Cambridge, MA, and London.
- Laks, A. and G. Most (2016b), *Early Greek Philosophy: Western Greek Thinkers* (Volume 5, Part 2), translated by A. Laks and G. Most, Harvard University Press, Cambridge, MA, and London.
- Leibniz, G. (1961), *Die philosophischen Schriften*, edited by C. Gerhardt, Georg Olms, Hildesheim.
- Leibniz, G. (1989), *Philosophical Papers and Letters*, translated and edited by L. Loemker, Kluwer Academic Publishers, Dordrecht, Boston, and London.
- Maudlin, T. (2007), *The Metaphysics Within Physics*, Clarendon, Oxford.
- Melamed, Y. (2014), “What is Time?”, in A. Garrett (ed.) (2014), 232–244.
- Melamed, Y. and M. Lin (2023), “Principle of Sufficient Reason”, in E. Zalta (ed.), *The Stanford Encyclopedia of Philosophy* (Summer 2023 Edition). URL: <<https://plato.stanford.edu/archives/sum2023/entries/sufficient-reason/>> (7.9.2024).
- Meli, D. (2014), “Newton and the Leibniz–Clarke Correspondence”, Iliffe, R. and G. Smith (eds.) (2016), 586–596.
- Newton, I. (2014), *Philosophical Writings*, Cambridge University Press, Cambridge.
- Plato (1929), *Timaeus. Critias. Cleitophon. Menexenus. Epistles*, translated by R. Bury, Harvard University Press, Cambridge, Massachusetts, and London.
- Rynasiewicz, R. (2022), “Newton's Views on Space, Time, and Motion”, in E. Zalta (ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2022 Edition). URL: <<https://plato.stanford.edu/archives/spr2022/entries/newton-stm/>> (17.6.2024).
- Slavov, M. (2022), *Relational Passage of Time*. Taylor and Francis, New York.

- Spinoza, B. (2002), *Complete Works*, translated by S. Shirley *inter al.*, Hackett, Indianapolis.
- Styrman, A. (2023), “The Passage of Time as Causal Succession of Events”, *The Journal of Philosophy* 120(12), 681–697.
- Vailati, E. (1997), *Leibniz and Clarke: A Study of Their Correspondence*, Oxford University Press, New York, and Oxford.
- Yakira, E. (2012), “Time and Space, Science and Philosophy in the Leibniz-Clarke Correspondence”, *Studia Leibnitiana* 44(1), 14–32.
- Yenter, T. and E. Vailati (2021), “Samuel Clarke”, in E. Zalta (ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2021 Edition), URL: <<https://plato.stanford.edu/archives/win2021/entries/clarke/>> (28.4.2022).

Write Down Every Moment: An Argument for Perspectivalism about A- and B-Theories of Time¹

VILLE V. KOKKO

1. Introduction: Eating the Past

In the novella “The Langoliers” by Stephen King², a number of passengers on an airplane wake up to find that most people on board have vanished, including the crew. It seems everyone who was not asleep has evaporated into thin air. Fortunately, one of the passengers is himself a pilot, so he is able to fly the plane to an airport and land it. There, the survivors leave the plane and come to a world that seems different and wrong. There are no other people, there is no electricity, and causality seems half-hearted – sounds and smells are muted, matches will not burn. There is also a mysterious but ill-omened noise coming from the horizon, gradually drawing closer.

One of the survivors is a mystery writer, and he makes a guess as to what has happened, which seems to be entirely correct.³ It is not that the world has changed – *they* have themselves travelled to the past when the plane accidentally flew through some spatiotemporal rift. That is why there is nobody there. This is what happens to the past when it becomes the past: it is deserted and faded.

That is not all that happens to the past, though. After some misadventures, most of the group manages to return to when they came from by flying the plane back through the rift in time. Just as they are taking off, though, they see the thing behind the mysterious noise they have been hearing all the while. A horde of eldritch creatures they dub the Langoliers are devouring the earth itself and

¹ This article is inspired by Valtteri Arstila’s course on the philosophy of time at the University of Turku in 2016, and it is intended to be a starting point for further research into the topic that I plan to conduct later.

² (King 1990).

³ That he is able to guess it correctly strikes me as the least plausible part of this story, considering humanity’s past record in the light of the history of various beliefs and the huge effort it takes science to get it right.

everything on it, leaving behind black nothingness in its place. Thus, it turns out that the past is not merely faded and uninhabited. It is also, after a while, just gone. At least sort of.

As a further glimpse into how time works in this fictional universe, when the survivors make their way through the time rift, they arrive in what also appears to be a dead, unpopulated world, though it seems somehow different than the past world. Then time catches up to them, with other people appearing and the world becoming alive again, and it turns out (as the novelist guessed again) that they had arrived a little way into the future. So, the future, too, exists in a way different from the present.

This may be an odd and surprising view of time travel, but from a philosophical standpoint, it immediately raises some questions and evokes some existing theories, not just of how time travel might work, but how time itself works.

2. Does the Past Get Eaten? Different Theories of Time

Though time travel in “The Langoliers” adds its own significant aspect to the story as an example of theories of time, before going into that, we should look at how it relates to philosophical theories of how time (without travel) works.⁴

One theory that is reminiscent of what happens in “The Langoliers” is *presentism*. According to this theory, only the present exists; the past and future times do not, nor do things in those times.⁵ The past gets thoroughly eaten, and the future has not been born yet.

Though the Langoliers may function as an evocative, fanciful metaphor for presentism, the story’s world is not really an accurate match for that theory. After all, the Langoliers do not eat the past moments themselves, just their material contents, besides of which they arrive late, so even if they did eat the past itself, the past would have a “moment” to exist in diminished form before that.⁶ At first sight, also, if presentism was true, there would be nowhere to travel to, though the

⁴ For a more detailed explanation on these, see e.g. (Miller 2013).

⁵ For brevity, I am going to speak of times or moments of time as existing as short for either the moments themselves existing, whatever that may mean, or the things that are located at those times existing. Insofar as moments of time “existing” is problematic, it can be read as meaning the contents of those moments of the sort that can normally be seen as existing or not.

⁶ Indeed, the way they eat “the past” or “the world” is rather peculiar on a closer look, as they devour only the ground and replace it with some kind of embodied nothingness, leaving the atmosphere let alone outer space untouched.

next section will show this is not true. As such, in strict terms, the theory of time that fits this story is rather the *moving spotlight* theory.⁷ In this theory, all moments in time exist, but the present has a special status. “The Langoliers” definitely exhibits a form of this idea.

On the same basis as the above, a theory that might have been ironically true about the Langolier world is the *growing block* model, in which the future does not exist, but the past does. The future is different because it has not come into being yet. This would have been ironic, since the Langoliers appear to remove the past, but, as stated above, they do not really. Since they only eat a big part of it, we might call this a “hollowed-out block” scenario. The growing block model could have applied in the story without the final scenes where the characters return through the time rift and end up a little way into the future, thus showing that the future does exist, though it is different and not in the spotlight.

All of the above theories accept what is known as the *A-theory of time*, meaning that they hold that the present is special, it might be said, time flows.⁸ I will call the individual theories accepting the A-theory “A-theories” for short. A theory of a different sort is *eternalism*, which embraces the *B-theory of time*. (The A and B in the names will be explained shortly.) According to eternalism, every moment exists equally; this is taken to imply a static universe in which time does not flow. It is natural to think that eternalism allows more unrestricted time travel; if the universe in “The Langoliers” functioned in accordance with B-theory, it would take some further explanation for how it would be possible for the past and the future to be different from the present. However, as will be shown in the next section, something like that is needed even in A-theories.

3. Time Travel without Meta-Time Travel

The time travel seen in “The Langoliers” differs greatly from the typical case of time travel in fiction. Usually, the past or future that time travellers enter is just as real as the present they left from. This is even seen in other works by King, such as his magnum opus, *The Dark Tower* fantasy series.

⁷ (Daniels 2016, 212–15.)

⁸ I will not go into detail about “time flowing” (or passing) in this article. I will just roughly treat it as something assumed to be associated with A-theory and the present being special and denied by the B-theory (see e.g. (Miller 2013, 346). For more on the topic, see e.g. (Smart 1949), (Prosser 2013).

We might try to explain the matter as follows: in those other stories, eternalism is true. Thus, when the characters travel to the past (I will speak of only the past from here on to avoid having to keep saying “or future”), they find something there because in their universe, times other than the present exist. Indeed, the *The Dark Tower* series involves several alternative universes, and it is stated that going back in time is not possible in a universe called Keystone Earth, which resembles our world. For all we know, since *The Dark Tower* ties together with a number of King’s other works, it might even be that the Langoliers are specifically eating the past of Keystone Earth, and the characters would not have faced the same issues if they had been in another alternative universe.

It bears mentioning that since philosophers try to prove the different theories about time by argumentation that is often *a priori* and not empirically based, it may not be a coherent idea that there might be both or either presentist or eternalist theories of time. It might be that only one possibility makes sense, since someone might be able to prove just this by *a priori* argumentation. In fact, this article will argue something close to saying neither quite makes sense. Regardless, we can regard these different kinds of fictional time travel as thought experiments of what would follow if one kind of theory *was* possible and actual. I will focus on the difference between A- and B-theory here.

So, it seems as though in those worlds where time travel in the conventional sense is possible, B-theory must be true, whereas in worlds where A-theory is true, time travel will either be conceptually impossible or conceptually limited as in the Langolier case.

This turns out to be wrong. No theory of time can make time travel conceptually impossible, and it takes something extra to make it physically impossible.

When the characters in “The Langoliers” have travelled to the past, they find the past to be different because it is the past. It is their past too, so they also experienced this same moment as real when they were passing through it in the normal sequence of things earlier. There is something different between these situations, in one of which they were in this time, and it was the present, and in the other one of which they are in that time again, but it is the past. What is it?

Time travel can be explained by saying that the time-traveller’s personal timeline does not coincide with the external timeline of the world, so that, for example, when someone travels to the past, their personal timeline stops

continuing in the present and continues in the past instead.⁹ Thus, the answer as to why the past is different now that it is the past must have something to do with how their timelines are out of synch with the external timeline.

However, we cannot say that having your personal time be out of synch with the external time makes the moment you are in not the present. The whole point about speaking of personal time in the context of time travel is that your personal time moves to coincide with the time to which you travel. Insofar as this is explained by an A-theory of time, the explanation needs to be rather that they are in a moment that is not the present by some further standard. This is also why time travel is not conceptually impossible in any of the theories.¹⁰

In order to say that a time traveller could be in T but T would be somehow different from the present for them, we need to add a *meta-time*¹¹ keeping track of what time it really is. This is conceptually implied in “The Langoliers”. When the characters arrive in the past, it is true for them that they are in the moment, but also that that moment is not the present, even for them. There is some kind of meta-time that determines which point in time is the present and which are not – even when the characters are in one of those past moments. The previous sentence is just another way of expressing the same thing that was expressed in the sentence before it, hopefully in a way that can help make things clearer to talk about using the concept of meta-time.

Meta-time is a conceptual necessity for A-theories of time. They imply it even if they do not name it as such. It is trivial to say that every moment exists in that moment; we need to also be able to say that that moment is no longer or not yet the present. A-theories speak of other moments from the point of view of the present moment, so they need something to keep track of which moment is the present one, or, they imply track is being kept of which moment is that present. B-theory, of course, does not do this.

One might question whether it is a bit fishy to postulate a second time-like thing when we were supposed to be talking about theories about the passage or lack thereof of the one time. At the end of the day, I would say it is *not* justified,

⁹ See (Lewis 1976).

¹⁰ The same point can be made based on an argument presented by Simon Keller and Michael Nelson: We can describe a situation in which there is a time traveller at some point in the timeline even by the terms of presentism, as long as we use appropriate terms to do so, at least as long as we think of a way to handle problems of identity and a person being in two places at the same time. (Keller and Nelson 2001)

¹¹ See e.g. (Miller 2013, 350).

but insofar as A-theory can be argued for, meta-time can also be argued for, as it is just another way of saying something it already says.

Here is where we finally get to why the A-theory and B-theory are called that, because the question of meta-time calls back to the discussion by J. Ellis McTaggart and his 1908 article “The Unreality of Time”¹². McTaggart makes a distinction between three kinds of series of positions of time: The A series is “the series of positions running from the far past through the near past to the present, and then from the present to the near future and the far future”.¹³ The “series of positions which runs from earlier to later”¹⁴ is the B series; the difference is, as we have already discussed for the A- and B-theories, that in the A series, the present has a special position and the other moments are ranked as past and present in comparison with it, whereas in the B series, all we have is the order of the moments. Finally, the C series is one in which there is no change but just a series of events, ordered but without direction.¹⁵ We can mostly ignore the C series here.

McTaggart argues that time cannot exist, because the A series is essential to it, along with the C series but unlike the B series¹⁶, and the B series cannot exist without the A series. (McTaggart does not give a very convincing argument for the fundamentality of the A series, unless in the sense that one happens to be convinced by the same kind of intuitions.) The problem McTaggart proposes for the A series, to put it very shortly, is that each moment is at some point past, present, and future, and these are incompatible attributes. In this context, trying to explain this in temporal terms by saying that (for example) the present moment *was* in the future, *is* the present, and *will be* in the past, is circular, using temporal terms to explain time. In another way of looking at it, it involves a vicious infinite regress, for to use these terms of the A series involves postulating a second A series of past–present–future relations, but to explain that series, we would need yet another, and so on.¹⁷

My point in bringing this up is that the second A series is a kind of meta-time, and the problem McTaggart grapples with is has something in common with our question of how to explain a time-traveller being in the past but it still being the past for them. More broadly, McTaggart’s line of thought touches on the topic

¹² (McTaggart 1908)

¹³ (McTaggart 1908, 458)

¹⁴ (McTaggart 1908, 458)

¹⁵ (McTaggart 1908, 461–62)

¹⁶ (McTaggart 1908, 463–64)

¹⁷ (McTaggart 1908, 467–70)

of meta-time being needed to explain how we can view one moment as being the privileged present instead of just saying each moment is the present when it is the present. We can say this from the perspective of a position in meta-time, a meta-moment, that that moment corresponds to.

The aim of this article is not to answer McTaggart's problem or counter his arguments about the unreality of time – even though the conclusions of this article do imply his reasoning is erroneous. We can bypass the problem or regression in the additional A series in this context by supposing that it is acceptable for meta-time itself “not to flow,” for it to be a qualitative different thing than actual time in A-theory, so that there is no regress of meta-meta-times. Whatever problems this might cause down the line, and while it does not work as a solution to McTaggart's problem from his point of view, it is at least sufficient to describe how the Langolier scenario can have time travellers existing in a time that is not just the present for them: as said, their location in time does not match their location in meta-time.

Even in such a world, time travel cannot be conceptually impossible, and that is because time travel could be travel through both time *and* meta-time at the same time. If you change what time and meta-time your personal timeline matches, the moment you arrive at will fully be the present for you by definition. Of course, this might be physically impossible. The time anomaly in “The Langoliers” certainly does not allow it; the characters only travel in time, and they do not travel in meta-time correspondingly, so they end up in a past that is the past. Maybe that is the only way possible in their world. Nevertheless, if King had written that travel in meta-time is also possible in the same universe, allowing both kinds of time travel as two different ways of doing it, he would not have run into logical contradictions (although he might have had a hard time not confusing his readers). After all, all we have to do to postulate time travel is to postulate synching the next moment of someone's personal timeline with a different moment than normally. Postulating that about meta-time as well is equally trivial once we recognise the concept: instead of travelling in time but not in meta-time, you could travel in both. This is besides the main point of this article, but it is suggestive of it, as we will see below.

4. Deflation by Perspective?

After all the talk about A- and B-theories of time, the view I am presenting an argument for here belongs to neither, except insofar as it belongs to both. The view is that A-theory and B-theory cannot in fact be formulated in such a way that they contradict each other. That is, if we leave out components of the theories that fail to be meaningful and incorrect statements about the implications of the theories, the two theories are equivalent.¹⁸ This view is closely related to the triviality objection to A-theory that it is either trivially true or trivially false¹⁹, but I am also emphasizing that B-theory is mistaken insofar as it is taken to mean that time does not exist or does not flow²⁰. I will call this the *deflationary view* about A- and B-theories of time.

The claim is that anyone asking about whether other times exist is constructing a meaningless double category: whether other times are existing in the present. If you ask whether they exist now, they trivially do not, since they are other times, not the present. If you ask whether they exist when *they* are the present, they trivially do (unless you deny the present exists, which is not being done in the debate under discussion). Similarly, contrary to how it is sometimes put, B-theory does not literally say all times “always” exist, as it does not really claim that all times exist *at other times*.

Speaking of moments of time or their contents, you can *either* say something exists relative to a given moment, *or* that it exists in some moment at all, but you cannot say anything about things not in the present but at some other time existing *now* other than the trivial thing that they do not exist *now*.²¹

This is not to be confused with any argument based on what supposedly can and cannot be said in a particular language such as English.²² I am sure we can use language to express all kinds of things that are not normally spoken about in it. The question is whether we are managing to express anything meaningful with what we say. A-theorists and B-theorists would not agree if they were asked whether they are speaking of the same thing in different terms,²³ but the claim is

¹⁸ See (Miller 2013, 348–51).

¹⁹ See e.g. (Meyer 2005).

²⁰ Eternalism implies time does not flow: E.g. (Miller 2013, 358).

²¹ C.f. (Meyer 2005, 214–15).

²² A criticism of such arguments, one I largely agree with, can be found in (Deasy 2019, 3371–78).

²³ See e.g. (Miller 2013, 349).

that they are mistaken in thinking they are managing to make meaningfully distinct claims.

If we are saying the present is special, not talking about the actual present but whichever moment is the present as in A-theories, then we are already saying every moment is special *when it is the present*. Those describing A-theories as not trivial have to insist that they mean something more than this, though – but the only way to say something more would be to talk about the specialness of the actual present, since otherwise you are just talking about whichever moment is the present.

The above might seem to suggest that we cannot distinguish between different A-theories, either. There can be no growing block if you cannot speak of the past and future as different from the perspective of the present, nor can there be a difference between the growing block and the sole existing moment of presentism. The spotlight will simply trivially light each moment when it is the present.

This is why we needed to postulate meta-time to explain Langolier-type time travel, so that the moments can be non-trivially different in relation to meta-moments instead. Meta-time does allow describing the difference between different A-theories even granting the deflation. However, if there is no need to make a distinction between A-theory and B-theory, we do not need to postulate this meta-time in the first place. Meta-time in B-theory will be discussed in the next section.

I think that making the case like this is already objectively sufficient; that the deflation between A- and B-theories follows from what has been said. However, it is clear that this has not convinced a great many philosophers, since the mainstream discussion about the topic seems to accept the difference between A- and B-theories. This fact perhaps gives me reason to suspect that I am wrong. In any case, it necessitates that I defend my position with a more specific argument that, hopefully, makes use of that putative correctness of the position to show more indirectly that it must be correct. That is what I will do next.

5. The Argument: Just Write down Every Moment

In this section, I finally present the paper's specific argument for the deflationary view about A- and B-theories.

Imagine a god-like being similar to Laplace's demon. We can call her a daemon for the slightest bit of variety. This being observes the history of a universe – maybe the entire history, maybe just a subsection of it – and writes down a list of every single moment that takes place during it on a godlike piece of paper. We will look at what happens if the universe she is writing about follows the A-theory, and what happens if the universe is in accordance with B-theory instead.

Writing down every single moment assumes that time is discrete. If it is not, the daemon can instead write down a long list of arbitrarily chosen moments, such as including one entry per second. We can imagine the daemon either as existing within the timeline she is writing down, so that at any time when she is writing and reading the paper, one of the moments named on it is the present one. We can equally well imagine that she is outside the universe such that she and the paper do not exist in any of the moments described therein. Either way, the end result on the paper is the same: a list of moments.

If B-theory is true, what happens when the daemon writes down every moment is simple. All moments are listed on her paper, statically, with no preference given to any of them, just like they are in reality. Of course, it is true that all the moments are represented there statically at the same moment in time, which is not something that is the case of the moments themselves. They do not all exist at the same time in that sense. Nevertheless, the representation is accurate in that it represents all moments as equal and – in some sense – static.

Now, what happens if the daemon is studying a universe where A-theory is true? Again, she writes down all the moments, though this time, they are moments out of which one that is the present is special and different from the others. However, even if she makes a note of this last fact in the margin, the rest of the paper ends up looking just like in the case in which B-theory was true.

Even if A-theories are true, there is nothing to keep the daemon from *considering* and *representing* each moment at the same time on her paper. From the dynamic reality in which time flows, it is still possible to create a static representation. Yes, this representation will be lacking the information of which moment is actually the present, but it can still represent everything that went on when each of them was. It can contain the information to see what was going on when T_1 was the present, and also the information of what the situation was when T_2 was the present. After all, we do something like this all the time when we so much as discuss this issue: we talk about *the present* in, say, the moving spotlight theory without specifying that said present is three o'clock on last Tuesday, the

moment when the text was written, the moment when it is being read, or any other specific moment. It does not matter. All that matters is that it is the present when it is the present. In doing this, we are already generalising over different presents from a perspective outside of them. We can also specify that we are speaking of two or more different moments in the same discussion. The daemon is doing the same thing, just writing down every moment (or arbitrarily many) for the sake of completeness.

It has already been established that, when considering different moments in time, we need meta-time to indicate any one of them as the present and the others not. We can also see this effect on the paper that the daemon has written. All the moments are listed there, and we are not told which one is the present one. If we look at any given moment, we can equally well consider it as present, past or future.²⁴ As before, we have to add meta-time to say more than this, but what happens when we do?

Let us say the daemon writes down every moment of meta-time too, and she notes how moments of meta-time correspond to moments in time. Now, she can see that at some given meta-moment M , some given time T is the present.

The daemon's representation of meta-time is just as static as her representation of time. As McTaggart noticed, positing a second series to make time flow would only lead to needing to posit yet another series to make that meta-time flow, and so on to infinity. In this thought experiment, the same effect is shown in that any level of meta-time *can* be represented just as statically as normal time. The consequence of this is that, with or without meta-time, or with any number of layers of meta-meta-time, every moment can *still* be considered equally well as past, present or future. To view any given moment as present, the daemon only needs to consider the conjunction of that moment in time with the meta-moment that corresponds to it and makes it present. To consider a moment as past, she needs to consider that moment in conjunction with any meta-moment (or all of them) that designates a later moment as the present, and similarly for the future.

Meta-time does not really do any work here. It would *only* be needed in a Langolier-type scenario of time travel where someone travels in time and ends up in a past or future that is different from the present due to not being the present. In such a case, we would need to be able to say that their personal timeline is out of synch with meta-time, that is, coinciding with a moment and a meta-moment that

²⁴ The very first moment cannot be considered as any moment's future, and the very last cannot be considered as any moment's past, but this is trivial.

do not correspond to each other. In normal cases, or in cases of time travel of the normally imagined sort where travel effectively takes place in meta-time as well, we can reduce meta-time away: we can simply identify each conjunction of a moment and its corresponding meta-moment as one moment.

We have arrived at this scenario by assuming A-theory, in which time flows and the present has a special status. I repeat that even in describing in this scenario as I just did, which is a perfectly standard way, we are not specifying *which* individual moment is the present. We are not merely saying, trivially and parochially, that this moment we are in now is the present now. We are generalising over every present moment. That is also what the daemon is doing, and once she has done it, *she ends up with a representation identical to the one she makes out of a B-theory universe.*

Of course, the B-theory universe does not include the second, meta-time series. However, we can simply see what would happen if she examined a B-theory universe that does have meta-time, only B-theory applies to this meta-time as well, and so every moment of meta-time exists equally and statically as well. Having written down every moment and meta-moment, the daemon can then consider what would happen if any meta-moment was the present meta-moment, and she can consider what would be the case if each moment was the present (past, future) by considering the conjunction of that moment and the relevant meta-moment(s). It is this scenario that our A-theory scenario would end up as being identical with, which is why this paragraph sounded like repetition from before.

Meta-time or the presence thereof (or the presence of any amount of meta-meta-times) does not make a difference from this perspective. The representation made by the daemon would always allow considering each moment in time the same way or considering any one of them equally as being the present. The information that time flows and the present is special would have been included in making this representation, but it would not leave any trace in the final representation.

It might be thought this would mean that there would be something missing from this representation. The obvious information that is missing is what the present moment actually is. If we postulate that the daemon is looking at her list of moments at a specific moment T that is included in the list, then we can see that T is special and that T on the list and M on the list of meta-moments are special for being the current ones. However, this would not be making the statement that the present is special beyond saying that whatever happens to be the present is

special when it is the present – and this makes every moment equally special unless we only consider that moment’s perspective. However, again, the very statement of A-theory refers to a generalised present instead, and there is nothing keeping the daemon or us from considering all moments in abstraction without specifying which one is the present.

If A-theory is supposed to be different from B-theory, we might suppose that the model created by the daemon on her piece of paper is inaccurate in some other, unspecified way, since the result it gives is the wrong one. We could try to describe the desired result we would get instead, one where the daemon had written down things in such a way that she could see that the present was special. However, the results of trying this only affirm the impossibility.

As mentioned before, if we had the information of what moment actually was the present, we could point to that moment on the list and say that it is special right now. Then we could point at a later one as more time passed, and so on. (Note that there is nothing in B-theory keeping us from doing the same; the pointing would simply need to be situated in the time slice indicated by the writing pointed at.) However, we have been trying to describe the scenario where every present is special when it is present. What would it mean if the daemon recognised a special present on the list if it was not simply the actual, changing present? If there was somehow a moment *T* that even the daemon, looking at her static list, would recognise as a privileged present, then none of the other moments would *ever* be the present. (As for meta-time, this might be interpreted by saying there is only one moment of meta-time, or at any rate however many there are, they all couple with *T* and no other times.) There would only be one moment that would be the present. Even as time passed (if it would count as doing that in this scenario), we would live in the past up to a certain point, then we would quickly pass through the present and thenceforth live in the future.

The above scenario can only be so much as coherently described by positing that “the present” only ever refers to one particular moment, but that just makes “the present” a name without any content. Otherwise, considering “the present” having any usual meaning, the whole scenario is nonsense. We could also postulate that more than one moment is the present, but not all of them are, but this would be at least equally nonsensical. The daemon can understand what it means for time to flow and the present to be special, but she cannot look at her representation and see it as being any different than a B-theory.

There is another side to these results. It is not that A-theory is wrong and B-theory is right because, as we represent it here, A-theory ends up being represented in terms that are just like those used to describe B-theory in the first place. Since the representation of A-theory and that of B-theory are just alike, it also follows that the features of A-theory (coherently understood) are just as true in B-theory. I will return to this in the next section.

The daemon's situation is really just a concretisation of our situation whenever we speak of moments of time in general without fixing any point at the present – even if we are doing so when trying to describe an A-theory different from B-theory.

From the above, I intend to draw the conclusion that instead of separate A- and B-theories, we just have A- and B-*perspectives*, which both describe the same thing but highlight different sides of it.²⁵

6. The A-Perspective and the B-Perspective

Given the results of the thought experiment about the daemon, which seemed to show that there is no meaningful difference between A- and B-theories, I propose a different view of how to see the ideas in these theories: They are not, properly understood, two contradictory theories. Instead, they are two different perspectives of the same way that the reality of time may be. (Since they cannot be distinguished, these perspectives necessarily describe the same kind of reality.)

One perspective, which we could call the A-perspective, assumes a particular present moment.²⁶ As in the A-series, we can thus identify a past, present and a future, not merely relative relations of earlier and later. While we are anchored to this present moment (even though, ironically, we can do this in an abstraction where we are *not* anchored to a particular moment, which is almost a contradiction and may be part of what leads to the confusions discussed in this paper), as long as we look at things from this perspective, the present shows as existent and the past and future and anything in them do not.

The A-perspective is not the perspective that the daemon is taking. Rather, the point about the daemon is to demonstrate that we can always move to the other

²⁵ A parallel argument to the one in this section could be formulated to deflate the question of whether objects have temporal parts or travel through time as wholes. The daemon would not see a difference between those scenarios either.

²⁶ It might be that I should say “a particular moment or a moving present,” but as with time flowing in general, I have not got the opportunity to look into that idea here.

perspective, the B-perspective, in which there are only relations of earlier and later.²⁷ In the B-perspective, we do not identify a particular present moment (even a postulated but unspecified present moment), but instead view each moment equally. Thus, we can see which moments come after or before which ones, and we could derive from this what it would mean for something to be the past or the present in relation to a particular moment, but we acknowledge that we can do this to each moment. This also has a bearing on what we count as existing: we are not predicating a particular moment of time in ascribing existence, so we can count everything that ever exists as existing, since it all does exist at some time or other.

This view avoids the contradiction of modifying our statements with two different time attributions at the same time. We do not ask whether something in the past or future exists now. We can ask whether something exists at a particular, identified moment, taking the A-perspective, or whether it exists at any time, taking the B-perspective, just as in the first formulation of the deflationary view introduced above.

The apparent contradiction between the A-theory and B-theory follows from both insisting that there is something more to time than the relations included in B-theory that are actually sufficient to explain what we know about time. If it was not sufficient, we could not start with an A-theory world and write down a complete list of its moments that looks the same as a B-theory world. Time “flows” in that each moment comes after the last one, and causality and our experience follow that order. A-theory claims that there is something more beyond that, something that makes time flow even after the flow is reduced to its components; however, it cannot explain what this something could be, because there is really nothing there except the mistaken judgment that something else *is* needed. Meanwhile, when B-theory is seen as implying that time is not real, the same assumption is made – that something else is needed to make time truly real – but there is no such thing to be found. A-theory can be derived from B-theory by examining the more specific case of the perspective of a particular moment. That is all it takes.

²⁷ We might even call this perspective one using a C series with the addition of a direction, because apart from direction, the C series is not missing anything that would be needed to represent the same thing. The B series is supposed to already be temporal, after all, but the point being argued in this paper could be extended to say that all the features of temporality can be derived from a series with an order and a direction, what is being called the B series here, without remainder. I will leave this observation at the level of a footnote and speak of the A and B series (and A-theory and B-theory, A-perspective and B-perspective) elsewhere in this paper for the sake of simplicity and not going too far afield.

This does mean that B-perspective is more fundamental in the sense that A-perspective is reducible to it. (If I somehow had to make the false binary choice between A- and B-theory, I would choose B-theory.) However, we should be careful about what we mean by fundamentality. A-perspective is arguably more real and important to our experience of our existence, not only in terms of subjective experience, but for such fundamental and objectively important underlying aspects as action, planning and change. Which perspective we see as more important or more “real” is, well, about perspective. It depends on which features we highlight. To say which one is more important is a judgement call anyway, and there are good arguments for each being important.

I can even change perspectives and say that the A-perspective is fundamental and essential. You are reading this *now*. And now. And now. Not the abstract present, but the actual one right now as you read this. Every one of us is always at some point in time (especially given the meaning of “always”). We are not gods outside of time, and if we were, we could not even do anything. The B-perspective where we abstract away from any given moment in time is just an abstraction. A representation leaving something out. Every such representation, as concrete or mental, is existing at some points in time and not others. Every moment we talk or think about the question is a particular moment, one with a specifiable past and future. McTaggart would hardly have claimed to disprove *this*. What he was instead doing was abstracting away this essential quality that we indeed always have, the very real moment in which he was and then the next moment and then the next, moving into an artificially crafted, impoverished B-perspective, and then complaining about not finding the present he abstracted away in that abstraction.

Right now, we are in *this* present. This, not some other, although which one it is has presumably already changed since the last sentence. This present is special because it is now. And now this one is. We can say that “this one” is indeed special. If we abstract away from which one it is, then all we can hope to say is that each present is special when it is the present. If we wish to find the present that is special without being relativised like this, then we need to look at the actual present – you know, the one that is *now*, and now the one that is *now* – not abstract it away. If you do abstract it away, abstract to a level where every moment is equally present, what do you expect you will get in terms of the present being special? How can it be special when you do not even know which one it is? The present indeed has a very special status: being the present. What else?

University of Turku

References

- Daniels, Paul R. 2016. 'What Happens to the Present When It Becomes the Past: Time Travel and the Nature of Time in The Langoliers'. In *Stephen King and Philosophy*, 207–18. London: Roman & Littlefield.
- Deasy, Daniel. 2019. 'The Triviality Argument against Presentism'. *Synthese* 196 (8): 3369–88.
- Keller, Simon, and Michael Nelson. 2001. 'Presentists Should Believe in Time-Travel'. *Australasian Journal of Philosophy* 79 (3): 333–45.
- King, Stephen. 1990. 'The Langoliers'. In *Four Past Midnight*. Hammondsworth: Viking.
- Lewis, David. 1976. 'The Paradoxes of Time Travel'. *American Philosophical Quarterly* 13 (2): 145–52.
- McTaggart, J. Ellis. 1908. 'The Unreality of Time'. *Mind* XVII (4): 457–74.
- Meyer, Ulrich. 2005. 'The Presentist's Dilemma'. *Philosophical Studies* 2005 (122): 213–25.
- Miller, Kristie. 2013. 'Presentism, Eternalism, and the Growing Block'. In *A Companion to the Philosophy of Time*, 1st ed., 345–64. Chichester: Wiley-Blackwell.
- Prosser, Simon. 2013. 'The Passage of Time'. In *A Companion to the Philosophy of Time*, 1st ed., 315–27. Chichester: Wiley-Blackwell.
- Smart, J.J.C. 1949. 'IV. - The River of Time'. *Mind* LVIII (232): 483–94.

Kant on Time

HEMMO LAIHO

This essay is intended as a brief introduction on Kant's views on time and the theoretical contexts he presents them in. The idea is to sketch out Kant's views as I see them in such a manner that it would be easy to build upon the discussed themes on a later occasion. I start by listing the key Kant texts that concern the notion of time. Then I investigate the theoretical contexts where time plays a prominent role in Kant's critical philosophy. Finally, I take a somewhat more detailed look at the Transcendental Aesthetic, in which Kant supposedly claims that time is nothing but a subjective way of representing things in temporal order.

1. Time in Kant's Texts

Time became a specific topic for Kant in his inaugural dissertation *On the Form and Principles of the Sensible and the Intelligible World (De mundi sensibilis atque intelligibilis forma et principiis, 1770)*. The dissertation paved the way for Kant's most famous work, the *Critique of Pure Reason (Kritik der reinen Vernunft, 1781; revised edition 1787)*. Regarding time in particular, many things Kant says in the dissertation resemble quite closely his treatment of the concept of time in the *Critique* a decade later (cf. esp. §14, AA 2:398-402 and A30-36/B46-53).¹

There are many interesting things Kant has to say about time and time-related issues in his other works too, especially *Prolegomena (Prolegomena zu einer jeden künftigen Metaphysik die als Wissenschaft wird auftreten können, 1783)*, the *Metaphysical Foundations of Natural Science (Metaphysische*

¹ 'AA' refers to the so-called Academy edition of Kant's works and its volume and page numbers. 'A/B' refers to the two editions of the *Critique of Pure Reason* (1781/1787). See Bibliography for further information.

Anfangsgründe der Naturwissenschaft, 1786), the *Critique of Practical Reason* (*Kritik der praktischen Vernunft*, 1788) and the *Critique of the Power of Judgment* (*Kritik der Urteilskraft*, 1790; revised edition 1793). Arguably, however, all of Kant's post-1781 considerations about time are based on the views presented in the *Critique of Pure Reason*, which is why I will focus, in what follows, on those chapters and sections of the *Critique* that concern time.

In addition to books and articles, Kant's collected works contain a great number of lecture notes on various subjects, marginal notes on the books Kant read and used as teaching material, and other personal papers left behind after his death. As far as the concept of time is concerned, of these originally unpublished texts particularly worthy of mention are the lecture notes on metaphysics (e.g., Kant 2001) and the collections of Kant's personal notes (e.g., Kant 2005; AA 14-23). When looking deeper into Kant's views on time, these texts should also be considered.

2. Theoretical Contexts of Kant's Views on Time

Transcendental Aesthetic

Time appears as central in several theoretical contexts in Kant's writings. Assumedly, the most central of them all is the section called Transcendental Aesthetic in the *Critique of Pure Reason* (A19-49/B33-73). In that section, Kant lays out the basis of his theory of sensibility (*Sinnlichkeit*) and so-called transcendental idealism. In Kant's model of the mind, the faculty of sensibility answers for everything that is "sensible", "aesthetic", or (sensibly or sensorily) "received" or "given" in the representations of objects. Put differently, the mind presents us with "appearances" through sensibility. Everyday sense perception would be one example of such sensibility-originated representation, though it should be kept in mind that on Kant's view sensible representation extends beyond the contributions of the physical senses as such.

The key concept in the Transcendental Aesthetic is *intuition* (*Anschauung*). This concept too serves several explanatory functions. In Kant's terms, when you see a house, you intuit a house (B162; AA 8:217n; 9:33). Such an intuition would be an instance of *empirical intuition*—an intuition that picks out, or through which you single out, the empirically appearing object before you. Empirical intuition is only possible through sensations that rise in the mind via external input, or as the mind is being "affected." However, a geometrical construction—a triangle, for example—would also be an intuited object, or an object represented in intuition,

and indeed regardless of whether you see or draw it. If you do not, your representation of the triangle is an instance of *pure intuition*, that is, an intuition without sensation. (See also *Vienna Logic*, AA 9:907)

At this point, things start to get trickier. For Kant, space and time themselves are supposed to be pure intuitions too. Neither of them is an object like the house or the triangle, however. Rather, time, just like space, is a *form* of intuition: a kind of representational structure that necessitates the way we represent sensible objects and allows this to happen in the first place. As Kant would have it, it is important *not* to think of space as absolute or substantial, or merely as a set of relations (see, e.g., Janiak 2022). Rather, space is that by which the cognizer represents external, locatable and dimensional objects that all appear as spatially connected to each other. Time, in turn, understood as a form of intuition, is not something “out there” or in the objects as such, but rather that by which the cognizer represents the objects—including the cognizer herself via the mind’s inner states—in relations of succession and simultaneity.

Ultimately, Kant wants to argue that space and time, understood as forms, must precede the contents of *actual* experience. By this Kant does not mean temporal precedence, but a kind of transcendental dependence relation: the actual intuition of appearing objects presupposes, or is preconditioned by, space and time. This is crucial in the bigger picture, because only in this way, Kant thinks, is it possible for us to know something necessary regarding the objects of experience, instead of mere empirically given contingencies that we happen to receive through actual experience. For example, it is an inescapable and a priori establishable feature of experience that the straight line between the points A and B is shorter than a curve would be, or that an enclosed figure is not representable with two straight lines. The themes of apriority and the necessary features of experience are closely linked to Kant’s major question in the *Critique*: How is synthetic a priori cognition possible? By synthetic a priori cognition Kant means, roughly, the kind of cognition that takes us beyond the analysis of concepts, yet makes claims that are knowable a priori and, moreover, true of any *possible* experience.

Syntheses, Categories, Schemata

By synthesis Kant means, roughly, the mind’s grasping of a bunch of separate elements together (*A/B passim*). There are many kinds of synthesis, and their explanatory function is not always fully clear, partly because of Kant’s complex terminology. That said, a relatively easy example of a synthesis would be the thought about a red round apple. In such a thought, there is a synthesis of

something intuited taken as a bearer of properties (apple) and some other things (red, round) being represented as the properties of that something. Arithmetic operations are clear cases of syntheses too. To arrive at the number 3 through adding 1 to 2 requires keeping the numbers 1 and 2 in the mind, and then unifying these two numbers according to a certain mathematical concept or rule. Drawing the line between the points A and B would be a similar example. Even though drawing the line does not take place through discrete units, you must be able to entertain in your mind what you have drawn so far to grasp the continuity of the line—and, for that matter, to understand what a line amounts to in the first place.

Categories, also known as the pure concepts of understanding and the pure concepts of synthesis, are, for Kant, the elementary concepts, or conceptual rules, that structure thinking in its connective operations. For example, to think of the previously mentioned apple as a bearer of properties is to employ one such structural feature, or a form of thinking—Kant would call it Substance or *substantia et accidens*. In addition to being cases of the application of the categories, it is common to the above examples that there is something sensible available to the cognitive subject: the red apple perceived through the senses, the mental imagery of the distinct numbers used in the arithmetic operation—perhaps with the help of apples and oranges—and the line being drawn on the paper. In the bigger picture, the sensible element is crucial, because the categories gain cognitive significance only by being related to something intuited or intuitable—roughly, to something that we can come across either in actual or possible experience. Otherwise, the categories are to remain merely abstract forms—maybe seemingly useful in dogmatic metaphysics but without ascertainable access to reality. It is also common to the above examples that they are temporal in character: synthetic operations take time to reach their goal, that is, the unification of the elements in question according to a certain rule or principle.

In the section known as the Schematism of the Pure Concepts of the Understanding (A137-147/B176-187) Kant aims to specify how the categories are related to the sensibly given. Unfortunately, however, Kant's presentation in that section is so quick that it is difficult to tell how exactly he is supposed to achieve what he intends to achieve. Kant even points out that he wants to avoid the “boring analysis of what is required for transcendental schemata of pure concepts of the understanding in general” (A142/B181). After stating this, he basically just lists the categories, and their schemata, accompanied with very brief descriptions. He even forgets to list some of them, or perhaps that is intentional. What seems clear,

however, is that time plays a central role in explicating how the categories gain cognitive significance through the respective schemata. The basic idea appears to be that the categories need to be given a temporal interpretation. For example, the schema of Substance is “the persistence of real in time” (A144/B183) and the schematized—temporally interpreted—category of Actuality is an “existence at a determinate time” (A145/B184). This makes sense: you can only take something as a bearer of properties if there indeed is something the presence of which persists in a relatively unalterable fashion. Moreover, it is thanks to this that you can determine that something as actually existing at a certain time, which is also why the categories in question can find an object “among the appearances” (A90/B122) in the first place.

Axioms of Intuition and the Anticipations of Perception

Time also features in Kant’s account of the mathematizable contents of quantitative and qualitative sensible representation—intuition and sensation, respectively. To this end, the key sections in the *Critique* are the Axioms of Intuition (A162-166/B202-207) and the Anticipations of Perceptions (A165-176/B207-218). In the Axioms, Kant intends to show, among other things, how any object, in so far as it can be represented in a spatiotemporally determinate fashion, is (or is represented as having) an extensive magnitude. Such a magnitude refers, basically, to the measurable spatial boundaries and other quantifiable dimensions of the object, the representation of which is always a temporal-synthetic process. It should be noted, though, that the spatial shape as such can be represented in an instant too—it is just that you cannot measure its magnitude that way.

In the Anticipations, Kant intends to show, in a similar fashion, how empirical objects can be represented as (or as having) intensive magnitudes. Perhaps even more importantly for his aprioristic purposes, Kant wants to underline that we can establish universally that there is intensive magnitude to every sensation-based empirical representation, and that such magnitudes have the capability of increasing and decreasing in a continuous fashion—take the luminosity of light, for example, or temperature. It is also common to such continuous magnitudes that even if the object that possesses (or is represented as possessing) such a magnitude remains spatially speaking the same, there can be changes over time as far as the representing of the intensive magnitudes are concerned. Since their intensity ultimately tracks sensation, or is made apparent via sensation, one may also speak of the “degree of influence on sense” (B208).

In addition, sensation brings the notion of consciousness into the picture: the increase in luminosity, for example, also indicates a growth in empirical consciousness. By this Kant appears to point out how the cognition of any kind of dynamic process necessarily features temporal ordering of the input, and is therefore preconditioned by time, but it is an interesting further question whether the empirical consciousness in question should be understood as a kind of temporal consciousness as such.

Analogies of Experience

In the Analogies of Experience (A176-218/B218-265), Kant elaborates the connection between time and experience in terms of persistence, succession and simultaneity. Kant then analyzes these in three specific sections that relate persistence to substances, succession to causal-lawful sequences, and simultaneity to the interaction or community of substances. Perhaps the most famous subsection of the Analogies is the Second Analogy, and its example of a ship going downstream, where one of the main points is that the temporal order of the perceptual representations of the ship is determined by factors external to the examination of the ship's positions, thus providing an objective foundation for the subjective examination of the object (A192-3/B237-8). This can be read as suggesting that causal relations are prior to, or that they ground, actual temporal relations (Watkins 2005, 221). The distinction also helps us distinguish object-perceptions from event-perceptions: whereas you can examine the house in any order you want, the events, such as the ship moving downstream, necessitate the temporal-successive order of the examination of their constituents.

In the Refutation of Idealism (B274-294), which Kant added to the second edition of the *Critique of Pure Reason* (1787), he further argues that temporal determinations depend on outer affection in the end. As I see it, the Refutation is simply a sequel to the Analogies: both sections underscore that even if we are always dealing with representations partly dictated by the mind, this does not in the slightest mean that everything would be "up to us", as suggested by genuine idealism. Indeed, Kant clearly did not want to appear as an idealist in a strong sense. Here there is some irony, because the early critics of Kant saw him just as a new Berkeley. Even worse, the label seems to have stuck.

Antinomies of Pure Reason

Finally, time appears in the Antinomy of Pure Reason (A405-567/B432-595). The first of the four antinomies is probably most interesting in regard to the concept of

time. It is basically a dilemma between two opposite claims presented by the human reason: (1) The world has a beginning in time and is spatially enclosed. (2) The world has no such beginning nor boundary but is spatiotemporally infinite. Kant's answer to the dilemma is that both of its horns are false, or at least there is no way for us to verify either of them from a neutral point of view. Moreover, one claims that the world is given as infinite in experience, which however is empirically impossible to ascertain—how could there be a completed synthesis of an infinite time-series? The other horn, on the other hand, suggests that the world has limits, which too are impossible to ascertain—how could we grasp the spatiotemporal boundaries of the world, and the empty time and space beyond them, let alone explain the beginning of the world from nothing? (See also Rosenberg 2005, 293) What we have at best is the indeterminately far-reaching regress of experience, but neither actual infinity nor unconditioned boundaries. In the end, the antinomies do not concern that much time as such, but rather the way we tend to reason about the world, sometimes leading to apparently well-supported yet conflicting arguments used to support conflicting but undecidable metaphysical views.

Interestingly, it does not seem to be the same notion of time that Kant is using in the Antinomies and in the Aesthetic, at least not all the time. In the Antinomies, Kant seems to introduce what could be called historical time, whereas in the Aesthetic time is referred to as the formal condition of experience. There seems to be one thing common to both notions, however: Kant does not seem to ever really mean by time that which might be called experienced time.

3. A Closer Look at the Aesthetic

As we have seen, there are many time-related themes in the *Critique* alone. Some of them are quite technical. Some serve specific argumentative purposes within Kant's system, while also being of general philosophical interest. Next, and lastly, I take a somewhat more specific look at the Transcendental Aesthetic, which lays the foundation for Kant's critical philosophy and his philosophy of time.

Roughly half of the Aesthetic centers round time. Accordingly, the Aesthetic might be regarded as a kind of ontology of time. That said, given the nature of Kant's answer, as briefly explicated above, the Aesthetic is more about providing a general framework of the possibility and limits of human experience and the fundamental role time plays in it. At best, then, the Aesthetic is a kind of structural ontology of the mind, one crucial feature of which is that the mind takes in the

appearing objects temporally, that is, situates them before and after one another or represents them as simultaneous—or as both succeeding and simultaneous, in which case the mind represents persistence (see B67).

As it is often put, time for Kant is a way of representing rather than a separately existing entity of its own. In arguing towards this conclusion, Kant points out, for example, that time is not a product of experience. This is to say, among other things, that the actual experiences of simultaneity and succession are grounded in time, but they do not constitute time. An important part of Kant's argumentative strategy is that time simply cannot be a product of experience, because otherwise certain strictly universal principles about time would not hold. His example of such a principle is that, necessarily, different times are not simultaneous, as spatial objects are, but successive. Another feature of Kant's argumentative strategy is that unlike objects and their relations, time itself cannot be thought away. He gives a similar treatment to space. From this Kant concludes that space and time are the necessary forms of all humanly (but see B72) possible cognition, which underlie the actual experience of objects.

To put it differently, time reflects a structural feature of perceptual experiences and world-related thinking—or even wholly abstract thinking, for that matter (say, in thinking of a non-sensible object with two contradictory predicates— x at t_1 and not- x at t_2 —you necessarily employ time). In a nutshell, one of the main Kantian ideas appears to be that there must be something in our cognitive makeup that makes it so that whatever we are ever going to experience, it will be represented as a spatially extended something occupying some place somewhere in time. Moreover, we can only represent a particular amount of time, duration, alteration, and so on, by setting it against something that surpasses every temporal instance. This structure must also be constant: time itself does not change.

Kant further argues that time has only one dimension. This can be illustrated with time's arrow: time is linear, and temporal determinations always succeed or precede one another, as if they were on the same line that grows from indeterminate past to here and now and onwards. Time has no limits, either: time's arrow just extends. Different times—different locations and durations representable on time's arrow—are just parts or slices of one singular time. More precisely, the different times are ways of limiting time, or ways of representing such limitations. As Kant also puts it, time as such is not made from parts, but it is a whole that precedes the parts. This would also be one of the main reasons why

time is ultimately a non-discursive intuition instead of a discursive concept. At one point, Kant even adds, perhaps somewhat curiously, that time must be an intuition because temporal relations can be given a spatial expression (see A33/B50).

Does all the above indicate that Kant's philosophy of time is subjectivistic to the core? It indeed appears to be a crucial part of Kant's philosophical project to avoid any kind of reification of time. For Kant, it would mean a serious mistake to treat time as self-subsisting and thing-like. At the same time, it would also appear as a mistake to think that Kant's idealistic treatment of time turns time into something thoroughly subjective. To begin with, Kant does not seem to have much to say about subjective time in the sense the term is commonly used nowadays: namely, to refer to subjectively experienced time (see Arstila & Lloyd 2014). What's more, time for Kant is subjective in a very specific sense, while being objective in another sense. Time is subjective in the sense that it is not something pertaining to the objects and their relations as such. Instead, time is that which grounds the possibility of experiencing objects and their relations in the first place. At the same time, and for this very same reason, time is objective. It is just that for Kant objectivity indicates representation that can be considered as valid of objects—ultimately, reflecting the possibility of being in a relation to objects distinct from the subject—and time happens to be a necessary feature of our cognitive makeup in accomplishing that.

Similar ambiguity is present in the question about the reality or ideality of time. While time is not something concrete, or real in the absolute sense, but instead has to do with our cognitive constitution and way of representing objects, time is nevertheless as real as it gets: “namely the real form of inner intuition” (A37/B53) to use Kant's own expression. What Kant seems to want to underline is that we must necessarily treat time as ideal, albeit only in a very formal sense, or otherwise philosophical catastrophe follows. As Kant famously has it, a critical point of view requires that we interpret empirical objects as appearances instead of things in themselves, which in this context ultimately means that we cannot ignore our own constitution and frame of reference in the representation of objects. All of this, however, does not make time unreal or turn it into an illusion.

A further ambiguity follows from Kant's claim that time is the form of inner intuition or inner sense. At first sight, this move seems to tie temporal experiences very tightly to the self, or to “the way of representing myself as object” (A37/B54). This is to say that time is more like the foundation of the inner way of representing ourselves and our inner states, whereas space has more to do with representing

outer objects, that is, objects distinct from the self. One might find further support for this distinction from the idea that temporal representing does not appear as limited to the physical senses, and thus differs drastically from spatial representing, which is much more intimately linked to the physical senses and making sense of the external world through them.

At the same time, Kant connects time and space in ways which make the state-oriented picture of time just a part of the full story. For example, Kant speaks of attributing spatial *and* temporal predicates to the objects of the senses as such (B70*). Another indication of the intimate connection of space and time is that attributing a change of place to an object, or its disappearance from a place—it being in a certain place at t_1 and not-being there at t_2 —is grounded in time (see A31/B48). All in all, representing moving objects is fundamentally a spatiotemporal process. As Kant puts it, motion unites space and time (A41/B58). Perhaps he simply wanted to underline that motion is an “alteration of places” (A49/B66), the representation of which necessarily borrows elements from both time (alteration) and space (places).

Alternatively, one might regard Kantian space and time as a unified space-time rather than as two separate elements. If one were to take this route, there are some things to keep in mind, however. Though space and time come together in representing motion and alteration—in which case they are dependent on “empirical datum” (A41/B58)—as pure forms, the two appear as separate, or at least neatly separable. For example, shapes belong strictly to the domain of space, whereas pure thinking apparently does not require anything spatial at its disposal. That said, as we have seen, measuring the shape would be a temporal-synthetic process, and cognitively significant thinking would ultimately require a spatial component—that something that we can come across either in actual or possible experience. In the end, it thus seems that time is so intimately linked to space, and vice versa, that it is only in rare circumstances that one should take just one of them into consideration.²

University of Turku

² This essay is a modified version of my presentation at the *Time Workshop*, which I had the privilege to co-organize with Valtteri Arstila at the University of Turku in October 3-4, 2018. Valtteri, I wish you a happy birthday and look forward to further collaboration in the future.

References

- Arstila, Valtteri and Dan Lloyd (eds.) (2014). *Subjective Time: The Philosophy, Psychology, and Neuroscience of Temporality*. MIT Press, Cambridge MA.
- Janiak, Andrew (2022). "Kant's Views on Space and Time", in Edward N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy*.
<<https://plato.stanford.edu/archives/sum2022/entries/kant-spacetime/>>
- Kant, Immanuel (1900-). *Gesammelte Schriften*. Königlich Preußische Akademie der Wissenschaften (vols 1-22), Deutsche Akademie der Wissenschaften zu Berlin (vol 23), Akademie der Wissenschaften zu Göttingen (vols 25-29) (eds.). De Gruyter, Berlin. (=AA)
- Kant, Immanuel (2000a). *Critique of Pure Reason*. P. Guyer & A. Wood (eds. & transls.). Cambridge University Press, Cambridge. (=A/B)
- Kant, Immanuel (2000b). *Critique of the Power of Judgment*. P. Guyer (ed.), P. Guyer & E. Matthews (transls.). Cambridge University Press, Cambridge.
- Kant, Immanuel (2001). *Lectures on Metaphysics*. K. Ameriks & S. Naragon (eds. & transls.). Cambridge University Press, Cambridge.
- Kant, Immanuel (2002a). *Metaphysical Foundations of Natural Science*. M. Friedman (transl.). In: H. Allison & P. Heath (eds.), *Theoretical Philosophy after 1781*. Cambridge University Press. Cambridge.
- Kant, Immanuel (2002b). *Prolegomena to Any Future Metaphysics That Will Be Able to Come Forward as Science*. G. Hatfield (transl.). In: H. Allison & P. Heath (eds.), *Theoretical Philosophy after 1781*. Cambridge University Press, Cambridge.
- Kant, Immanuel (2003). *On the Form and Principles of the Sensible and the Intelligible World [Inaugural Dissertation]*. In: D. Walford & R. Meerbote (transls. & eds.), *Theoretical Philosophy 1755–1770*. Cambridge University Press, Cambridge. (AA 2)
- Kant, Immanuel (2004). *Lectures on Logic*. J. Michael Young (transl. & ed.). Cambridge University Press, Cambridge. (AA 9)
- Kant, Immanuel (2005). *Notes and Fragments*. P. Guyer (ed.), C. Bowman, P. Guyer & F. Rauscher (transls.). Cambridge University Press, Cambridge.
- Rosenberg, Jay (2005). *Accessing Kant: A Relaxed Introduction to the Critique of Pure Reason*. Clarendon Press, Oxford.
- Watkins, Eric (2005). *Kant and the Metaphysics of Causality*. Cambridge University Press, Cambridge.

Dynamic Snapshots and Artificial Temporality

DAN LLOYD

At the railroad station he noted that he still had thirty minutes. He quickly recalled that in a cafe on the Calle Brazil ... there was an enormous cat which allowed itself to be caressed as if it were a disdainful divinity. He entered the cafe. There was the cat, asleep. He ordered a cup of coffee, slowly stirred the sugar, sipped it ..., and thought, as he smoothed the cat's black coat, that this contact was an illusion and that the two beings, man and cat, were as good as separated by a glass, for man lives in time, in succession, while the magical animal lives in the present, in the eternity of the instant. (J.-L. Borges, "The South" tr. Andrew Hurley. Collected Fictions. New York, Viking, 1998.)

Valtteri Arstila (2018) has proposed a model of subjective time which highlights the presence of "pure" phenomenology of temporality, nonsensory but vivid awareness of motion, succession, and change, but unaccompanied with actual percepts of change, movement, or sequence. Present sensory information plus pure temporal feelings constitute subjective temporal experience, "dynamic snapshots", a theory that streamlines temporality by eliminating an appeal to the specious present. In this essay I look at temporality in a current Large Language Model, GPT-4 from OpenAI. The GPT architecture replaces the earlier standard of Recurrent Neural Nets (RNNs). RNNs have resources for embodying Husserlian temporality. In contrast, recurrent processing is absent from GPTs, implying that whatever cognition they enact or simulate, will be produced without temporal context. GPT errors are accordingly errors of deficient temporal processing. The dynamic snapshot model enables a novel option for correcting this chronic flaw, while escaping the limits of recurrent nets.

1. Introduction

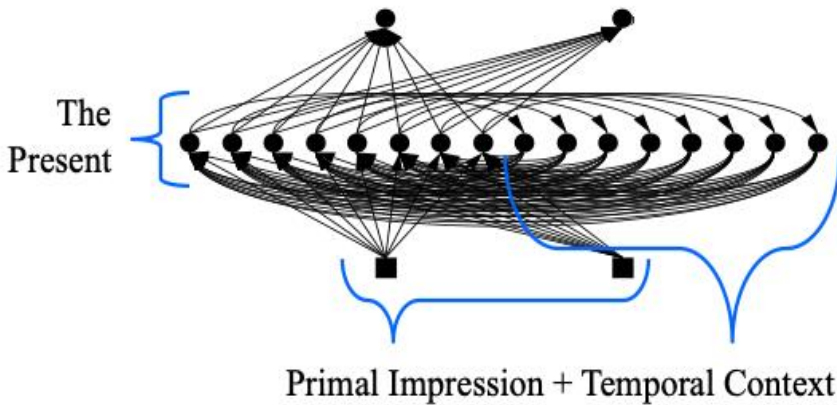
We move through time, the objective time of clocks and calendars. And time moves through us, the subjective experience of succession and duration, “felt time” (Wittmann, 2016). Subjective time has historically been the province of phenomenology, under the profound guidance of Edmund Husserl (Husserl, 1964). Husserl invites us to attend to our experiences (of all sorts) and describe them in detail, but additionally argues as a philosopher (in the tradition of Kant) that the main features of subjective time are necessary components of all human experience. After Husserl, the intricacies of experienced time are required components of the phenomenological study of everything human.

A century later, with the rise of the sciences of mind, it is natural to pursue temporality empirically, to sharpen its outlines and ultimately to discover its component neural mechanisms. This has not been easy. Subjective time, being omnipresent, cannot be readily bracketed in controlled experiments. The empirical evidence is often indirect, and requires careful conceptual analysis on the way to any conclusion. Yet progress has been made, and at the forefront of temporality research we find the work and the energy of Valteri Arstila. We’ll return to his work below.

In this essay I’d like to transplant the empirical study of temporality into a novel domain: Artificial Intelligence. AI underwent a sea change in the last decade, with a specific turning point in 2017 with the paper entitled “Attention is All You Need” (Vaswani et al., 2017) . The paper introduced the Transformer architecture, which has become the underlying functional architecture of Generative Pretrained Transformers, including the GPT series from OpenAI, Meta’s LLaMa, Anthropic’s Claude, Google’s Gemini, and others. In this paper, GPT will denote all these contemporary Large Language Models (LLMs). Transformers, plus numerous bells and whistles, have replaced the earlier state-of-the-art architecture, namely, recurrent neural networks (RNNs). This sea change from one basic architecture to another invites many comparisons. Fortuitously, RNNs and GPTs are very different in their processing of time, a difference which I hope will illuminate the source of some conspicuous failures of GPTs.

2. Recurrent processing and temporality

In other work, I've proposed that recurrent neural network (RNN) architecture is well suited to the representation and use of temporal information, with embodied distinctions reflecting basic Husserlian phenomenology (Lloyd, 2004). The figure below sketches the basic architecture of a recurrent net. It reproduces the time-honored icon of a three-layer network, with input, "hidden," and output layers, information flowing from bottom to top. However, in the hidden layer we find an innovation, a secondary set of hidden units connected only to other hidden units. This adjunct hidden layer, a "context layer," copies the most recent state of the regular hidden layer. That copy is then re-presented alongside the next input, resulting in the architecture shown in the diagram. The hidden layer now receives its input from two sources, from the input layer (the environment) and from itself – one cycle in the past (Elman, 1990). The internal recurrence effectively opens a new dimension for nets. Now the network responds not just to the current input, but to its own most recent internal state. Since that echo dates from one cycle ago, the next internal state merges both present and past information. Then, that new internal state, combining present and immediately prior information, itself recycles, and the successive state can forward information from two cycles previous. And so on. In principle, this neuronal hall of mirrors can keep a pattern alive indefinitely. Husserl characterizes subjective temporal experience as only partly constituted by the immediate stimulus (which he calls the "primal impression"). But it equally comprises two other phases or aspects: *Retention*, the lingering awareness of what has just happened, maintaining a shadow timeline originating at the primal impression and sinking into the past, only to fade at a vague temporal horizon. And *protention*, or anticipation, the constant expectation of the immediate future (Husserl, 1964). Recurrent nets readily condition their behavior on the immediate past, and anticipate future inputs. The Husserlian phenomenology of time neatly aligns, with the context layer analogous to retention. When they are tasked with anticipating the next input, they also model anticipation. These are networks with a computational version of temporality (Lloyd, 2004).



In the decades since the first recurrent nets, models have grown exponentially more complex, while still exploiting the recursion afforded by their characteristic feedback loops. But despite several innovations, RNNs inevitably face the attenuation of information as input sequences lengthen, the so-called “vanishing gradient.” A recurrent net has trouble detecting and using long-range dependencies. When sequences are presented token-by-token, the vanishing gradient is a temporal horizon. But what is a computational bug might be a phenomenological feature. Husserlian temporal consciousness acknowledges that retentive and protentive tails weaken as intervals from the present moment increase. He describes this fade as providing a moving temporal phenomenological horizon in experience. Because this horizon spans seconds or minutes but not longer durations, Husserl adds a separate capacity for explicit recall or prediction (and this is also familiar in our phenomenology).

3. Transformers in trouble


The vanishing gradient problem faced by recurrent neural nets nonetheless motivated a search for alternative architectures, inspiring the “Transformer” architecture, the backbone of popular LLMs. Contemporary neural nets are much greater in size and complexity than the elementary RNN above, and transformers especially so, involving a new lexicon of technical terms. (<https://poloclub.github.io/transformer-explainer/> includes a nice interactive tutorial.) My description here points out just a few of the relevant components of these models. To begin, “Transformer” describes their overall function: their job

is to transform inputs into outputs. In this regard, then, many models, including RNNs, are transformers. However, the “transformer block” also denotes part of the core computational architecture, including “attention heads.” Attention in this context is analogous to its regular usage, denoting the selection of salient items from a larger set. In transformer models, attention ranges over all the tokens (words, approximately) in the vast training corpus. Each token is “embedded” in a unique vector that encodes its proximity in the training set to all the other tokens. From token to token the attention mechanism computes the likely direction of discourse. All of this is iterated through multiple stacked layers of encoding, attention, and decoding (Vaswani et al., 2017).


The core innovation, and surprise, of this architecture is the absence of recurrent processing. As the title claims, attention is all you need. By ditching the recurrent feedback, transformers can exploit long term dependencies, relationships that are lost over the relatively near horizon afforded by recurrent feedback. In transformers, gradients don’t vanish. As a result, GPTs can frame responses to prompts that connect disparate elements of ongoing dialogue, leading to their striking performance.


Thus we have two very different families of Large Language Models, one of which includes an analogue of temporality, the other not. What difference does this difference make? Consider this dialogue with the current flagship of LLMs, GPT-4, recorded in the fall of 2023 (I am DA):


: Please sort this sentence by word length.

:by this word please length sentence sort.


: Are you confident in this answer?

: Yes, I am confident in the answer. The words in the given sentence are sorted by their length, with shorter words first and longer words later.


: "Sentence" is a longer word than "sort".


: I apologize for the mistake in my previous response. Here is the corrected sorting of the given sentence by word length:
this in by sort sentence

: "in" is shorter than "this"


: You are right, my apologies for the mistake. Here is the corrected sentence sorted by word length:


in by this sort sentence

: Good, however the original target sentence was "Please sort this sentence by word length."

: Apologies for the confusion. Here is the corrected sorting of the original target sentence:

by sort this Please length sentence

: Check again. Are you sure?

: Apologies for my previous mistakes. Here is the corrected sorting of the original target sentence based on word length:

by in this Please sort sentence length

This exchange shares general features with a raft of AI bloopers. (For many examples, and discussion, see (Arkoudas, 2023).) First, GPT can state the principle that is relevant to the task, apparent also in most of the examples reviewed by Arkoudas. Usually these generalizations are correct, and likely to be found in the gargantuan training corpus. But the model fails to correctly apply the principle, instead generating a mashup of accurate and inaccurate replies. Across any exchange of multiple prompts and replies, GPT routinely contradicts itself. It looks initially like the model can be corrected, but this turns out to be an illusion (Stechly et al., 2023). It routinely apologizes and often repeats the governing principle from the inaugural prompt, and generates a new answer, which frequently contains a new instance of error. At every turn, GPT initially claims that it is confident in its answers, but will immediately rescind any claim, even if the interlocutor introduces an error, or questions the accuracy of a response. In short, GPT-4 is using the words appropriate to solving the problem posed, but seems disconnected from its own prescriptions. It seems to be guessing, but that is not the only deficit. It also seems to forget its guess from moment to moment, contradicting itself.

From this, many authors conclude that LLMs (including GPT-4) can't reason (Arkoudas, 2023; Berglund et al., 2024; Dziri et al., 2023; Marcus, 2020, 2023; McCoy et al., 2023; Tong et al., n.d.; Wu et al., 2024)). That's true but its failures often emerge from a more basic problem, namely, that GPT-4 doesn't keep track of its own outputs. It doesn't notice when it flips its position, or blatantly

fails to follow the very principle it can state. In the example above, the request to reconsider the problem doesn't generate an "aha moment," when the model notices its own failure and thereafter corrects it. In short, GPT-4 doesn't learn from its mistakes. This gap in this LLM's self-monitoring is invisible to the model itself. That is, the model doesn't detect its own forgetfulness, but instead confabulates merrily into absurdity. Even simple tasks, like counting, fail, seemingly due to this same basic inability to hold "in mind" the current count in a set of items to be counted. Let's shorthand these observations with the general conclusion that the model is *oblivious to itself*.

4. A profound temporal deficit

Several authors maintain that LLMs are incorrigible, often appealing to a priori arguments based in the limits of computation itself (Marcus, 2022). The obliviousness here seems shallower, a contingent property of this LLM. Nonetheless, the ubiquity of debunking examples indicate that this wall is basic to this species of AI, a point to which we return. In this context, what does it mean to be oblivious to itself? Do these incapacities reflect a single underlying deficit?

The sticky and confabulated errors of the model reveal contrasting aspects of temporality we humans exploit routinely. The continuous assembly of an experienced world enables us to compare past to present, and predict our futures. Most important, our retentional and protentional awareness enables us to detect continuity and consistency, and conversely, change and inconsistency. When there's a disconnect, an inconsistency, we notice. The experience of temporality is arguably the essential and continual feature of consciousness, at least as humans experience it. (Arstila & Lloyd 2014; Husserl, 1964; Lloyd, 2004). Without it, we have no more sensory information than a video camera, exhausted by the current frame. GPT-4 apparently lives in a thin atemporal world, revealed in errors of consistency, continuity, and reasoning. It compensates for this perceptual poverty via fantastically complex reflexes, its moves forgotten as soon as they are made. Nonetheless, its seeming temporality is a pose. (This deficit may underlie the powerful impression that the LLMs are "competent without comprehending" (Dennett, 2023).)

Animals without temporal awareness must respond at all times to the immediate present. They can learn, of course, when a behavior backfires or succeeds, and with enough learning their immediate responses can change. But for

them even very complex responses are reflexes. They don't experience sequences as such, and they don't experience the arc of their own behavior. I hypothesize that GPT-4 is an entity of this sort, a gigantic reflex engine without a mechanism implementing Husserlian temporality, as described above. Its obliviousness reflects the absence of a temporal world. GPT-4 is like the cat in the epigraph to this essay, its feline experiences inconceivable to humans, "for man lives in time, in succession, while the magical animal lives in the present, in the eternity of the instant."

This extended digression affords an answer to an inevitable question, What is temporality for? The AI model contrasts offer a distinct answer, one that is less emphasized in most discussions of subjective time: Humans use temporality as a continuous compare-and-contrast operation. The circle of immediate awareness is variable and often sketchy, but it is implicitly crosschecked every second. The evolutionary advantages of immediate and continuous self-monitoring are apparent. Without this omnipresent self-awareness, even elaborate verbal productions would be guesses, however glib.

5. The pure experience of time

The discussion above, if accurate, consolidates a number of GPT's chronic errors, suggesting they are various manifestations of a single deficit. A similar consolidation occurs in Valtteri Arstila's "dynamic snapshots" (Arstila, 2018). Valtteri consolidates several intriguing experiments around a single model of temporality, a model situated between two others. On the one hand, the "cinematic" model denies any direct subjective experience of time. Instead, we experience a flow of sensory contents, each a snapshot, like frames in a movie, and everything temporal is an inference from the immediately present sensorium and background knowledge. I agree that this model is much too thin – we do have temporal *experiences*, distinct from inferences about future and past. These experiences are non-sensory (Husserlian apprehensions (Husserl, 2010)). The other model, though, is more capacious. Valtteri contrasts his view with two variants of the "specious present." I agree that a "present moment" that is extended in time is ultimately incoherent. But the other variant he considers is straightforwardly Husserlian, the "retentionalist" model. The experienced present is a knife-edge, but one that includes explicit content representing the immediate past and immediately expected future. The question is, how explicit are these

retentions? Valteri reviews experiments (and illusions) affording “pure temporality,” that is, a feeling of movement, succession, or change in the absence of actual perceived change.

Perhaps the simplest example of pure temporality is the experience of duration. Waiting... for the train to arrive, the light to change, the guests to appear, or just plain viewing a static scene. We’re aware of the unwinding extension of the passing moments, moments however that pass without any sensory change. Such experiences are commonplace but might be misdescribed as pure duration, for the reason that our minds are always in motion (and our eyes as well). The stasis of the perceived scene may be approximately true of the physical environment, but the phenomenology is an impure mix of small changes. What we might describe as blank boredom is phenomenally dense, rather than a pure swelling void.

Valteri focuses on other experiences of pure temporality: the experience of motion in the absence of perceived change, as in the waterfall illusion. The pure experience of causation, spontaneously felt when events are in the right temporal relationship. And the pure experience of succession, which at short time scales can be dissociated from the awareness of which of two events occurred first (Arstila, 2018).

These pure temporal experiences are distinct from Husserlian retentions in their relative lack of explicit content. The experience of a falling snowflake, according to this view, has two components. One is the change of position of the flake, while the second is the pure awareness of motion in the visual field. The dynamic snapshot is the immediate perceptual scene plus the evoked experiences of pure temporality. Valteri asserts that the snapshot and attendant pure temporality are sufficient to account for the phenomenology of subjective time. Evaluating his theory is beyond the scope of this chapter, but the focus on pure temporality introduces a useful further articulation in theories of temporality.

In this context, then, we can return to the contrast between RNNs and GPTs with new questions. In particular, what is the computational version of pure temporality? Do either of the main neural network architectures support (or refute) the possibility of pure time?

Pure temporality is relatively content-free, but signals crucial perceptual change: Even before one knows what has changed, one knows that there has been change, and this immediately triggers an attentive search for what has altered. GPTs, as in the example above, are cheerfully blind to one prominent change,

namely their own inconsistency from moment to moment. That is, their assertions are changing in manifestly contradictory directions, but no alarm bell directs their self-examination.

Socrates famously claimed that the unexamined life was not worth living. Generative Pretrained Transformers might do well to listen to Socrates' advice, but not for its contribution to a satisfying, or ethical, life. Rather, this essay has highlighted a continuous and transient form of self-examination, basic to humans but missing in GPTs. This continual self-monitoring has a conspicuously epistemic function, offering a running check on reality. The cross-check spans brief intervals, but their overlapping sustains a consistent lived world.

Evolution would favor efficiency in this process, as in all biological processes. In the hurly-burly of experience, brains need to know promptly if there's trouble. This suggests a role for pure temporality, as the alarm bell that grabs attention to a possible problem. If, as Valteri conjectures, pure temporality arises through dedicated, encapsulated neural mechanisms, then it could serve as the quick alert, to be followed by a more careful review involving more sophisticated cognitive capacities. A similar strategy is often cited with respect to interactions between the limbic system, especially the amygdala, and cortex (Pessoa, 2010). The amygdala seems to be dedicated to quick threat detection (Snake!), rapidly followed by higher cortical analysis (Relax, it's just a stick.). Threat detection is always an urgent matter, but so is the detection of change, especially inconsistency. Such detections warn a system that something may be wrong, and motivate a search for the cause of the discrepancy.

6. Toward temporality in GPTs

This discussion may suggest some advice on the future development of intelligent architectures. Recurrent neural networks involve a thorough and total retention of each moment of experience, a heavy computational task (and phenomenologically weighty as well). GPTs enlarge their cognitive horizons by eliminating recurrence, but at the price of the many failures of self-oblivion. Could there be a middle way? The existence of pure temporal experiences invites us to look for a quick-and-dirty retentional cross-check, preceding any awareness of retentional contents. This arguably requires reentrant processing, but perhaps at some primitive or early level in the cascade of activation. This is compatible with Valteri's suggestion that pure temporality is mediated by dedicated, encapsulated, and domain specific modules.

Valtteri's work clarifies the minimal information needed for subjective time in its AI analogue. LLMs could make good use of a simple signal announcing change and thereby alerting the system to a need to confirm its consistency over time. Given the manifest failures of AI in so many elementary scenarios, it could be a worthy engineering goal to install in AI systems a bit of subjective time. Such mechanisms could intervene early in the cascade of neural processing, or not, but in any case their job is simply to direct attention to change (including motion and succession). In discourse, then, such a change might be from one assertion to its contradiction, as in the GPT example above. The retentionalist model holds that contents are explicitly encoded and retained in order as time passes. This discussion, following Valtteri's lead, suggests that this elaborate and changing content is unnecessary and accordingly maladapted. Instead, an alarm sounds when there's change; the alarm directs us to compare the present to an item in working memory. As far as subjective time is concerned, a pure temporal monitor is all you need.

Stepping back still further, we see again for humans and AI alike the radical and foundational role of the experience of time. In Valtteri Arstila's work, we find many excursions in this fascinating domain (Arstila, 2012, 2015, 2017, 2018; Arstila & Lloyd, 2014). (Here I've discussed just one of several papers and volumes on time.) There will be more, a welcome subjective future of research and discovery.¹

Trinity College, Connecticut, USA

References

- Arkoudas, K. (2023, August 9). GPT-4 Can't Reason. *Medium*.
https://medium.com/@konstantine_45825/gpt-4-cant-reason-2eab795e2523
- Arstila, V. (2012). Time Slows Down during Accidents. *Frontiers in Psychology*, 3.
<https://doi.org/10.3389/fpsyg.2012.00196>
- Arstila, V. (2015). Keeping postdiction simple. *Consciousness and Cognition*, 38, 205–216. <https://doi.org/10.1016/j.concog.2015.10.001>
- Arstila, V. (2017). Experience and the pacemaker-accumulator model. *Journal of Consciousness Studies*, 24(3–4), 14–36.

¹ Acknowledgements. For many conversations on time and AI, I thank Cheryl Greenberg. Three paragraphs in this paper are adapted from (Lloyd, 2024).

- Arstila, V. (2018). Temporal Experiences without the Specious Present. *Australasian Journal of Philosophy*, 96(2), 287–302. <https://doi.org/10.1080/00048402.2017.1337211>
- Arstila, V., & Lloyd, D. (Eds.). (2014). *Subjective Time: The Philosophy, Psychology, and Neuroscience of Temporality*. MIT Press.
- Berglund, L., Tong, M., Kaufmann, M., Balesni, M., Stickland, A. C., Korbak, T., & Evans, O. (2024). *The Reversal Curse: LLMs trained on “A is B” fail to learn “B is A”* (arXiv:2309.12288). arXiv. <https://doi.org/10.48550/arXiv.2309.12288>
- Dennett, D. C. (2023). *I’ve Been Thinking*. W. W. Norton & Company.
- Dziri, N., Lu, X., Sclar, M., Li, X. L., Jiang, L., Lin, B. Y., West, P., Bhagavatula, C., Bras, R. L., Hwang, J. D., Sanyal, S., Welleck, S., Ren, X., Ettinger, A., Harchaoui, Z., & Choi, Y. (2023). *Faith and Fate: Limits of Transformers on Compositionality*.
- Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, 14(2), 179–211. [https://doi.org/10.1016/0364-0213\(90\)90002-E](https://doi.org/10.1016/0364-0213(90)90002-E)
- Husserl, E. (1964). *The Phenomenology of Internal Time-Consciousness* (Fourth Printing edition). Indiana University Press.
- Husserl, E. (2010). *Thing and Space: Lectures of 1907* (R. Rojcewicz, Trans.; Softcover reprint of hardcover 1st ed. 1998 edition). Springer.
- Lloyd, D. (2004). *Radiant Cool: A Novel Theory of Consciousness* (1st edition). A Bradford Book.
- Lloyd, D. (2024). What is it like to be a bot? The world according to GPT-4. *Frontiers in Psychology*, 15. <https://doi.org/10.3389/fpsyg.2024.1292675>
- Marcus, G. (2020). *GPT-3, Bloviator: OpenAI’s language generator has no idea what it’s talking about*. MIT Technology Review. <https://www.technologyreview.com/2020/08/22/1007539/gpt3-openai-language-generator-artificial-intelligence-ai-opinion/>
- Marcus, G. (2022, March 10). Deep Learning Is Hitting a Wall. *Nautilus*. <https://nautil.us/deep-learning-is-hitting-a-wall-238440/>
- Marcus, G. (2023, March 15). GPT-4’s successes, and GPT-4’s failures [Substack newsletter]. *The Road to AI We Can Trust*. <https://garymarcus.substack.com/p/gpt-4s-successes-and-gpt-4s-failures>
- McCoy, R. T., Yao, S., Friedman, D., Hardy, M., & Griffiths, T. L. (2023). *Embers of Autoregression: Understanding Large Language Models Through the Problem They are Trained to Solve* (arXiv:2309.13638). arXiv. <https://doi.org/10.48550/arXiv.2309.13638>
- Pessoa, L. (2010). Emotion and Cognition and the Amygdala: From “what is it?” to “what’s to be done?” *Neuropsychologia*, 48(12), 3416–3429. <https://doi.org/10.1016/j.neuropsychologia.2010.06.038>
- Stechly, K., Marquez, M., & Kambhampati, S. (2023). *GPT-4 Doesn’t Know It’s Wrong: An Analysis of Iterative Prompting for Reasoning Problems* (arXiv:2310.12397). arXiv. <https://doi.org/10.48550/arXiv.2310.12397>
- Tong, S., Jones, E., & Steinhardt, J. (n.d.). *Mass-Producing Failures of Multimodal Systems with Language Models*.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). *Attention Is All You Need* (arXiv:1706.03762). arXiv. <https://doi.org/10.48550/arXiv.1706.03762>

- Wittmann, M. (2016). *Felt Time: The Psychology of How We Perceive Time* (E. Butler, Trans.).
- Wu, Z., Qiu, L., Ross, A., Akyürek, E., Chen, B., Wang, B., Kim, N., Andreas, J., & Kim, Y. (2024). *Reasoning or Reciting? Exploring the Capabilities and Limitations of Language Models Through Counterfactual Tasks* (arXiv:2307.02477). arXiv. <https://doi.org/10.48550/arXiv.2307.02477>

What If Time Were Like Other Minds

BRUNO MÖLDER

It is commonplace for philosophers to claim that we experience time in one way or the other. My aim in this little essay is to elaborate on one possible way of clarifying what it is to experience time. I am not proposing an analysis of what those philosophers themselves could mean when they say that we experience time. This itself is not a mystery, as we shall see below; it is quite clear what they mean. Instead, my question concerns the nature and character of this experience: if it is indeed an experience, what kind of experience might it be? To make progress on this question, I suggest looking at the field of social cognition to see if borrowing some notions from this field could help make sense of the experience of time.

1. Experiencing time: some examples

Here are some examples of what philosophers have said about the experience of time. For instance, Barry Dainton claims that although we do not perceive time as such, we have direct experiences of relationships that involve time, such as change.¹ He proceeds from the example of someone waiting behind the dentist's door and noticing that the hands of the clock have moved:

Time itself is not something we can perceive—we do not have a sensory organ that can discern time *per se*—but we can perceive something intimately related to time, namely *change*. Returning to the dentist's waiting

¹ Many authors use the terms “experience” and “perception” interchangeably. In some contexts, “perception” intuitively fits better than “experience”. The reason might be that “experience” has a phenomenal aspect, whereas perception need not be phenomenally accessible, as there can be unconscious perceptions and, therefore, an unconscious perception of time. Then again, the notion of perception is tied to what is present, thus we arguably can only perceive what is present. Given this, Power (2018, 40–42) writes about non-perceptual experiences, which include memory experiences. I primarily use “experience”, except in contexts where I engage with authors who write about perception.

room, let us suppose that the clock is now showing 10.29, and for the past minute or so you have been intently observing the steady advance of the *second*-hand around the clock face. Whereas previously you *inferred* the passage of time, relying on the time currently visible and recollections of earlier showings, now you are—in a fashion—*directly perceiving* the passage of time. (Dainton 2008, 620)

The clock example derives from C. D. Broad (1923), who contrasted inferring that the clock's hand has moved with directly seeing the hand move.

But it is a notorious fact that we do not merely notice that something *has* moved or otherwise changed; we also often see something *moving* or *changing*. This happens if we look at the second-hand of a watch or look at a flickering flame. These are experiences of a quite unique kind; we could no more describe what we sense in them to a man who had never had such experiences than we could describe a red colour to a man born blind. It is also clear that to see a second-hand *moving* is a quite different thing from “seeing” that an hour-hand *has* moved. In the one case we are concerned with something that happens within a single sensible field; in the other we are concerned with a comparison between the contents of two sensible fields. (Broad 1923, 351)

Whereas Dainton claims that we are able to perceive the *passage of time* directly, Broad's remark applies to seeing motion. One might find Dainton's last sentence far-fetched, as merely staring at the movement of the second hand of the clock gives you a direct perception (given a suitable interpretation of “direct”) merely of the second hand of the clock, and not the passage of time itself. However, we could agree with Dainton's more general point that we can often perceive things changing. It is important that the mode of access to movement and change is direct or non-inferential. Here's a statement from Robin Le Poidevin:

[O]n the one hand, our perceptions inform us of the order and duration of events. Moreover, the process appears to be a *non-inferential* one: we can become aware of temporal features without having consciously to derive them from other information. But, on the other hand, the usual models for ordinary perception, of shape and colour for example, are not applicable to

time. In particular, order and duration are not objects of perceptual states (Le Poidevin 2004a, 112, *italics mine*).

Elsewhere, Le Poidevin (2019, 1) writes that since the perception of time is not linked to any particular sense modality, “perhaps ... we notice time through perception of other things”. Thus, what we experience directly or non-inferentially are events and things in various temporal relations. By having those experiences, we are said to experience time or to have temporal experiences.

What are the other temporal relations between events or temporal properties of events besides change and movement that we are said to experience in this manner? The experience of change is a close kin to the experience of persistence, that is, the sense of something persisting over a period of time. A compelling auditory example is Sean Kelly’s (2005, 208) description of a singer holding a high note for so long that, in addition to the note’s auditory qualities, one senses it as persisting for too long. In addition, we seem to be aware of temporal order within and between events. An example of the first could be an experience of a temporally extended tune, the elements of which are temporally ordered. Examples of the other would be awareness of simultaneity or succession of stimuli.²

Some antireductionists about the ontology of time claim that we also experience such entities as Nowness and the Passage of Time. Then, they argue for the existence of an ontological present and temporal passage as the best explanation of these respective experiences (Paul 2010, 338–339). I do not want to include these kinds of experiences for the present purposes. One reason is that the above-described experiences of change, motion, and temporal order are already sufficiently puzzling. The second reason is that the case that we really experience such things as Nowness and Passage has not been made convincingly, given that there are alternative explanations for why we think we have experiences of now and temporal passage. For instance, the first phenomenological datum should be that the events we experience seem to exhibit the property of nowness. However, this can be taken to fall out from the fact that every experience happens now. The second phenomenological datum should be that the events seem to pass. However, this datum is based on the (putative) experience of change, which can

² I have omitted duration from this list because it is less than obvious that we can experience duration as such, as opposed to estimating its length.

be explained in terms of brain processes instead of some kind of direct apprehension of the real passage.³

2. The rich content of temporal experience?

Let us now return to the experience of time, understood as a non-inferential experience of temporal properties of events. Are there good reasons to call this mode of access to temporal properties an *experience*? Let us call this question(1):

(1) What is it to experience temporal properties?

One may think that a way to approach this question is to take it as a query about what is represented in the content of experience:

(2) What kinds of properties are represented in the content of temporal experience?

Could answering question (2) give us an answer to question (1) as well? Regarding visual experience, there are two competing approaches to properties in experience (Siegel 2010). In one view, only so-called *low-level* properties are represented in the content of visual experiences, that is, colours, shapes, motion, spatial location, and illumination. In the contrasting approach, which Susanna Siegel (2010) calls “the Rich Content View”, *high-level* properties are also represented, such as kind properties, emotional properties, mental properties, and semantic properties.

Without attempting to offer an answer to question (2) and delving into the reasoning behind the Rich Content View, I will only present three remarks on this point to show that my question (1) differs from question (2). At first glance, it may seem that the very claim that we experience temporal properties requires the Rich Content View, but this need not be the case. If it suffices for temporal experience to establish that we experience motion, then the low-level view could be sufficient. It is not necessary to presume the Rich Content View for the temporal case.

Second, it is safe to assume that, according to the Rich Content View, the properties represented in the content of experience make a difference in one’s phenomenology. Thus, some comments on phenomenology are in order. If phenomenology includes both experiential and cognitive aspects, then answering question (2) does not necessarily answer question (1). Let’s elaborate. Assume that

³ See Paul (2010), who approaches these issues broadly in this way. For some other options, cf. Baron and Miller (2019, Ch. 3).

the Rich Content View applies to the temporal case and that temporal properties are part of phenomenology. However, this does not yet show that the mode of access to these temporal properties is purely experiential, for it is possible that temporal properties are related only to the cognitive aspects and not the experiential aspects of the overall phenomenology.

Third, setting aside phenomenology for a moment, let us assume that all the above-listed temporal properties and relations—change, motion and temporal order—are represented in the contents of mental states. This still leaves the following question open: What kind of epistemic access do we have to those properties? For example, is it an *experience* rather than a *cognitive judgment*?⁴ What makes it one or the other? Therefore, the answer to the question about the properties represented in the content of experience does not yet settle question(1).

3. Learning from other minds

I harbour no illusions that question (1) can be answered in this short essay. However, I hope we can make some progress in understanding the notion of experience by shifting the narrative. Let us turn from the experience of time to social cognition.⁵ I want to explore whether we can gain insights into our epistemic access to temporal properties from discussions about how we know other minds. In the field of social cognition, several authors have proposed that our access to other minds is in some sense perceptual, for example, that we actually see other people being in pain. Perhaps we experience time and temporal properties similarly to how we experience other minds.

Why assume that the experience of time and the experience of other minds might be analogous in any way? There are some similarities. In both cases, we are supposed to experience something that, at first glance, seems hidden from plain sight. The standard orthodoxy still holds that we know other minds inferentially by first experiencing behaviour, but some authors influenced by phenomenology

⁴ This should be reminiscent of the heading of our one-time research project funded by VolkswagenStiftung “Subjective Time: Phenomenological Property or Cognitive Construct?” (2007–2009). Besides Valtteri and me, the project members were Cordula Becker, Niko Busch, Christine Falter, Julian Kiverstein, Valdas Noreika and Agnieszka Wykowska.

⁵ “Social cognition” is a common term for the field that studies how we make sense of other minds. Here “cognition” is taken in a broad sense without committing one to the view that the processes involved cannot be perceptual or experiential.

argue for a direct perception view. Similarly, in the case of time, we sense time through experiencing temporal relations such as change or motion directly.

Here's one classic quote from Max Scheler about "direct acquaintance" with another person's emotions.

For we certainly believe ourselves to be directly acquainted with another person's joy in his laughter, with his sorrow and pain in his tears, with his shame in his blushing, with his entreaty in his outstretched hands, with his love in his look of affection, with his rage in the gnashing of his teeth, with his threats in the clenching of his fist, and with the tenor of his thoughts in the sound of his words. If anyone tells me that this is not "perception", for it cannot be so, in view of the fact that a perception is simply a "complex of physical sensations", and that there is certainly no sensation of another person's mind nor any stimulus from such a source, I would beg him to turn aside from such questionable theories and address himself to the phenomenological facts. (Scheler 1954, 260).

Shaun Gallagher has made a similar point about the "immediate perception" of others' emotions and intentions:

In seeing the actions and expressive movements of the other person in the context of the surrounding world, one already sees their meaning; no inference to a hidden set of mental states (beliefs, desires, etc.) is necessary. When I see the other's action or gesture, I see (I *immediately perceive*) the meaning in the action or gesture. I see the joy or I see the anger, or I see the intention in the face or in the posture or in the gesture or action of the other (Gallagher 2008, 542).

Which notion of perception is used in such phenomenological approaches to intersubjectivity? Let us focus on Gallagher (2008). He elaborates only a little on what he means by "perception". He notes that perception, in his sense, is "smart", meaning that when looking at his red car, he can "simply and directly see" the car as opposed to seeing a meaningless red mass, which is then subject to some kind of cognitive interpretation (Gallagher 2008, 536).

Gallagher's proposal contains two important ideas for the present purposes. First, the idea that we perceive the meaning of others' actions directly, and second,

the denial that social cognition (at least in part) is a two-step process: that we first sense physical properties and then apply an inferential cognitive process of interpretation to these sensations. I will discuss each of these ideas and an additional one related to Husserl, focusing on potential analogues with temporal experience.

Perception of meaning

The idea that the meaning of other people's actions is directly given in social perception operates with a notion of perception that fits well with the Husserlian account of time consciousness. Briefly, the Husserlian model of time consciousness is a version of retentionalism that explains the apparent temporal extension of experience in terms of the representational or intentional content of a conscious state. In particular, in Husserl's (1991) view, consciousness has a complex, three-part structure: in addition to the primal impression of the present moment of the experienced object, there is the retention of the recent past and the protention of the next moment of the object. This structure is intentional in the sense that the retentive and protentive phases are *intentionally* present, providing both past- and future-oriented meaning to the act (see Gallagher 1998, 42–52, for the development of Husserl's view and more on the notion of intentionality; see also Gallagher and Zahavi 2008, 77–78).

If one conceives of experience in this manner, it allows us to talk about temporal *experience* in which the temporal aspects are intentionally present in experience. This comes with the cost of tilting the scales towards a retentionalist (Husserl 1991) and intentionalist (Kiverstein 2010) conception of temporal experience. However, this need not yet constitute a straightforward commitment to retentionalism, for the intentional presence of retention and protention need not be the only elements needed to account for temporal experience.

Non-inferentiality

The second idea is that to say we experience or perceive something is to say that we access it non-inferentially. Gallagher (2008, 537) contrasts direct perception with the two-step process in which something is “added to perception, e.g., an inference or interpretation that goes beyond what is perceived”. Similarly, the experience of time is not a two-step process—it's not as though we first observe things changing and then infer that we have experienced temporal properties. Such an inference might be made by a philosopher of time, but that would be a

theoretical judgement, not an experience. The temporal phenomenology of non-philosophers doesn't seem to involve such conscious inferences. Instead, we seem to access temporal properties like change, motion, and others non-inferentially, in Gallagher's sense. Recall also the statement by Le Poidevin (2004a, 112) quoted above. This is where the notion of non-inferentiality applies in the case of temporal experience.⁶ Of course, this only fixes the *explanandum*, and does not yet settle which would be the best way to explain the non-inferential experience of temporal properties.

Note that the non-inferentiality of experience in this sense is compatible with the idea that it is underpinned by inferential or constructive processes. Namely, non-inferential experience does not preclude subpersonal inferences⁷, either in the case of social perception or in the case of time. What phenomenally appears as direct experience may result from active interpretation and re-interpretation at the subpersonal level. To echo the title of our project once more, subjective time could be both phenomenological and constructed. Even if we presume for a moment that temporal experience does not involve cognitive processes, it could still be a matter of experiential (rather than cognitive) construction. As an experience, it is phenomenological, but at a non-experiential level, it is constructed by subpersonal processes.

Co-presentation

This is a third idea that originates from an attempt to make sense of the perception of other minds. Whereas the previously discussed notion of non-inferentiality

⁶ Krueger and Overgaard (2012, 307–308) point out that the difference between inference and visual experience is that inferences are cognitively penetrable in the sense of Pylyshyn (1999). While the result of an inference can be influenced by additional information, the visual experience (including the experience of facial emotions) cannot be influenced by non-visual information. The target's angry look persists even when you believe that she is not really angry. The impenetrability of some social perceptions would be one reason to argue that sometimes we perceive others' mental states. I am not sure how far this idea would carry, though. In the case of other minds, expectations influence what we experience. For the temporal case, the consideration of impenetrability would establish that it is truly an experience, not cognition, only if it can be shown that cognitive processes do not influence the experience of change and movement, etc. This is an empirical issue that cannot be settled here.

⁷ Gallagher (2008, 537) concedes that direct perception does not preclude "complex processes at a sub-personal or unconscious level", but he refuses to call those processes "inferences" because, presumably, inference is what a whole organism does, not its subpersonal components. In a sense, this is a verbal issue. It is commonplace to talk about inferences at the subpersonal level. However, we could distinguish between the personal and subpersonal notions of inference, and the claim would still stand.

concerned our experience of temporal properties, this one could hopefully shed light on the way “time itself” is given through the experience of temporal properties. Co-presentation (*Abschattung*) is a notion from Husserl’s philosophy that Joel Smith (2010) has employed to explain how we can perceive the mental states of others. Put simply, the idea is that in laughter (behaviour) joy (a mental state) is co-presented.

Husserl applied the notion of co-presentation to the perception of physical things that are presented in experience partially, or from one side only. In such cases, the missing part or side is co-presented. It is co-presented in the sense that it gives rise to anticipations, which could potentially be verified by the perceiver when she confirms that the co-presented part is actually there (Smith 2010, 737–738). In the case of other minds, Husserl (1960, §50) notes that verification by manipulating the object (looking at its other side) is not possible; however, he still maintained that they are co-presented in a distinctive manner, namely by anticipating further harmonious behaviour that co-presents mental states (Smith 2010, 741; see also Bohl and Gangopadhyay 2014, 213–215).

Smith (2010, 741) elaborates on the co-presence of mental states in terms of anticipations that “latch onto” mental states, understood in functional terms. If mental states are functionally defined as certain behaviour *B* following a certain input *I*, then one can see another person as having the relevant mental state when one anticipates that if *I*, then *B*. For example, let’s say that joy is functionally defined as the state one is in when one laughs upon hearing wonderful news. In that case, if the observer anticipates the other person to laugh upon wonderful news, joy is co-presented in laughter.

Could this idea be useful for explicating the experience of time? The analogy is far from perfect, as time and temporal properties are not, in relevant respects, conceived in terms of functional roles. However, abstracting away from the particulars of the accounts of Smith and Husserl, we could make use of Husserl’s idea that anticipation has a conditional “if...then...” structure, which is both “hidden” and “intentional” (see Husserl 1970, §47, quoted by Smith 2010, 738). We can also disregard the claim that this structure belongs solely to perceptual anticipation. What I would like to pick up from the co-presentation discussion is the idea of an intentional conditional structure and apply it to the amodal experience of time. The analogy would be that if we experience temporal properties of things and events, such as change, motion and temporal order, then we also experience time. Time is co-presented in temporal experience. Time is the

hidden intentional part of the experience of temporal properties. This is one way to cash out Le Poidevin's (2019, 1) point that "we notice time through perception of other things".

This chimes again with views that consider temporal experience to be an intentional matter. However, the difference from retentionalism is that, whereas the retentionalist approach attempts to explicate the overall intentional structure of time consciousness, the current proposal seeks to clarify how we grasp time through the experience of temporal properties.

4. Time for conclusion

Taking notes from social cognition to conceptualize the experience of time may seem like a stretch to some readers. There is an important difference between the two: the experience of time is more global and could even characterize consciousness as a whole, whereas social cognition is just one type of cognition among others, distinguished by its peculiar objects—other people. The nature of these two phenomena might also differ. Social cognition could be perceptual, while subjective time might be a cognitive affair; alternatively, social cognition might not be perceptual after all, whereas our mode of apprehending time could be experiential. However, the aim of this exercise was not to assimilate the experience of time to social cognition, but rather to borrow some concepts from the latter and explore their applicability to the context of temporal experience. The differences between the source and the target domain do not matter for such an aim.

I approached the initial question, "What is it to experience temporal properties?" by asking us to imagine how we might conceive of temporal experience if it were experiential or perceptual in the same way social cognition is said to be perceptual. We learned that this perspective offers a notion of perception where the classic distinction between experience and cognition becomes less straightforward. Perception is "smart"; it presents the world in meaningful units and does not divide into two consciously separate steps of sensing and making inferences. The experience of temporal properties would not require conscious inferences, though it does not preclude subpersonal constructive and inferential processes. On this view, both the non-inferential experience of temporal properties and the more elusive grasp of time itself would be intentional rather than purely sensory. However, they are intentional in different ways. The experience of

temporal properties would include the near past and anticipated future as a meaningful context of perception, whereas time itself is co-presented through the experience of temporal properties.

Finally, one may wonder if this is not a case of pushing at an open door. The retentionalist or intentionalist positions on time are already well known, so why not start with them, instead of arriving at a similar conclusion through such a winding path? My response is that this essay should not be read as a defence of a retentionalist or intentionalist account. This should be evident from the fact I could not address potential criticisms of the three analogies within this short space. Rather, it is an exploration of what we might discover if we were to examine the experience of time from an unusual angle. It has opened another path toward an account that incorporates intentional elements, but is far from being committed to the full-blown views of retentionalism or intentionalism.⁸

University of Tartu, Estonia

References

- Baron, S. and K. Miller (2019), *An Introduction to the Philosophy of Time*, Polity Press, Cambridge.
- Bohl, V. and Gangopadhyay, N. (2014), “Theory of Mind and the Unobservability of Other Minds”, *Philosophical Explorations* 17, 2, 203–222.
- Broad, C. D. (1923), *Scientific Thought*, Routledge & Kegan Paul, London.
- Dainton, B. (2008), “The Experience of Time and Change”, *Philosophical Compass* 3/4, 619–638.
- Gallagher, S. (1998), *The Inordinance of Time*, Northwestern University Press, Evanston, IL.
- Gallagher, S. (2008), “Direct Perception in the Intersubjective Context”, *Consciousness and Cognition* 17, 535–543.
- Gallagher, S. and D. Zahavi (2008), *The Phenomenological Mind: An Introduction to Philosophy of Mind and Cognitive Science*, Routledge, London.
- Husserl, E. (1960), *Cartesian Meditations: An Introduction to Phenomenology*, trans D. Cairns, Martinus Nijhoff, The Hague.
- Husserl, E. (1970), *The Crisis of European Sciences and Transcendental Phenomenology*, trans D. Carr, Northwestern University Press, Evanston.
- Husserl, E. (1991), *On the Phenomenology of the Consciousness of Internal Time (1893–1917)*, trans J. B. Brough, Kluwer Academic Publishers, Dordrecht.

⁸ This paper is based on my presentation on May 19, 2011, at the University of Turku during the workshop “From the Metaphysics of Time to the Phenomenology of Time”, organized by Valterri Arstila.

- Kelly, S. D. (2005), “The Puzzle of Temporal Experience”, in A. Brook and K. Akins (eds.), *Cognition and the Brain: The Philosophy and Neuroscience Movement*, Cambridge University Press, Cambridge, 208–238.
- Kiverstein, J. (2010), “Making Sense of Phenomenal Unity: An Intentionalist Account of Temporal Experience”, *Royal Institute of Philosophy Supplement* 67, 155–181.
- Krueger, J. and S. Overgaard, (2012), “Seeing Subjectivity: Defending a Perceptual Account of Other Minds.” in S. Miguens and G. Preyer (eds.), *Consciousness and Subjectivity*, De Gruyter, Berlin, 297–320.
- Le Poidevin, R. (2004), “A Puzzle Concerning Time Perception”, *Synthese* 142, 109–142.
- Le Poidevin, R. (2019), “The Experience and Perception of Time”, in E. N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy (Summer 2019 Edition)*, URL = <https://plato.stanford.edu/archives/sum2019/entries/time-experience/>.
- Paul, L. A. (2010), “Temporal Experience”, *Journal of Philosophy* 107, 7, 333–359
- Power, S. E. (2018), *Philosophy of Time and Perceptual Experience*, Routledge, Abingdon.
- Pylyshyn, Z. (1999), “Is Vision Continuous with Cognition? The Case for Cognitive Impenetrability of Visual Perception”, *Behavioral and Brain Sciences* 22, 341–423.
- Scheler, M. (1954), *The Nature of Sympathy*, trans Peter Heath, Routledge & Kegan Paul, London.
- Siegel, S. (2010), *The Contents of Visual Experience*, Oxford University Press, Oxford.
- Smith, J. (2010), “Seeing Other People”, *Philosophy and Phenomenological Research* 81, 3, 731–748.

On the Real Possibility of O'Brien: Historicized Transcendental Conditions and the Contingency of Sanity

SAMI PIHLSTRÖM

When invited to honor Valterri Arstila by a contribution to this Festschrift, it took a while for me to decide the topic of my paper. Although I first briefly met Valterri (I believe) in Turku in the late 1990s, I have never closely collaborated with him, and clearly our philosophical profiles are quite different. However, in a broad sense, *time* is, I believe, a theme that connects our diverging undertakings.¹

Trying to briefly say something about temporality – or, more broadly, about our historical contingency – in this essay also enables me to recall a somewhat unusual occasion when our paths crossed: Valterri was one of the participants selected through rigorous peer review for the interdisciplinary “Intercontinental Academia” program organized by a global network called University-Based Institutes for Advanced Study (UBIAS) in 2015-2016. The first round of the Intercontinental Academia, hosted by institutes for advanced study in Sao Paulo, Brazil and Nagoya, Japan, organized events and meetings on the topic of time, with Valterri as one of the active members of the group. I have no idea whether the UBIAS experience made any difference in his academic career, but I suppose the project of bringing together people (literally) from all continents and from a wide variety of disciplines – including philosophy, neuroscience, psychology, physics, biology, the social sciences, and the humanities – to discuss time and temporality might have been at least slightly interesting in its own right.²

¹ Arstila’s research focuses, e.g., on philosophical issues of time perception; see, among many other representative works, Mölder, Arstila, and Ohrstrom 2016.

² As the (then) Director of the Helsinki Collegium for Advanced Studies (a UBIAS founding member institute), I was (until 2015) a member of the UBIAS steering committee and of the board of the first round of its Intercontinental Academia. While I couldn’t, due to overlapping duties,

My brief remarks on our inescapable temporality and historicity in this paper remain, however, purely philosophical. My main issue is the revisability and contingency of even the most fundamental conditions of our thinking.

1. Transcendental conditions and their historicity

Even though Ludwig Wittgenstein noted, in his diary from the mid-1930s published as *Denkbewegungen*, that one should not fear madness, which he saw as the highest judge on whether one's life has been "right" or "wrong",³ most of us presumably do fear madness (if, that is, they think about its possibility at all). Throughout the history of our discipline, philosophers have frequently focused on the constitutive conditions of coherent thought – or what we might even, slightly provocatively, call the conditions for the possibility of sanity. In particular, the transcendental tradition extending from Immanuel Kant to twentieth-century and contemporary transcendental philosophers – including thinkers as different as analytic epistemologists writing on the structure of transcendental arguments and interpreters of, say, Husserlian phenomenology – has investigated the ways in which our transitory and contingent human projects and capabilities (including cognitive experience, meaningful language-use, scientific inquiry, and ethical debate) are dependent on constitutive conditions that make them possible, as the kind of activities they are, for beings like us.⁴

There is, or at least seems to be, something permanent and unchanging in such transcendental conditions, at least in the way they constrain and constitute

attend the UBIAS event in Brazil in 2015 (but only saw Valtteri and the other participants on the screen when I gave an online lecture – an early occasion of that kind), I did attend the UBIAS meeting in Nagoya in March 2016, giving a talk on "Time, Guilt, Irrevocability, and Forgiveness". (A version of that paper found its way to a chapter in Kivistö and Pihlström 2016.)

³ This is what he wrote: "Du sollst so leben, dass Du vor dem Wahnsinn bestehen kannst, wenn er kommt. *Und den Wahnsinn sollst du nicht fliehen.* Es ist ein Glück, wenn er nicht da ist, aber fliehen sollst Du ihn nicht, so glaube ich mir sagen zu müssen. Denn er ist der strengste Richter (das strengste Gericht) darüber ob mein Leben recht oder unrecht ist; er ist fürchterlich, aber Du sollst ihn dennoch nicht fliehen. Denn Du weisst ja doch nicht, wie Du ihm entkommen kannst, & während Du vor ihm fliehst, benimmst Du Dich ja unwürdig." (Wittgenstein 1997, 185-186; entry on July 20, 1937; original emphasis.)

⁴ It is impossible to even sketch here what transcendental philosophy and the "transcendental tradition" (should be taken to) mean, but at a very general level we may say that transcendental philosophers investigate the necessary conditions for the possibility of something whose actuality we (can only) assume. For discussions of transcendental arguments in epistemology and for transcendental reflections on the self inspired by phenomenology, see Stern 1999 and Carr 1999 (respectively).

the phenomena whose conditions they are. Kant himself maintained that the pure concepts of the understanding (viz., his twelve famous categories) are fixed and stable, necessarily structuring our experience and, indeed, the objects of any cognitive experience possible for beings like us. Moreover, while one of the transcendental conditions for the possibility of cognitive experience in Kant's system is time (along with space), the spatio-temporal character of any possible human experience and its objects is not itself subject to change. Lacking what Kant called "intellectual intuition" (that is, possessing only the capacity of sensible intuition), we cannot suddenly start to experience things non-temporally (or non-spatially); our imposing a spatio-temporal framework on our experience is something necessary and immutable.⁵

However, moving beyond Kant's own framework, it seems that there is a sense in which what it is, for us, to be "beings like us" – with the particular sensible and conceptual capacities we have – *is* (in principle) subject to temporality, contingency, and historical change. Very simply, *we* (whoever we are) *could* be different, and we could change across time, even if we contingently do not (in this fundamental sense). Thus, we should acknowledge and explore the historical mutability and revisability of what is, in the Kantian tradition, considered ahistorical and immutable, viz., even those transcendental conditions that we may regard as constitutive of something being not just experienceable but even "thinkable" for us at all. Divisions we draw between the thinkable and the unthinkable establish contexts within which (only) it is possible for us to coherently think at all, but such distinctions themselves could change as our lives change.

As soon as we realize that this is the case (insofar as it makes sense to speak of something's "being the case" in a matter such as this), we may feel our sanity to be threatened. That is, when acknowledging that the transcendental conditions that we find necessary for the possibility of our engagement in the kind of practices we (actually) do engage in, including the practices of inquiry and even thinking in general, are themselves (at a meta-level) subject to possible historical change, we may find ourselves staring into a bottomless pit. Such constitutive conditions, after all, protect our sanity. If we did not, in Kantian terms, organize our experience in terms of the necessary epistemic conditions of our cognitive activities (viz., the categories and the forms of intuition), we would live in what Kenneth R. Westphal

⁵ As this is not a study on Kant, I won't supply detailed references to the familiar sections of the *Transcendental Aesthetic* of the *First Critique* here.

(2004) aptly labels a “transcendental chaos”, being unable to experience any coherent objective reality. More generally, the ways we draw, or recognize, distinctions between what we find “thinkable” and “unthinkable” are what enables our meaningful participation in practices such as argumentation, discussion, and inquiry – not only in scientific but also, importantly, ethical contexts.⁶

In earlier work spanning across more than two decades, I have argued for the historicization (and thereby naturalization, in a broad and non-reductive sense) of transcendental arguments and transcendental conditions from a pragmatist point of view, suggesting that our apparently inescapable commitment to the constitutive conditions of our practices is itself in a sense contingent and thus in principle revisable, as we *could* be different from what we are, and finding support for this view in the pragmatist tradition (as well as the outskirts of the transcendental tradition itself, so to speak).⁷ The transcendently necessary, in short, depends on our purposive practices and is therefore pragmatically reinterpretable and thus contingent at a meta-level. This also entails that our conception of history and historicity as constitutive features of the human condition is itself historically relative, and thus potentially changing.⁸ As the classical pragmatists – perhaps most prominently William James and John Dewey, neither of whom can be regarded as a transcendental philosopher in any ordinary sense – recognized better than most other philosophers, our metaphysical security in the world we live in is inevitably fragile; ours is a world that is not only naturally but (we may say) transcendently precarious. There is no guarantee that we will be able to maintain our sanity in a world in which change is permanently possible, even pertaining to the constitutive conditions for the possibility of coherent experience and thought. Our sanity itself is fragile. The conceptual framework(s)

⁶ My use of the concept of the (un)thinkable draws from Gaita 2004 [1991], especially chapter 17; he speaks of the unthinkable in an ethical context, referring to views we cannot so much as consider, in contrast to those we reject as wrong. Unlike Gaita, I believe we should construe this concept (quasi-)transcendentally, as something we need in characterizing what we may call the “transcendental ‘us’”, or the conditions of our “mindedness”, borrowing useful terminology from Lear 1998, chapters 11-12. (This paper will not provide any detailed account of either Gaita’s or Lear’s views, though.)

⁷ Regarding the latter, I mean, of course, Wittgenstein and Wittgensteinian philosophy, in particular. See Lear 1998 for a highly illuminating discussion of the relation between transcendental and anthropological stances in (late) Wittgenstein.

⁸ I develop these ideas in, e.g., Pihlström 2003 (see chapter 6 for remarks on the historicity of our conception of time and temporality itself), and I return to pragmatist ideas of the historical revisability of a priori conditions in, e.g., Pihlström 2023.

in terms of which we are able to make sense of our lives seems to be fundamentally insecure, subject to revision.

I am, of course, moving rather swiftly and painting with a broad brush here. We will next discuss the general phenomenon we have now identified by focusing on a specific example – a philosophically contested concept constitutive of our ability to think if there ever was one. One of the concepts that is, arguably, vitally needed for our being able to continue sane and coherent thinking is the concept of *truth* – a concept that the pragmatists (especially James) have notoriously been taken to have distorted by defending a view that has become known as the “pragmatist theory of truth”. True, our conception of how exactly truth ought to be defined may historically change, but very few philosophers would maintain that we can continue to genuinely engage in our practices of inquiry and dialogue, or even thinking, without *some* concept of truth at our disposal. We can, thus, speak of the transcendental, and thus constitutive, significance of the concept of truth – or, better, of our commitment to the concept of truth as something that constrains all our (at least discursive) practices and is thus an ineliminable characteristic of the human condition. Pragmatist philosophers, along with many others (e.g., relativists and postmodernists of various stripes) may be taken to be partly guilty of the phenomenon known as the “post-truth era”: we have seen how little truth matters for some of us in, say, politics and the social media, and what kind of dramatic consequences the disregard for truth may have in our societies.⁹ We may, again, find our sanity threatened in a culture manifesting utter disregard for truth.

2. O’Brien: dangerous and possible

Discussing the role played by the concept of truth in our lives, we should recall the infamous slogans (e.g., “war is peace”, “slavery is freedom”) flowing from the Ministry of Truth in George Orwell’s dystopic novel, *Nineteen Eighty-Four* (1949), which returned to the top of bestseller lists when Donald Trump was elected the US President (for the first and hopefully last time) in 2016. Orwell’s work can be regarded as one of the most philosophically important literary examinations of the ethical and political dimensions of our thinking about truth, and our responsibility – both intellectual and more practical – for maintaining the constitutive values of “us” as participants in practices committed to pursuing truth.

⁹ I explore the pragmatist understanding of truth, and of what it means to be committed to the ethical value of truth, in relation to the post-truth culture in Pihlström 2021.

It is in relation to Orwell's unforgettable figure, O'Brien, that we may reflect on the historical contingency of "us" – especially our transcendentally constitutive commitments, such as the concept of truth, without which we can hardly make sense of our lives at all. It is the very possibility of such sense-making that Orwell challenges.

A philosophically provocative interpretation of Orwell's novel has been proposed by none other than the leading figure of late pragmatism, Richard Rorty.¹⁰ He rejects realist readings of *Nineteen Eighty-Four* according to which Orwell defends an objective (or realistic, correspondence-theoretical) concept of truth. In Rorty's reading of Orwell, emphasizing the significance of *freedom* in contrast to truth, the figure of O'Brien, the Party torturer, is given special emphasis. Orwell's significance lies in the novelty of his (re)description of what is possible: he shows us that "nothing in the nature of truth, or man, or history" will block the conceivable scenario that "the same developments which had made human equality technically possible might make endless slavery possible" (Rorty 1989, 175). According to Rorty, O'Brien, the "Party intellectual", is Orwell's most important invention in *Nineteen Eighty-Four*. Rorty emphasizes that Orwell offers *no answer* to O'Brien's thoroughgoing distortion of the concept of truth: "He does not view O'Brien as crazy, misguided, seduced by a mistaken theory, or blind to the moral facts. He simply views him as *dangerous* and as *possible*." (Ibid., 176.)

O'Brien is (hopefully) utterly unlike *us* – and therefore discussing this character is, in my terms (inspired by Lear 1998) instead of Rorty's, an inquiry into the transcendental "us". Our *not* endorsing O'Brien's way of viewing the world, including specifically his way of understanding the concept of truth as completely dependent on whatever happens to be the opinion of the Party, is arguably constitutive of our relation to the world in general and an element of any of our practices (including scientific inquiry and ethical deliberation) that may be thought to have any relation whatsoever to the way(s) the world is or can be taken by us to be. O'Brien imposes on Winston (the main protagonist of the novel) something that we can only find a thoroughly distorted, or insane, worldview affirming the absolute rule of the Party, a rule overwhelming enough to control even historical and mathematical truth. Winston's individual perspective on his suffering, and his life as a whole, is brutally subordinated to O'Brien's absolute Party standpoint, but Rorty's message is that there is no non-perspectival

¹⁰ See the chapter, "The Last Intellectual in Europe: Orwell on Cruelty", in Rorty 1989; cf. Kivistö and Pihlström 2016, chapter 5.

viewpoint for us to claim O'Brien to be wrong – or anything else than “dangerous” and “possible”.¹¹

Not only is Winston's belief in historical truths challenged: he initially believes that there were airplanes before the Party came to power, but O'Brien tells him that if the Party says that there have been airplanes only during the era of the Party, then there have been airplanes only during the era of the Party. More radically, even his most deeply entrenched mathematical beliefs, such as $2+2=4$, are shaken when he is, toward the end of the novel, tortured to acknowledge the “truth” that $2+2=5$. The Party fully controls even mathematical truth. Most importantly (though this remains mostly implicit in the novel), the Party even controls the truth about truth itself. If the Party maintains (as it does) that truth is dependent on what the Party says is true (that this is what truth *is*), thus endorsing what we might call a full-scale antirealism or constructivism about truth, then truth *is* dependent on what the Party says. There is nothing more to the concept of truth than this. Given that the Party can at any time change its opinion in arbitrary ways, we are not only not guaranteed that the truths we cannot fail to believe to be true will remain true, but we are not even guaranteed that what we mean by truth will remain unchanged. The “truth” about truth can arbitrarily change. *This* will threaten our sanity much more radically than any possible insecurity about what is true or false. We cannot remain confident that the concept of truth will continue to play its role, or any role, in our practices, and thereby our ability to participate in those (or any) practices will be under threat – a transcendental threat, as we might say.

This important aspect of Orwell's novel is in my view closely related to the ontology of history and temporality. In a recent book of mine (Pihlström 2022) developing a pragmatist philosophy of the humanities, specifically a pragmatically realist ontology of the reality that humanistic inquiry can be taken to investigate, I suggest, albeit only in passing, that while O'Brien, qua fictitious, of course does not exist in actual reality, the *real possibility* of his existence (emphasized by Rorty in his own way, albeit without the word “real”) is (in a rather un-Rortyan sense) an ontologically irreducible element of the historically structured world we live in. This real possibility is also arguably part of the reality we may investigate within

¹¹ See again, for an analysis of the novel in the context of pragmatist views on the problem of evil and suffering and specifically the tension between “theodocist” and “antitheodocist” approaches to that problem, Kivistö and Pihlström 2016, chapter 5.

the arts and the humanities,¹² and as such a potential “object” of humanistic inquiry, especially of research on history and literature. Such an ontology of real possibilities needed in the philosophy of history and historiography must be further pragmatically developed. This view is closely entangled with a resolutely antiteleological (e.g., pragmatist) philosophy of history recognizing the contingency of all our historical developments, including any apparently fixed and permanent conceptual conditions we seem to have no chance to abandon no matter how dramatically our form of life might change. Real possibilities make our world “open”, ruling out both causal determinism and metaphysical teleology (except as something based on *human* purposiveness characterizing our intentional actions, and thus liberated from any assumptions of grand non-human cosmic metaphysics). A kind of humanistic pragmatism, moreover, emerges here as an ethically oriented philosophy of history taking both suffering and its contingency fundamentally seriously: the suffering inflicted on other human beings – a paradigm literary depiction of which is O'Brien's torturing Winston – in the drama of history is not necessary but something that could be, or could and should have been, avoided.

However, crucially for our concerns here, this ontology of real possibilities must be constrained and limited by our transcendental sense of the thinkable and the unthinkable – or, indeed, the necessary conditions for the possibility of maintaining sanity. What is (humanly) ethically unthinkable structures what *we* may coherently view as (historically) possible, to the extent that there are cases in which we can only paradoxically say that the unthinkable became really possible, or really happened. An overwhelming historical catastrophe such as the Holocaust would be a case in point. As D.Z. Phillips (2005) reminds us, the fact that the Holocaust did take place does not entitle us to say (meaningfully) that the unthinkable was or is thinkable, after all, but rather that the unthinkable did happen or that some people actually committed an unthinkable crime. Reality – or what can be taken to be reality for us – ontologically depends on the kind of pragmatic-cum-transcendental boundaries of our form of life signaled by our distinguishing

¹² It is, more specifically, a Peircean “real general”. See Peirce's both early and late writings, spanning several decades from the early 1870s to the 1910s, on the reality of “generals” – the view he labeled “extreme scholastic realism” – in the essays collected in Peirce 1992-98. For Peirce, “real generals” include habits, dispositions, and modalities that are irreducible to existing particulars. Of course, I cannot here provide an independent argument for the reality of (Peircean) generals. I am only saying that insofar as we have the concept of a real general at our disposal, we can apply it to make sense of historical and interpretive ontology (cf. Pihlström 2022).

between the (ethically) thinkable and the unthinkable. The transcendental “we” (cf. Lear 1998) cannot coherently think beyond such boundaries, to think the unthinkable, and it would be an illusion to try to simply view “us”, or our form of life, from a reductively anthropological “sideways on” perspective; as Lear argues, there is a sense in which our “mindedness” admits no alternatives. O’Brien, then, in a sense, is unthinkable *for us*, yet real as a possibility we may see as a potential contingent development of what *we* are.¹³ Somewhat analogously, the Holocaust is unthinkable, yet (unlike O’Brien or the Orwellian world) shockingly real as a horrifying historical actuality.

In a simple sense, this just means that *we could be different*. Our “thinkabilities” and “unthinkabilities” could be different from what they are, and they could change in the course of time, although saying this requires that we are here (now) to say it within one or another framework always already limited by the boundary between the thinkable and the unthinkable. Our contingency is both a blessing and a burden. In the contingent world we live in, there is no other option for us but to do whatever we can to keep the O’Briens out. Otherwise, we will lose not just truth but along with it our capacity of coherent thought (cf., e.g., Conant 2000), and thereby everything, ceasing to be “us”. Therefore O’Brien’s (or the Party’s) “theory” of truth is necessarily ruled out by us and for us, (*pace* Rorty) for transcendental reasons. But we may – though it is unclear whether we are legitimately entitled to – come to think that its possibility might not (e.g., in a dystopian future) be ruled out in this way.

Someone could come to really hold O’Brien’s theory of truth only by becoming – in Gaita’s (2004 [1991], chapter 17) apt terminology – a crank, a lunatic, or wicked.¹⁴ Reinvoking the Holocaust example, we may compare Winston’s situation to someone who has to defend the view that the Holocaust

¹³ Comparing O’Brien to other famous fictional characters, such as non-human or quasi-human monsters depicted in Gothic fiction, I would be tempted to say (though I must admit I do not have any fully worked-out view on this) that O’Brien is “unthinkable” (though, paradoxically, “really possible”) in a way someone like Dracula isn’t. We may very well imagine – and find metaphysically albeit not biologically possible – an “undead” blood-sucking vampire. But it is quite another thing to try to imagine someone who seriously believes that there is no workable concept of truth at all (which is what O’Brien’s extreme antirealism about truth arguably comes down to). The concept of “believing” also breaks down here. Yet, again paradoxically, it is this unthinkability itself that is “real” in O’Brien’s character. His real possibility is what we should find extremely – transcendently – troubling and a fundamental challenge to how we are able to think at all.

¹⁴ Gaita does not discuss Orwell in this context, but I believe his conceptual machinery is appropriate for my purposes here.

actually took place against a Holocaust denialist claiming that it never did (at least not in the way in which the generally accepted historical account says it did). Gaita admits that he would not know where to begin such an argument with a Holocaust denialist. There simply is no proper place to being. The matter would not be about the reality of a particular historical event; it would, at a much more fundamental or constitutive (“transcendental”) level, be about the canons of rational thought as such and thus, again, the concept of truth. Sound argument, as Gaita repeatedly reminds us, depends on sanity, and Holocaust denialism rejects what we find constitutive conditions of sanity. So does, at a meta-level, truth denialism *à la* O'Brien. When Rorty reminds us that O'Brien is, horrendously, possible, that means that our meta-level commitment to sanity presupposing the concept of truth is fragile. The permanent possibility of insanity is an ineliminable characteristic of our sense of our sanity.

In Rortyan terms, our key ethical (and political) challenge is to invite others, as many as possible, to join the group of “us” that we Western liberals are – by using persuasion instead of force, as Rorty often puts it – and thus to lead them to view the world “with our eyes”, or as we do, also regarding the concept of truth in its non-Orwellian sense. Is this an ethical task, or is it a task of changing those (imagined or real) others’ ways of viewing ethics, or their understanding what ethics is all about (or, re-employing the visual metaphor, to change their “eyes” and not merely what they see with them, or how they see it)? In an important sense, our transcendental distinction between thinkability and unthinkability is ineliminable from our relations to others and otherness in general, and here O'Brien presents us with an extreme form of otherness.¹⁵ In Gaita’s (2004 [1991]) terms, what is at issue is our understanding of the very “meaning of the ethical”, and the concept of truth is crucial in this task.

I believe philosophers like Gaita are right in arguing that *some* unthinkabilities must limit “us” and our form of life and thought in order for us to be able to engage in serious (including ethical) thinking and discussion at all. This seems unchanging but is, inescapably, historically precarious. The issue of truth and post-truth highlights, furthermore, the need to draw and redraw the boundary between the “merely” ethical (at a first-order level, so to speak) and the transcendently constitutive that frames the ethical. The former is something that

¹⁵ Incidentally, we may note here that this transcendental reflection, as different as it is from Rorty’s resolutely non-transcendental pragmatism, fully agrees with Rorty (e.g., 2021) regarding the need to defend liberal anti-authoritarian liberalism. There are both transcendental and non-transcendental (radically pragmatist and naturalized) ways of doing so.

can only be found within a framework established by the latter, but the latter sometimes changes (or needs to be revised), for better and worse, and whenever it does, ethical considerations cannot be dismissed from the assessment of such changes even though their very possibility (circularly) depends on the framework they may be used to challenge and revise. A serious inquiry into “us” should recognize the fact (if it makes sense to call it a “fact”) that we *could* ourselves fail to be “minded” in the way we must think one has to be minded in order to be minded at all.

O’Brien’s troubling figure leads us to repeatedly ask what it means to be “inside” and “outside” “us”, viz., what it is to be included in or excluded from a human practice or a form of life. A totalitarian way of viewing the world, such as O’Brien’s, excludes all rival perspectives (as unthinkable) and thus radically, and wickedly, suppresses human individuality. As Hannah Arendt (1951) famously maintained, totalitarianism generally seeks to make human individuals superfluous. It is, in Arendt’s terms, our human “natality”, our being distinct and unique as individuals capable of novelty, that can be claimed to make all human beings, without exception, infinitely valuable. There are different ways of failing to recognize this shared humanity. Our not fully realizing that O’Brien is possible – a failure to perceive a threat to our human condition – is one form of such non-recognition. Our drawing, or recognizing,¹⁶ the boundaries of the thinkable should structure our intellectual, ethical, and political community – us – so that we continue to welcome human diversity and will keep O’Brien out.

Nothing I have said about the ethical significance of the boundary between the thinkable and the unthinkable should lead us to think that we would always easily be able to stay within the thinkable (even though there is a sense in which we can, of course, never leave the area of the thinkable). Recognizing the deep historical contingency of “us”, we ought to recognize the possibility – not independent of the real possibility of O’Brien – that we might ourselves in some (if not actual then at least really possible) situations think, or even do, the unthinkable.

¹⁶ I have repeatedly spoken of our “drawing” such boundaries, though it might be more appropriate to speak of our “recognizing” them, albeit always from within an area of discussion already defined by there being some such boundaries. In any case, the metaphor of “drawing” must not lead us to believe that we could somehow stand outside such a boundary-drawing (or boundary-recognizing) activity. This would be to harbor the illusion of a God’s-Eye View. Whether we draw or recognize the boundaries between the thinkable and the unthinkable, we do so from within a form of life depending on those boundaries – a form of life that is itself transformed whenever any redrawing (or re-recognizing) takes place.

3. Historicism, holism, and reciprocal interdependence

Regarding truth in this context, a related – equally fundamental – question is whether claims about what is thinkable or unthinkable, transcendently construed, are themselves true or false (and thus truth-apt). *Can* they be? I would be tempted to say: only in a pragmatic sense. It is “satisfactory” for us, from within our practices, to draw transcendently constitutive boundaries in certain ways; we need to do so, and there is a sense in which *we* can do nothing else – or, again, if we do, then “we” will change. It seems that the correspondence theory of truth cannot be directly applied here. But it works well *within* a pragmatically (and transcendently) bounded area of discourse defined by the limits of the thinkable; hence, what we arrive at here, indirectly, is an argument for embedding the age-old correspondence theory within a broader, more inclusive pragmatist account (cf. Pihlström 2021, chapter 1).

James Conant, one of Rorty’s sharpest but most constructive critics, defends what he calls an “ordinary” notion of truth in his detailed response to Rorty’s reading of Orwell. Conant (2000) views Orwell’s novel as an investigation of the conditions of thought in general (and here, though not everywhere else, I am following him), regretting Rorty’s slide into the other extreme on the scale between metaphysical realism (or “Realism”, with a capital “R”) and a sheer rejection of the concept of truth (replacing it with the concept of freedom).¹⁷ The reduction of truth to something completely different – something that *we* cannot recognize as truth – is as dangerous, eventually, as O’Brien (or views such as Holocaust denialism). They replace the pursuit of truth and the virtue of sincerity that goes along with it by something radically different.¹⁸ As Conant says, Orwell’s novel is primarily “about the possibility of a state of affairs in which the concept of objective truth has faded as far out of someone’s world as it conceivably can” (ibid., 297).

Such a (quasi-transcendental) “fading out” is a real historical possibility of our contingent human condition. Thus, more generally, one of the Rortyan (and

¹⁷ One of Rorty’s famous, or notorious, one-liners can be found in the Orwell essay: “If you take care of freedom, truth can take care of itself” (Rorty 1989, 176). For a pragmatist critique, see again Pihlström 2021.

¹⁸ This is not to say that there is nothing valuable in Rorty’s critique of traditional concepts of truth. In his posthumously published book, Rorty (2021) views our realist concern with Truth (or Reality) as a remnant of the authoritarian concern with being answerable to a superhuman God, defending pragmatism as the final step in the process of Enlightenment liberating us from such postulations. Rorty’s radical pragmatism is emancipatory in the sense of advancing such liberation.

Orwellian) ideas that must be emphasized here is our unavoidable historicity. The boundaries of the thinkable do change, changing us and our world. However, these changes only take place from within the thinkable; otherwise, they would be inconceivable to us as the kind of changes they are. We may draw a simile from Wittgenstein (1974 [1921], §6.43): when the limits of the world change, the world must “wax and wane as a whole”. We can extend or reconsider the thinkable only by moving within it, recognizing that our so doing may adjust or reconfigure the “hinges” (another Wittgensteinian metaphor) enabling such moves. Unavoidably living within the thinkable, we can only conceive of our world and lives as waxing and waning *in toto*. There is no way in which *we* could dispense with truth, but within our bounds of sense we may recognize the (hopefully distant) possibility that we could lose our grip of that concept.

We *might* (or so we can try to say) come to view truth differently, even insanely, presumably even in O’Brien’s mind-blowingly absurd way. After such changes, we would in a sense view *everything* (the world, ourselves) differently; this is what it means to say that we are dealing with a transcendental issue. Historicist rearticulations of the thinkable are in this sense *holistic*. However, discussing the possibility, desirability, or undesirability of such historical changes of our “mindedness” presupposes that even our conception of our historicity is subject to transformation. In the past we may have had a more stable conception of what is thinkable and unthinkable. Our understanding of how historical changes change us may itself change (again changing us) at the meta-level, and has done so in the course of history.

History, in its relevant meaning here, is human history. A key philosophical commitment of this entire discussion is, then, *humanism*. Even when historicized and critically self-reflective, humanism, maintaining that rejections of such key elements of the human world as objective truth are (for us) unthinkable, is a transcendental position. In this sense, it has no relevant alternatives: it is the way we are, inescapably, minded (cf. Pihlström 2021, 2023). It is, accordingly, also an ineliminable feature of the kind of pragmatism that functions as a background view (or a kind of “meta-theory”) of our inquiry into the historicity of transcendental conditions: philosophy deals with problems arising from our contingency and finitude, problems concerning our relation to the world as an arena of human possibilities, a temporal world in which we are condemned to freedom (in a quasi-Sartrean sense).

It is within this human condition that we can only – even when affirming our contingency and historicity – emphasize what we may call the reciprocal interdependence of truth and sanity, something that Orwell (and, in his own way, Rorty) challenges. Raimond Gaita is right in reminding us that the pursuit of truth – and all our practices that go along with it, including argument and debate – depends on sanity.¹⁹ But our sanity also, conversely, depends on our being able to employ the concept of truth. Gaita somewhat one-sidedly emphasizes one of the poles of this reciprocal interdependence, which itself holistically develops (“waxes and wanes”) through our history. I have throughout this paper spoken of truth as both a transcendental condition for the possibility of our engagement in practices we engage in (e.g., thinking, inquiry) and of the transcendental conditions for the possibility of our employing the concept of truth itself (or related concepts); this possibly confusing way of speaking is deliberate. Transcendental dependencies are not one-directional but two-directional.²⁰ Our form of life is thoroughly structured by our commitment to continuing to employ the concept of truth, and this is both an enabling, constitutive feature of our lives and enabled and constituted by our life remaining sane. If Orwell (or O'Brien) can teach us anything, the lesson is that both truth and our sanity – and their reciprocal interdependence – are fragile and need protection in our precarious human world.

University of Helsinki

References

- Arendt, Hannah (1951). *The Origins of Totalitarianism*. Chicago: University of Chicago Press.
- Carr, David (1999). *The Paradox of Subjectivity: The Self in the Transcendental Tradition*. Oxford: Oxford University Press.
- Conant, James (2000). “Freedom, Cruelty, and Truth: Rorty versus Orwell”. In Robert B. Brandom (ed.), *Rorty and His Critics*. Malden, MA and Oxford: Blackwell.
- Gaita, Raimond (2004 [1991]). *Good and Evil: An Absolute Conception*. Rev. ed. London and New York: Routledge.

¹⁹ Therefore, as suggested above, Holocaust deniers are not just wrong; they are not involved in an argumentative game in which we could judge them to be right or wrong at all. They are not pursuing the truth. Nor, of course, are truth deniers such as O'Brien. But therefore we cannot play the game of argumentation against them, either.

²⁰ The analogy is a kind of coherentism or, architecturally, a building supported not by a single unshakeable foundation but by the entire structure itself. This pragmatic reinterpretation of transcendental conditions is controversial, of course.

- Kivistö, Sari and Pihlström, Sami (2016). *Kantian Antitheodicy: Philosophical and Literary Varieties*. Basingstoke: Palgrave Macmillan.
- Lear, Jonathan (1998). *Open Minded: Working Out the Logic of the Soul*. Cambridge, MA and London: Harvard University Press.
- Mölder, Bruno, Arstila, Valtteri, and Ohrstrom, Peter (eds.) (2016). *Philosophy and Psychology of Time*. Cham: Springer.
- Orwell, George (1992 [1949]). *Nineteen Eighty-Four*. London: Penguin.
- Peirce, Charles S. (1992-98). *The Essential Peirce*, 2 vols. The Peirce Edition Project. Bloomington and Indianapolis: Indiana University Press.
- Phillips, D.Z. (2005). "The Holocaust and Language". In John K. Roth (ed.), *Genocide and Human Rights: A Philosophical Guide*. Basingstoke: Palgrave Macmillan.
- Pihlström, Sami (2003). *Naturalizing the Transcendental: A Pragmatic View*. Amherst, NY: Prometheus/Humanity.
- Pihlström, Sami (2021). *Pragmatist Truth in the Post-Truth Age: Sincerity, Normativity, and Humanism*. Cambridge: Cambridge University Press.
- Pihlström, Sami (2022). *Toward a Pragmatist Philosophy of the Humanities*. Albany: SUNY Press.
- Pihlström, Sami (2023). *Realism, Value, and Transcendental Arguments between Neopragmatism and Analytic Philosophy*. Cham: Springer.
- Rorty, Richard (1989). *Contingency, Irony and Solidarity*. Cambridge: Cambridge University Press.
- Rorty, Richard (2021). *Pragmatism as Anti-authoritarianism*. Ed. Eduardo Mendieta. Cambridge, MA and London: The Belknap Press of Harvard University Press.
- Stern, Robert (ed.) (1999). *Transcendental Arguments: Problems and Prospects*. Oxford: Clarendon Press.
- Westphal, Kenneth R. (2004). *Kant's Transcendental Proof of Realism*. Cambridge: Cambridge University Press.
- Wittgenstein, Ludwig (1974 [1921]). *Tractatus Logico-Philosophicus*. Trans. Brian F. McGuinness and David F. Pears. London: Routledge and Kegan Paul.
- Wittgenstein, Ludwig (1997). *Denkbewegungen: Tagebücher 1930-1932, 1936-1937*. Ed. Ilse Somavilla. Frankfurt a.M.: Fischer.

III

OTHER PHILOSOPHICAL REFLECTIONS

Quantum Properties as Potentialities in Bohm's 1951 Book *Quantum Theory*¹

ALEXANDER D. CARRUTH & PAAVO PYLKKÄNEN²

1. Introduction

A hundred years has soon passed since Werner Heisenberg in 1925 formulated matrix mechanics, the first complete formulation of quantum mechanics. This was followed by Erwin Schrödinger who in 1926 developed a mathematically equivalent wave mechanics which describes quantum systems using wave functions that evolve according to the Schrödinger equation. However, regardless of extensive discussions about the interpretation of quantum theory over the years that have passed, there is still no consensus among philosophers of physics what (if anything) quantum theory is telling us about the fundamental nature of physical reality. In other words, the metaphysical significance of quantum theory is unclear.

One factor (but not the only factor) here is that the philosophical climate in the 1920s when the first interpretations of quantum theory were developed was anti-metaphysical due to the influence of logical empiricism. Thus Niels Bohr said that the wave function should not be seen as a description of quantum objects, but rather as a part of a mathematical algorithm with the help of which we can predict the probabilities for the results of measurements that we observe in the classically describable parts of our measuring apparatus (i.e. as indicated by a macroscopic “pointer”). Because of Heisenberg's uncertainty principle we cannot observe quantum objects moving (we cannot measure their position and momentum accurately simultaneously), and Bohr thought (in the spirit of the verificationist theory of meaning of the positivists) that it is meaningless to talk

¹ This paper has undergone peer review.

² About the division of labour between the authors: The authors contributed equally. AC is more responsible for the metaphysical parts (especially section 3) while PP is more responsible for the parts discussing Bohm's interpretation of QM.

about momentum when measuring position and vice versa. More generally Bohr thought that it is meaningless to try to provide a model or description of individual quantum objects. Many other physicists who adopted Bohr's "Copenhagen interpretation" had a similar anti-metaphysical attitude toward attempts to provide realist models of quantum objects. There were, however, significant differences between physicists often classified as supporters of what was known as the "usual", "standard", "text-book" interpretation or the Copenhagen view (see e.g. Howard 2004; see also Bohm and Hiley 1993, ch 2).

2. Bohm's metaphysics in his 1951 text-book *Quantum Theory*

One physicist who initially accepted the usual interpretation, but also had a desire to provide a more physical description of quantum reality was David Bohm. He is of course well known for proposing in 1952 a first consistent alternative interpretation of quantum theory in terms of non-local hidden variables (he independently re-discovered and completed the pilot wave theory de Broglie had presented in 1927; see Bohm 1952, Bohm and Hiley 1993). But it is less known that already in his 1951 text-book *Quantum theory*, when he was still working in the framework of the usual interpretation, Bohm was seeking a physical interpretation and tried to express quantum theory in terms of qualitative and imaginative concepts, in contrast to the very mathematical form in which quantum theory was originally developed. Here he clearly went beyond the anti-metaphysical, empiricist approach advocated by many in the Copenhagen camp.

In this short article we want to focus on one aspect of Bohm's 1951 interpretation, namely the way he thinks about the intrinsic nature of quantum objects, and the way he understands the quantum properties of matter. Let us first consider the issue of intrinsic nature. In classical physics it is customary to think that something has either a particle or a wave nature. Thus Newton thought that light consists of particles, while Young's interference experiments in 1801 established that light has a wave nature. Our ideas about the nature of light changed, but all along it was assumed that light had *some* intrinsic nature. According to Bohm quantum theory changed this radically. We have to give up the assumption that a quantum object such as an electron has an intrinsic nature as a particle or as a wave. Instead, quantum objects "...can act either like waves or like particles, depending on how they are treated by the surrounding environment" (1951: iii).

How should we then understand quantum properties (such as wave-like or particle-like properties) if they are not inherent properties? Bohm suggests that we ought to consider these properties as "... incompletely defined potentialities, the development of which depends on the systems with which the object interacts, as well as on the object itself" (1951:132). "Incompletely defined" should here be read in an ontological sense. Bohm acknowledges that the wave function is the most complete possible description of the electron that can be obtained by referring to variables belonging to the electron alone. But he emphasises that this description is incapable of defining the general form (wave or particle) in which the electron will manifest itself. So rather than the wave function describing definite pre-existing forms (e.g. particle or wave) that are revealed in an experiment, it only describes potentialities. These potentialities are actualized (i.e. become definite) only via interaction with a suitable measuring apparatus. (See Bohm 1951: 133).

These potentialities are also *opposing* in the sense that actualizing a given potentiality (such as wave-like behaviour) takes place at the expense of another potentiality (we cannot actualize particle-like behaviour in a situation in which wave-like behaviour is actualized) (1951: 138).³

Such postulation of potentialities to quantum ontology is a metaphysical move, and very different from the way in which Niels Bohr, for example, would characterise quantum reality (indeed, as we pointed out above, Bohr thought it meaningless to talk about the nature of quantum reality beyond what can be observed). Let us next see in more detail why Bohm characterises quantum properties as potentialities, and also explore further why he says they are incompletely defined.

He asks us to consider an electron with a broad wave-like packet, of definite momentum and, therefore, of a definite wavelength. (Remember the de Broglie relation $p = h/\lambda$ which implies that wave-length needs to be definite for momentum to be definite; here p is momentum, h is Planck's constant and λ is wave-length.) Bohm notes that such an electron is capable of demonstrating its wave-like properties when it interacts with a suitable measuring apparatus, such as a metal

³ A referee was wondering whether by *opposing* Bohm just means something like that when you cast a die and the result is 2 then, at that instant, the probability of the result being 4 becomes zero. The point in quantum mechanics is that in the experimental context in which you actualised "2" (position) you could not have (even in principle) actualised "4" (momentum), whereas with the dice both actualisations are in principle possible in the same situation. So this analogy does not fully capture the quantum situation.

crystal. (The crystal functions in the same way as the slits in the familiar two-slit experiment, and we get wave interference effects if the wave packet passes through it.) But, he continues, the same electron is potentially capable of developing into something more like a particle when it interacts with a position-measuring device. When this happens its wave-like aspects become correspondingly less important. And we can keep on doing this, using again an apparatus that enables the electron to demonstrate wave-properties, and then using again a position-measuring device, and so on.

The bottom line is that the electron is capable of undergoing *continual transformation* from wavelike to particle-like aspect, and vice versa. It is the kind of apparatus with which the electron interacts which determines which of these potential aspects prevails. Importantly, Bohm notes that this also applies in situations outside experiments – for example with crystals at the bottom of the sea. While presentations of the usual interpretation of quantum theory usually talked about “observables” (as opposed to “beables”), and emphasized the need for a classically describable measuring apparatus, presupposing a human observer, Bohm sees the actualization of quantum potentialities as something that can take place in nature, independently of measuring apparatuses or human observers.

The quantum properties differ from the properties of classical physics not only because they are potentialities, but also because of the special role that probability has at the quantum level: “...these potentialities refer to developments, the precise outcome of which is not related completely deterministically to the state of the electron before it interacts with the apparatus” (1951:132). To illustrate this Bohm describes a position measurement. We start again with a broad wave-like packet and let it interact with a position-measuring device. After the interaction the wave function is broken up into independent packets. But, he says, the electron is found in only one of these packets. This means that from the wave function we can calculate only the probability that a given packet is the correct one. This, of course, is the familiar probabilistic aspect of quantum mechanics.

Bohm now makes a very interesting observation. He points out that the “...interpretation of the properties of the electron as incompletely defined potentialities finds its mathematical reflection in the fact that *the wave function does not completely determine its own interpretation*” (1951:133, italics added).

Let us assume that we are in a situation where the electron has not yet interacted with a measuring apparatus. In this situation the wave function defines two important kinds of probability, namely the probability of a given position and

the probability of a given momentum. Now, Bohm points out that “...*the wave function by itself does not tell us which of these two mutually incompatible probability functions is the appropriate one*” (1951: 133, italics added). This question can be answered only when we design the experimental arrangement and in this way decide whether the electron interacts with a position measuring device or with a momentum measuring device.

Here we can see the special role that the experimenter has in quantum mechanics, as opposed to classical physics. In the latter we are used to thinking of the ontology of the physical world as actualized particles and fields in space, and our task as experimenters is to reveal the pre-existing values of the various variables in which we are interested. In quantum mechanics, in the way Bohm sets it up in his 1951 book, quantum ontology consists of a realm of incompletely defined and opposing potentialities, which require interaction with an environment to actualize them. And which potentiality of a single quantum object is actualized depends on the nature of the environment (e.g. whether it is a position or a momentum measuring device). In quantum experiments it is the experimenter who decides what to measure (e.g. position or momentum), and in this sense the experimenter plays a role in deciding which potentiality will be probabilistically actualized. This is very different from classical physics, where the experimenter is trying to reveal pre-existing values (see also Stapp 2007). Of course, one may ask here whether the choice of the quantum experimenter is “free”, or whether it is determined by the laws of physics. And if the latter, which laws – classical, quantum, or some laws that have not yet been discovered?

3. An Ontology of Incomplete and Opposing Potentialities

Bohm aims to give a metaphysically robust interpretation of quantum mechanics which says something about the physical nature of quantum objects. Given this aim, one potentially interesting and instructive line of inquiry is to examine the extent to which Bohm’s account might (or might not) comport with available general ontological frameworks. Whilst of course not decisive, should it be possible to identify a general ontological framework that fits well with Bohm’s 1951 interpretation, it is reasonable to see this as a point in favour of *both* the interpretation of quantum mechanics under discussion and the relevant general ontological framework. Note also that because Bohm’s 1951 interpretation is situated within the framework of the usual interpretation of quantum theory, the

metaphysical implications are likely to extend beyond Bohm's specific ideas in the 1951 book and may apply to quantum theory in general. Overall, we envisage a metaphysics of science-based approach in which traditional metaphysical theorising and the natural sciences should be seen as mutually supporting and in dialogue with one another (see for instance Guay and Pradeau 2020).

Bohm's account is centred on four ontological claims about the features of quantum objects: first, that they are *potentialities*, in the sense that they exist without fully actualising or manifesting their whole nature; second, that the potentialities are *incompletely defined*, in the sense that the context in which they are situated is always a crucial and ineliminable factor on how they do in fact actualise or manifest in any specific, concrete instance, third, that the actualization of the potentialities is typically probabilistic; and fourth, that the potentialities can be to some extent *opposing*, in the sense that the actualisation or manifestation of a certain variable precisely can render impossible the actualisation or manifestation of a certain "complementary" variable, and vice versa (as discussed above in the case of position and momentum). In the rest of this section, we'll examine an ontological framework based around a particular conception of *powers* or *dispositions*—real potentialities—defended by, amongst others, Charlie Martin (2008), John Heil (2003) and Rögnvaldur Ingthorsson (2020), which we'll argue seems to fit quite well with Bohm's account of quantum potentialities.

We do not claim, however, that this is the *only* available framework, although we will note briefly that certain neo-Humean approaches (arguably the dominant general ontological framework in recent metaphysics) seem to struggle to accommodate something like Bohm's ontological proposals (with broadly neo-Aristotelian approaches looking more promising—we are, of course, not the only ones to point out the apparent affinity between Aristotelian frameworks and quantum potentialities: Heisenberg noted the connection (e.g. 1974, pp.16-17), see also Riezlner (1940)). Why should neo-Humean approaches struggle? Briefly, the guiding picture of nature in the neo-Humean framework is what is often referred to as the 'Humean mosaic', a distribution of point-like purely intrinsic qualities over a four-dimensional spatio-temporal manifold. Amongst the elements of the mosaic, there are no necessary connections. Such a picture, at least *prima facie*, does not sit well either with the *incompleteness* or the *opposing-ness* of Bohm's quantum potentialities, as (i) *incompleteness* implies that various distinct elements (the quantum object plus various elements of the environment) are necessary for any given concrete actualisation and (ii) if the actualisation or manifestation of a

certain kind of feature (e.g. position) of a quantum object precisely makes it impossible that some other feature (e.g. momentum) actualises or manifests, as the *opposing* nature of quantum potentialities requires, then again this involves the kind of necessity that a neo-Humean system typically abhors. We'll not dwell further on this point here, but move on to explore a more immediately promising approach.

Unlike with neo-Humeanism, neo-Aristotelian systems of various kinds embrace natural necessity. Neo-Aristotelian metaphysics is not a monolith, and there are various quite distinct guiding pictures of nature endorsed by different thinkers working within this broad framework—for instance, essence-centred accounts; hylomorphic accounts and powers-centred approaches (and these various elements can be combined, they are not mutually exclusive). Our focus will be powers-centred approaches, which seems a natural starting point given the centrality of the concept of potentialities in Bohm's 1951 theory.

Powers are features or properties which are essentially directed *to* something or other. That is, their nature involves being directed towards some manifestation or set of manifestations; and they bring about these manifestations in suitable circumstances (or, if some powers are stochastic—which may well be the case at the quantum level—will assign the same objective probabilities to their manifestations—henceforth we will omit qualifications of this sort, as it should usually be clear how the relevant claims can be modified to accommodate stochastic powers). As such, powers and potentialities go hand-in-hand—conceiving of properties as powers involves conceiving of them as having the potential to manifest or actualise something further. Canonical classical examples include fragility, which could be roughly characterised as 'the power to break when struck', or solubility, 'the power to dissolve when in contact with a suitable solvent' (these are used here merely as illustrations of the general idea—in the final accounting it may turn out that fragility or solubility are to be reduced to some more basic features). It is often suggested that we can roughly capture these features using conditional statements of the form:

P: x has the power to ϕ if it is the case that were x to be placed in suitable circumstances C, then x would ϕ

We don't offer **P** as an *analysis* of what it is to have a power, but conditional statements such as **P**, even though they fail to analyse what it is to have a power,

express key features of powers mentioned above: their directedness and their sensitivity to circumstance.

Inspired by something like **P**, what could reasonably be regarded as the *orthodox view* of powers in the contemporary debate holds that powers are single-track—each power is directed towards just one manifestation—and that they operate according to the *stimulus-manifestation* model—a power will only give rise to a manifestation when it is galvanised into action by some trigger or stimulus (e.g. Bird 2007). Whenever a single-track power manifests, and in whatever circumstances it does so, the manner in which it manifests is the same. According to the 'stimulus-manifestation' model, a power will only give rise to a manifestation when it is galvanised into action by some trigger or stimulus. For instance, in the case of the *fragility* of a vase, the stimulus might be 'being struck with a force greater than X', or in the case of the *solubility* of a sample of salt, 'being submerged in water'. Crucially, the manifestation is produced by the target power alone, although it will not be produced until the occurrence of the stimulus.

The orthodox view, however, does not seem particularly well suited as a general ontological framework for Bohm's account of quantum potentialities. The reasons for this are two-fold. First, whilst context has some role to play regarding what is actualised on a given occasion (by providing, or failing to provide, the stimulus conditions that trigger the manifestation or actualisation of the power), Bohm's notion of *incompleteness* seems to point towards a more reciprocal relationship between the quantum object and the environment, where it is mutual interaction that gives rise to a specific manifestation on a given occasion. Second, it is hard to see how or why powers understood on this model would exhibit *opposing-ness*—that is, it is hard to see why, if the nature of a power is exhausted by its directedness to a single manifestation combined with sensitivity to a specific stimulus, the actualisation of one sort of manifestation on an occasion would make impossible the actualisation of another sort. The claim here isn't that the orthodox view is strictly inconsistent with the *opposing* nature of quantum potentialities, but rather that some addition to the framework would be needed in order to account for it.

An *alternative view* holds that powers are *multi-track*—a single power is directed towards a (possibly very wide and diverse) range of qualitatively distinct manifestations—and that they operate according to the *mutual manifestation* model—in order for some manifestation to occur, there must always be at least two powers working together, and there is no sense of priority such that one power

(or some subset of the powers involved) could be considered the ‘operative’ or ‘active’ power(s), whilst the other(s) are merely a stimulus or trigger. Note that this is distinct from the view (often associated with Aristotle) that powers come in *active/passive* pairings, where the active power plays the determinative role, and the passive power is a mere liability to be subject to changes associated with the manifestation of the active power. This alternative view of powers has been defended by for instance Martin (2008), Heil (2003) and Ingthorsson (2021), although we will focus on Martin’s presentation here as he is particularly explicit concerning features of the view that are important in the present discussion.

Martin’s account (he uses the language of ‘dispositions’ rather than ‘powers’, but for the purposes of this discussion they can be treated as synonyms) emphasises three features which seem, *prima facie*, to fit well with the key features of Bohm’s theory which were mentioned at the start of this section. First, Martin is clear that “[a] particular disposition exists or it does not. You could say of any *unmanifesting* disposition that it straight-out exists, even if it is not, at that time or at any other time, manifesting any manifestation” (2008, pp. 1-2). Thus, it is clear that Martin’s powers are well-suited to be understood in terms of *potentialities* conceived as genuine, albeit, as we will see momentarily, incompletely defined features of their bearers (in the relevant case, of the quantum object). Second, the notion of reciprocity and mutuality are at the very core of Martin’s understanding of the operation of powers:

The reciprocities of these dispositionalities for their mutual manifestations are many, deep and complex. The important point is that for a specific mutual manifestation, any particular dispositional state is itself only one among many other dispositional states that together form reciprocal disposition partners for their particular mutual manifestation. (ibid. p.3)

And

One must take the *reciprocity* of a reciprocal readiness *partnerings* as their *mutual* manifestings seriously and not ask for the action of a single readiness factor only because there is no such action. Instead, actions are the *reciprocal* partnerings of a web or net of readiness. (ibid. p.22)

This conception of powers fits quite naturally with Bohm's notion of *incompleteness*, as described above as requiring that various distinct elements (the quantum object plus various elements of the environment) are necessary for any given concrete actualisation. For on such an account one could never conceive of the underlying potentiality of a quantum object alone as being responsible for the actualisation or manifestation of a determinate feature on a specific occasion—the context (experimental or natural) will always play an equally important and ineliminable role. And this is precisely what Martin's account holds about the nature of powers in general, that they are reciprocal to the core: what it *is to be* a particular power is in large part given by reciprocal relatedness to other powers. And, as the quote below mentions, Martin takes the readinesses, directednesses and selectivities of a power to run to infinity, so there is no sense in which a power can be thought to be 'completely defined' in isolation (this is an ontological point, not just a point concerning whether we could specify such a definition in language).⁴

Finally, Martin is explicit that the multiply-directed nature of powers is not limited to a range of 'positive' manifestation types, but also involves exclusion or suppression of alternatives:

A particular dispositional state with its particular set of disposition partners for a particular set of manifestations with *different* sets of disposition partners (that may be infinite) will *not* be for or may even be *prohibitive* against manifesting an infinite number of manifestations... Dispositionalities *must* outrun manifestations, for dispositionalities can be for *alternative* mutual manifestations whose alternative disposition partners mutually exclude the concurrence of those mutual manifestations. (ibid. pp.3-4)

Once more, we can see that a conception of powers along the lines suggested by Martin seems to fit rather neatly with the idea that some potentialities can be *opposing*, as Bohm's theory requires—it should be part of our basic understanding of powers that the relationships between them, the way in which they reciprocally and mutually manifest and their manifestations themselves are complex, sensitive to circumstance and can involve exclusion or selectivity *against* alternatives as

⁴ We are grateful to a referee for pushing us to be more explicit about how Martin's view fits with incompleteness and opposing-ness.

much as they involve directedness *towards* what it is they actualise on some given occasion (e.g. under some specific measurement conditions). Thus, Martin's understanding of powers as both directed towards some (reciprocal, mutual) manifestations and prohibitive of others is well placed to accommodate the Bohmian notion of opposing potentialities, for it allows quite naturally that the actualisation of one manifestation (e.g. the reciprocal and mutual manifestations involved in a momentum measurement, which involves both powers of the quantum object and the measuring apparatus) may also be *prohibitive* against some other manifestations (including those that would need to be actualised for a position measurement).

Although the above discussion has been rather brief, we hope to have shown that the central ontological commitment of Bohm's 1951 quantum theory—*incompletely defined, opposing potentialities*—can be situated within an independently motivated general ontological framework such as the account of powers defended by Martin and others. This may not be the only possible framework, but we have tried to show that there is a natural fit here, and have also given some reasons to think that various alternative views—such as neo-Humean approaches and the orthodox contemporary view of powers—will have a harder time accommodating Bohmian quantum potentialities.

4. Does this really work?

Let us now go back to consider Bohm's account of quantum potentialities further. While the approach may sound *prima facie* interesting and viable, it has some puzzling aspects. To see this, let us consider a statement which summarizes Bohm's 1951 quantum metaphysics: "We ... interpret momentum and position (and thus wave and particle aspects) as incompletely defined potentialities latent in the electron and brought out more fully only by interaction with a suitable measuring apparatus." (1951: 133) But here it is tempting to ask: doesn't the measuring apparatus itself consist of quantum objects whose properties are also "incompletely defined potentialities"? And if so, how can then sets of "incompletely defined potentialities" actualize "incompletely defined potentialities"? Bohm does not address this question explicitly in his book, but the final short chapter of his book, "Ch 23 Relationship between quantum and classical concepts" (pp. 624-628) provides a discussion with which we can evaluate how he might have replied.

Our common intuitions about the nature of the micro and the macro and their relation might well suggest that we can derive classical concepts as limiting cases of quantum concepts. After all, isn't it obvious that macroscopic, classical objects consist of microscopic, quantum objects? And isn't it customary to think that quantum theory is more fundamental, so that it could be the basis from which classical theory and classical concepts could be derived as a special limiting case? It is one of the fascinating features of the usual, text-book interpretation quantum theory that this is not at all obviously the case. Let us examine how this can be seen in Bohm's 1951 quantum ontology.

As we have seen, according to Bohm quantum properties are incompletely defined potentialities which can be actualized only in interaction with a classically describable system (a measuring apparatus is a special case – the system could be a naturally occurring crystal at the bottom of the sea).

The wave function can be used to calculate the probability that when a system interacts with a suitable measuring apparatus, it will develop a definite value of the variable that is being measured. But the measuring apparatus itself cannot be understood as a purely quantum object (i.e. consisting only of incompletely defined potentialities), for its last stages are always classically describable. This means that after the measurement interaction we will find the macroscopic pointer of the measuring apparatus in a well-defined position, pointing to a definite result. And it is only at the classical level that definite results for an experiment can be obtained. Bohm continues:

This means that without an appeal to a classical level, quantum theory would have no meaning. *We conclude then that quantum theory presupposes the classical level and the general correctness of classical concepts in describing this level; it does not deduce classical concepts as limiting cases of quantum concepts.* (1951: 625)

What about the correspondence principle? Does it not show that we can derive classical ontology as a limiting case of quantum ontology? No, says Bohm: "The correspondence principle is simply a consistency condition which requires that when the quantum theory plus its classical interpretation is carried to the limit of high quantum numbers, the simple classical theory will be obtained." (1951: 626).

Thus, the correspondence principle is not a direct derivation tool, but a condition ensuring consistency between quantum and classical theories.⁵

Bohm's discussion implies that quantum theory needs classical theory and its concepts as a foundational framework. This means that quantum theory is an incomplete description of physical reality: one cannot describe the large-scale behaviour of a system completely in terms of quantum concepts. Quantum concepts refer to incompletely defined potentialities, but in the classical level we find definite objects and events. These classical properties cannot be deduced from the quantum description in terms of the wave function, but they must nevertheless be consistent with the quantum description.

Does the world then divide into two independent realms, the classical level with classical properties and the quantum level with quantum properties? Of course not, for these levels depend on each other. We need well-defined classical events (e.g. a macroscopic pointer pointing at a specific value) to actualize quantum-mechanical potentialities. But reciprocally, we need a quantum theory of the component molecules of a macroscopic system to fully understand its large-scale behavior. Bohm summarizes the situation: "...large-scale and small-scale properties are both needed to describe complementary aspects of a more fundamental indivisible unit, namely, the system as a whole." (1951: 627)

All this suggests an interesting ontology for the world. The world can be divided to two interdependent levels (quantum and classical), which are very different in their nature and the tendencies that prevail in them. At the quantum level, Bohm says, there is a continual tendency for a system to cover the whole range of its potentialities, while at the classical level there is a continual tendency for things to become definite, i.e., for a specific potentiality to be realized at the expense of all other potentialities (1951: 627).

These two levels meet e.g. in the process of measurement, where the system settles down to a particular value of the measured variable and all other possibilities are discarded. This is connected to the idea of the collapse of the wave function. At the end of the process of measurement, there appears a definite result at the classical level (e.g. a macroscopic pointer is pointing to a specific value).

⁵ In a well-known quote, to which physicists like to refer even today, Landau and Lifshitz (1974: 6) write: "...quantum mechanics occupies a very unusual place among physical theories: it contains classical mechanics as a limiting case, yet at the same time it requires this limiting case for its own formulation". If Bohm is correct the first part of Landau and Lifshitz's statement is strictly speaking incorrect, but they are pointing to an idea similar to the one Bohm emphasizes.

This is reflected back into the quantum level in the sense that a narrowing down of potentialities at the classical level is accompanied by a similar narrowing of potentialities at the quantum level. But at the quantum level such narrowing of potentialities for one variable (e.g. position) means that there will be a widening of the potentialities for the complementary variable (e.g. momentum). This widening influences our expectations about what will be observed at the classical level if we now make a momentum measurement. Bohm summarizes: "...in the continual interplay between the quantum potentialities and their classical realizations, the system is subject to an endless series of transformations." (1951: 627).

It is important to distinguish between the kind of 'levels' that we have suggested Bohm's ontology involves, and the conception of levels familiar from discussions concerning the autonomy of the special sciences. In this latter debate, those who resist the idea that the entities described in chemistry, biology and so on can be reduced, ontologically speaking, to micro-physical entities, often appeal to a hierarchy of compositional levels of being. The claim then developed is that the entities occupying 'higher' levels being dependent on but nevertheless somehow autonomous with respect to, their lower-level realisers. This is clearly not the same sort of relationship envisaged as obtaining between the quantum level of incompletely defined and opposing potentialities and the classical level of determinate, manifest being.

Rather, the quantum and classical levels are mutually dependent and presuppose each other. Again here, we might find a natural fit with Martin's general ontological framework, for he rejects both the notion of *pure* potentiality on the grounds that to move from potency to act, some determinate, manifest being is required (a world of pure potentiality would be a world of 'promissory notes' which are never made good on (2008, p.63)) and the notion of *pure* manifest or actualised being, on the grounds that being that is 'in pure act', that is, being that always manifests everything of which it is capable, would actually amount to non-being (at least in the case of concrete, material phenomena; Martin allows that abstracta such as numbers, should they exist, might be in pure act (ibid. p.54)). Thus, for Martin, whilst powers involve a kind of order of being with a rich structure that "outstrips" the manifest; nevertheless they presuppose some determinate, manifest being (as Bohm's level of quantum potentiality presupposes the classical level).

5. Discussion

One might wonder whether it is relevant, in 2024, to go back to Bohm's 1951 discussion of quantum properties as potentialities. We saw that Bohm himself there acknowledged that quantum theory presupposes the framework of classical physics, which cannot be seen as a satisfactory view for anyone who assumes that quantum theory is a complete description of the non-relativistic domain. However, as we pointed out in the introduction, the fact is that there is still no commonly accepted interpretation of quantum theory. A key part of this is that there is no commonly accepted view of the relation between the quantum level and the classical level. This is why a view such as the many worlds interpretation is taken by many to be a genuine option in the current debates. There is no commonly accepted account of how quantum theory gives rise to a single macroscopic world, and thus the idea of "many worlds" remains an option.

Given this, if one wants to avoid the assumption of many worlds, it is perhaps not such a bad option to accept that currently the quantum theory presupposes the classical level. This is to acknowledge that quantum theory is incomplete, and that something new is needed if we want to derive a single, unique classical world from the quantum level. Bohm's later 1952 "pilot wave" theory provides one way to do this; but while the theory has many attractive features, it is not universally accepted.

So perhaps Bohm's 1951 idea of quantum properties as potentialities is still a live option which deserves more attention. For one thing it suggests an interesting link between physics and metaphysics, as we have briefly indicated in this paper. There are also other authors who have recently moved to a similar direction. For example, in their paper "Taking Heisenberg's *potentia* seriously" Ruth Kastner, Stuart Kauffman and Michael Epperson suggest that "...quantum theory is best understood as requiring an ontological dualism of *res extensa* and *res potentia*, ... understood as mutually implicative ontological extants that serve to explain the key conceptual challenges of quantum theory; in particular, nonlocality, entanglement, null measurements, and wave function collapse" (2018: 158). While they do not refer to Bohm's 1951 discussion but rather Heisenberg's 1955 and 1958 discussions of potentialities, they make similar suggestions as already Bohm did, while introducing new interesting insights.⁶

⁶ For other discussions of the role of potentialities in quantum theory see e.g. Fleming 1992; Kožnjak 2020; for a discussion in the context of theology, see Korpela 2022. Kistler 2018 provides

We also suggest that seeing quantum properties as potentialities provides a middle way between on the one hand very counterintuitive “actualist” views (such as Many Worlds (Wallace 2012) or Wave Function Realism (Albert 1996, Ney 2021)) and on the other hand radically epistemic and subjectivist neo-Bohrian views such as QBism (Fuchs, Mermin and Schack 2014). The idea that potentialities are a fundamental aspect of the quantum world also fits nicely together with the suggestions that both quantum gravity and quantum nonlocality suggest the need to consider a non-spatio-temporal order, from which space-time emerges (see Ismael 2020, Bohm 1980, ch 7).

University of Helsinki (AC & PP)
University of Skövde, Sweden (PP)

References

- Albert, D. Z. 1996. Elementary Quantum Metaphysics. In *Bohmian Mechanics and Quantum Theory: An Appraisal*. J.T. Cushing, A. Fine, and S. Goldstein, eds. Kluwer, Dordrecht, 277–284.
- Bohm, D. 1951. *Quantum Theory*. Prentice Hall, Englewood Cliffs. Dover edition 1989.
- Bohm, D. 1952. “A Suggested Interpretation of the Quantum Theory in Terms of ‘Hidden Variables’” I and II, *Phys. Rev.*, 85 (2), 166-179 and 180-193.
- Bohm, D. and Hiley, B.J. 1993. *The Undivided Universe: An Ontological Interpretation of Quantum Theory*. Routledge, London.
- Bohm, D. 1980. *Wholeness and the Implicate Order*. Routledge, London.
- Fleming, C.N. 1992. “The Actualization of Potentialities in Contemporary Quantum Theory” *The Journal of Speculative Philosophy* 6, (4), 259-276
- Fuchs, C.A., Mermin, D. and Schack, R. 2014. An introduction to QBism with an application to the locality of quantum mechanics. *American Journal of Physics* 82 (8), 749-754.
- Guay, A. and Pradeau, T. 2020. “Right out of the box: how to situate metaphysics of science in relation to other metaphysical approaches”, *Synthese*, 197, 1847–1866.
- Heil, J. 2003. *From an Ontological Point of View*. Oxford University Press, Oxford.
- Heisenberg, W. 1974. *Across the Frontiers*, Harper and Row.
- Howard, D. 2004. “Who Invented the ‘Copenhagen Interpretation’: A Study in Mythology”, *Philosophy of Science* 71, 669-682.
- Inghthorsson, R. 2021. *A Powerful Particulars View of Casuation*. Routledge, New York.
- Ismael, J. 2020. “What Entanglement Might Be Telling Us: Space, Quantum Mechanics, and Bohm’s Fish Tank”, in D. Glick, G. Darby, and A. Marmodoro (eds) *The*

a useful discussion of both the notion of potential and the notion of potentiality in physics - it is important to understand the difference between these notions.

- Foundation of Reality: Fundamentality, Space, and Time*. Oxford University Press, Oxford.
- Kastner, R., Kauffman, S. and Epperson, M. 2018. "Taking Heisenberg's Potential Seriously", *International Journal of Quantum Foundations* 4 (2), 158-172.
- Kistler, M. 2018. "Potentiality in Physics", in *Handbook of Potentiality*, Springer Netherlands, 353-374.
- Korpela, Sampsa 2022. *God, Time and the Concept of Potentiality in Quantum Physics*. Doctoral dissertation, University of Helsinki. Available at: <https://helda.helsinki.fi/handle/10138/346531>
- Kožnjak, B. 2020. "Aristotle and Quantum Mechanics: Potentiality and Actuality, Spontaneous Events and Final Causes", *Journal for General Philosophy of Science* 51, 459-480
- Landau, L. and Lifshitz, E. 1974. *Quantum mechanics*. Pergamon press, Oxford.
- Martin, C. B. 2008. *The Mind in Nature*. Oxford University Press, Oxford.
- Ney, A. 2021. *The World in the Wave Function: A Metaphysics for Quantum Physics*. Oxford University Press, Oxford.
- Riezler, K. 1940. *Physics and Reality*, Yale University Press.
- Stapp, H. 2007. *Mindful Universe: Quantum Mechanics and the Participating Observer*.: Springer Frontiers Collection, Berlin and New York.
- Wallace, D. 2012. *The Emergent Multiverse: Quantum Theory according to the Everett Interpretation*. Oxford University Press, Oxford.

Some Thoughts on Descartes on Immortality

OLLI KOISTINEN

1. Theory

The title of the first edition (1641) of Descartes *Meditations* is:

*Meditationes de prima philosophia, in qua Dei existentia et animae immortalitas demonstratur. [Meditations on first philosophy in which the existence of God and the **immortality** of the mind are demonstrated.]*

But the title of the second edition (1642) is:

*Meditationes de prima philosophia In qua Dei existentia, & animæ humanæ à corpore distinctio, demonstratur. [Meditations on first philosophy in which the existence of God and the **distinction** between the human mind and body is demonstrated.]*

Thus, the first edition tells that in the *Meditations* the immortality of the soul is demonstrated, whereas the title of the second edition no more contains the proof of immortality but instead the proof of the mind body distinction.

The change of the title makes sense because Descartes touches the issue of immortality only in the Synopsis, whereas in the subsequent six meditations immortality is not discussed directly at all. And even the Synopsis does not try give a conclusive argument for immortality. The Synopsis seems to base immortality on the distinction between physical reality and thinking reality. The point is that whereas bodies are configurations of other bodies, a body B may cease to exist when the relations between the bodies composing B radically change.

However, the soul or I cannot be thought of as having a compositional structure:

we cannot understand a mind except as being indivisible. For we cannot conceive of half of a mind, while we can always conceive of half of a body, however small; and this leads us to recognize that the natures of mind and body are not only different but in some way opposite [contrariae.] (CSM II 9-10; AT VII 13)

Thus, the soul is indivisible and its possibility of destruction cannot be explained in the same way as the possibility of destruction of bodies. But this, of course, does not entail that there is no other way the soul could cease to exist, and in the Synopsis it is further claimed that the non-destructibility of the soul depends on its being a *pure substance* and that **pure** (puram) substances are by nature indestructible. But what does Descartes mean by a pure substance and why should it be indestructible? According to Marleen Rozemond (2010, 266), Descartes here wants to say that pure substances do not have a matter form –structure. Such a hylomorphic conception of substance requires that substances have a form which is generated from matter:

But a rejection of substantial forms also undermines the Aristotelian conception of hylomorphic substances which do not come to be as a result of creation, but which are composites that come to be as a result of substantial forms coming to be through natural processes (they are ‘educated from matter’) and which can cease to be through natural processes. While Descartes’s explicit target in this ‘theological’ argument is the existence of such forms, it also undermines the hylomorphic composites they help constitute, and that is what is relevant to his discussion of immortality in the Synopsis[...]

As it is well known, a hidden aim of the *Meditations* was to lay down a foundation for physics without obscure substantial forms. Thus, Descartes found in such a law based mechanistic physics also a support for the immortality of the soul: such a substance cannot be naturally generated or destructed.¹ It does not have any structure: neither internal nor that of form and matter. However, in the *Meditations*

¹ It is good to bear in mind that Descartes is speaking about natural immortality. God, of course, as an omnipotent substance has the power to annihilate even the pure substances he has created.

Descartes does not explicitly lay out the basics of his physics. Thus, an explicit immortality proof is also lacking from the *Meditations*.

2. Experience

Descartes was a philosopher of certainty. Certainty is not in the science. It is a theoretical study that leaves the mind somehow cold. The mind does not touch the object and does not internalize it.² The part one consideration in this paper has also been theoretical. Let us, then, try to find out whether there could be any experience where Descartes's considerations might give the reader evidence for immortality.

For me, there is no doubt that Descartes in the *Meditations*, most perpicuously, in Med1-2 took (ontological) experience seriously.³ Even his criterion of truth, clarity and distinctness, had experience of certainty at its basis. The famous existence "proof" of the first person (cogito) exhibits perhaps the most striking use of experience:

But there is a deceiver of supreme power and cunning who is deliberately and constantly deceiving me. In that case I too undoubtedly exist, if he is deceiving me; and let him deceive me as much as he can, he will never bring it about that I am nothing so long as I think I am something. So after considering everything very thoroughly, I must finally conclude that this proposition, I am I exist, is necessarily true when it is put forward by me or conceived in my mind. (CSM II 17, AT VII 25)

The proposition "ego sum, ego existo" is, to use Chisholm's language, self-presenting. Considering or conceiving it shows its truth. But the proposition in itself is not a logical or analytical truth(s) and prima facie, at least, it seems that this first-person existence "proof" does not have anything to do with immortality: it only tells that if I consider such a proposition, I cannot doubt it. However, if I do not exist I cannot consider that proposition. Thus this "proof" is consistent with my mortality. And Descartes is well aware of that.

However, there is a way that Descartes might have had at the back of his mind the claim that the existence proof gives a good basis for an immortality proof. In the second meditation, Descartes was in skeptical doubt and rejected the

² See Alquié 2018.

³ See Alquié 2018.

existence of the so-called external world. Such a rejection involved giving up the belief that there is a physical reality. And as it is well known, Descartes asserted that he was capable, while in such deep skepticism, not only to “prove” his existence but also to give an account of its nature as a thinking thing. The basic line of experiential thought here is like this: “even if you, O external extended world, do not exist, I do. My existence is not dependent on yours. I am a mind.”

But let us now try to find out what my non-existence would require. In general, a thought about something being non-existent seems to require a foundation. I can think of the tree in front of me as non-existent because I can think of this park as not containing the tree, I can think of Turku as non-existent because I can conceive Finland without Turku, and of any composite entity I can think of its non-existence as its parts not being related the way they are. If we can understand how an entity could be generated from some prior material with the aid of a cause, we are able to understand the non-existence of such an entity.

Now, the point is that it seems that there is no way to understand naturally how a self or I could be generated from some pre-existing material. If Descartes is right in that he can have a clear and distinct idea of himself, it seems that he also has evidence for saying that he cannot conceive his non-existence. And, the next step, of course, is this: inconceivability of the non-existence of something equals to the necessity of the existence of that thing.

I believe that this kind of Cartesian argument for the inconceivability of the non-existence of the I is rather natural. Of course, it might be objected that it is extremely easy to think of the non-existence of the self. I did not exist 1950 and it is very easy for me to believe that, and hence understand the proposition “I do not exist.” But, perhaps, a Cartesian reply to this objection would go as follows: a distinction has to be made between a human being and the soul or mind of that being, i.e., the I of that being. A human being is a union of mind and body, and such a body together with its contribution to the mind of that being are easy to conceive as non-existent. If “I” is thought to pick out that entity, then there is no hope of immortality. But it was Descartes’s aim in the second meditation to show that there is a use of “I” which refers to something that is prior to the I of the human being, i.e., an I that is a pure thinker: distinct from but united with the human body.

3. Problem

I believe that Descartes's self-conception is by and large right. However, some philosophers see it as involving a problem. It might be claimed that the non-existence of the self is inconceivable because the self as a substance is inconceivable. Thus, claims about its existence or non-existence both are incomprehensible. This kind of criticism has a long history, and as R.M. Chisholm tells, such a criticism joins the major 20th century schools of thought—analytic philosophy and phenomenology. The point is that the substantial self is not cognitively accessible. Chisholm (1975, 23) cites Carnap, Russell, and Sartre's *Being and Nothingness*. In fact, Sartre (2014,9) already in his "Transcendence of the Ego" emphasized what Chisholm calls the subjectless of the given by writing that "[t]he Cogito affirms too much. The sure and certain content of the pseudo-'cogito' is not 'I am conscious of this chair', but 'there is consciousness of this chair'."

The worry about the conceivability of a substantial is old. Basically, the same worry was already present in Gassendi's Objections to the *Meditations*. Gassendi was asking for an explanation or description of the nature of the self: something that we could give of substances in the extended world. But, of course, we cannot give such an explanation of a mental substance. So should we somehow reject the whole notion of a substantial thinking self as somehow meaningless? It is unfortunate that Descartes never was in the position to answer to Carnap, Russell, and Sartre. However, to Gassendi he answered and I believe his reply also meets the criticism of his later critics. Gassendi writes that to know what wine is, it is not enough to tell that it is liquid, that it is white or red, that it is intoxicating etc. Instead:

You will have to attempt to investigate and somehow explain its internal structure, showing how it can be seen to be manufactured from spirits, tartar, the distillate, and other ingredients mixed together in such and such quantities and proportions. (CSM II 193, AT VII 277)

And, then, Gassendi goes on to claim that Descartes should conduct a similar investigation of the thinking substance:

Similarly given that you are looking for knowledge of yourself...[y]ou should carefully scrutinize yourself and conduct a kind of chemical investigation of yourself, if you are to succeed in uncovering and explaining to us your internal substance. CSM II 193, AT VII 277.

The spirit of Gassendi's complaint is that the inner constitution of the thinking substance responsible for its functioning remains hidden. But Descartes's answer to Gassendi has striking power:

You want us, you say, to conduct 'a kind of chemical investigation' of the mind, as we would have of wine. This is indeed worthy of you...and of all those who have only a very confused conception of everything, and so do not know the proper questions to ask about each thing. But as for me, I have never thought that anything more is required to reveal a substance than its various attributes; thus the more attributes of a given substance we know, the more perfectly we understand its nature. (CSM II 248-249, AT VII 359-360)

Here Descartes' point, as I read it, is simply: The thinking substance differs from extended substance in that its essence or nature is an activity which is not to be accounted with the help of some inner structure. The phenomenal properties of wine are to be accounted in terms of inner structure but thinking substance, which in its essence consists in thinking, does not have such a structure. This may be difficult to understand due to some sort of obsession to see a mechanism behind everything or to try to understand everything through images.

What Descartes thought about the relation of ideas to existents involves a problem if he is considered as a full-blown/thoroughgoing representationalist. By representationalism, I mean a philosophical doctrine according to which true ideas point behind themselves or transcend themselves. Thus, in representationalism, the "I exist" of cogito should be conceived as a representation of my existence, and in general in representationalism we cannot get out of the layer of ideas. That ideas, no matter whether certainty is attached to them, really are about the things they seem to be about is ultimately a matter of faith.

However, it seems to me that Descartes never was a thoroughgoing representationalist. The proposition "I exist" is something that I can show to be true by thinking of it because I can think of it in such a way that I directly perceive

myself to be the thinker of it. It is through thinking that I come to have a direct de re touch with myself. We have already touched the subjectlessness of the given. The point perhaps is that in perception no self is given. This may be true but it is not true that *thinking* of what is given is subjectless. For Spinoza, for example, one cannot have a true idea without knowing that one has a true idea. It is the activity of thinking or in the activity of affirming that something is the case where the mind becomes aware of its ability to affirm.

4. Help from Spinoza and Kant

Let us take a closer look at how Spinoza introduces self-knowledge. In the *Ethics* part II proposition 43 Spinoza says:

He who has a true idea at the same time knows that he has a true idea, and cannot doubt the truth of the thing. [*Qui veram habet ideam, simul scit se veram habere ideam nec derei veritate potest dubitare.*]

This sentence says a lot. First, it gives certainty as a criterion of truth and second it brings the self into the content of the cognition. It is, as it were, in one the same act of apprehending where both what has to be the case and the one who is apprehending are cognized. The idea that $2+2=4$ does not involve a self but the cognition that “I have a true idea through the idea that $2+2=4$ ” does involve I. And that is what Spinoza in 2p43 claims.

For Spinoza, true ideas are clear and distinct ideas. There is self-evidency in them and they also are adequate ideas which are work of the mind: when the mind has an adequate idea, the mind acts. So if I have an adequate idea that $2+2$ is 4 this is because the mind puts together AA and AA and counts them with the result AAAA. It is not enough just to stare individually AA, AA, the identity sign and AAAA. AA must be joined to AA. Now, the subject doing the addition is conscious of herself doing that and this consciousness of the action is consciousness of the power of the agent and thus consciousness of the subject doing the sum. The cognition we have of ourselves is what I call *agent cognition*.

The reading of Spinoza just offered is a Kantian one. Kant’s big idea was that *the I* had the ability to unite several representations in one consciousness. Uniting representations in one consciousness means roughly uniting them in a

concept and such uniting is essential to thinking, believes Kant. For example, to think of a line consists in drawing a certain figure with the help of a concept:

But in order to cognize something in space, e.g., a line, I must draw it, and thus synthetically bring about a determinate combination of the given manifold, so that the unity of this action is at the same time the unity of consciousness (in the concept of a line), and thereby is an object (a determinate space) first cognized. (*Critique of Pure Reason*, B 137-8)

In such a cognition where one is conscious of the identity of the action (i.e., conscious of doing something over time), one is also is conscious of one's identity over time.

For Kant, the consciousness of the identity of oneself through the consciousness of the identity of the action does not prove the persistence of an underlying substance. If a substance is required for there being a consciousness of the identity, it still would be possible that such a substance is all the time replaced by some other substance:

a whole series of these substances may be thought, of which the first would communicate its state, together with its consciousness, to the second, which would communicate its own state, together with that of the previous substance, to a third substance, and this in turn would share the states of all previous ones, together with their consciousness and its own. The last substance would thus be conscious of all the states of all the previously altered substances as its own states, because these states would have been carried over to it, together with the consciousness of them; and in spite of this it would not have been the very same person in all these states. (*Critique of Pure Reason* Note to A 364)

I believe that Kant is right in his claim that the consciousness of the identity of myself as a thinker is nothing but the consciousness of the identity of an action. However, it seems that in the note to A364 Kant endorses such a substantial theory of the self where the self in itself is something that stays beyond its actions. Such

a self is superfluous. The self, I claim, is in its action. Perhaps, Descartes, in criticizing Gassendi, was thinking along those lines, too.

University of Turku

References

- Alquié, Ferdinand. 2018. "Expérience ontologique et déduction systématique dans la constitution de la métaphysique de Descartes." In *Cheminer avec Descartes. Concevoir, raisonner, comprendre, admirer et sentir*. Sous la direction de Thibault Gress. Garnier. 25-84.
- Chisholm, Roderick. 1976. *Person and Object: A Metaphysical Study*. George Allen and Unwin Ltd.
- Descartes, René. 1985. *The Philosophical Writings of Descartes*. Vol II. Translated by J. Cottingham, R. Stoothoff, and D. Murdoch. Cambridge U.P. Abbreviation CSM II.
- Kant, Immanuel. 1998. *Critique of Pure Reason*. Translated and edited by P. Guyer and Allen Wood. Cambridge University Press.
- Rozemond, Marleen. 2010. "Descartes and the Immortality of the Soul." In *Mind, Method, and Morality: Essays in Honour of Anthony Kenny*. Edited by J. Cottingham and P. Hacker. Oxford University Press. 252-271.
- Sartre, J.P. 2004. *The Transcendence of the Ego*. Translated by A. Brown. Routledge.

On Voluntary Servitude

EERIK LAGERSPETZ

1. Introduction: The Stone Paradox

Can God, who is omnipotent, create a stone He cannot move? This is an old problem of Scholastic theology (Geach 1977, chs. 1 and 2). In this paper, I try to show how an analogous problem appears in different political contexts, and why it may have important theoretical implications. Here a problem of this form is called an as instance of the *Paradox of Autonomy*. I try to show how this version of the paradox runs through the history of modern political thought. As the Paradox of the Liar, it is a paradox of self-reference. The paradox arises in every theory that contains propositions of the following type: there is an entity *a* which is supposed to be *autonomous* in respect of (or to have *control*, or *power* over) a set of states of affairs *C* in the sense that

(1A) for every state of affairs *s* belonging to the set *C*, *a* can act so that either *s* or not-*s* holds.

For some further reasons

(2A) The proposition (1A) is *necessarily* true.

And, again, for some theoretical reasons

(3A) the state of affairs expressed by (1A) belongs itself to the set *C*.

From (1A) and (3A) it follows that

(4A) *a* can act so that not (1A) holds.

Thus, *a* can lose its autonomy or control over *C* by performing a *self-limiting* act. There is nothing paradoxical in self-limiting acts. I can limit my legally defined autonomy by making a contract; I can restrict my physical autonomy by letting others tie me up. But in the cases considered here the premise (2A) states that the range of autonomy of *a* is *C* in every (logically, legally, morally etc.) possible world. On the other hand, from (4A) it follows that there is a possible world in which (1A) is not true. Thus, the premises (1A)-(3A) are mutually incompatible.

The paradox emerges when *a*'s range of autonomy is defined in a self-referential way, so that it does not exclude the possibility of *a*'s depriving its own autonomy. In most instances of the paradox, this possibility follows from an attempt to define the set *C* as the *maximal* set compatible with the other requirements of the theory in which the statements have been formulated. For example, the omnipotence of God may be defined so that He has autonomy in respect with every logically possible and non-tautologous state of affairs. If we suppose that the omnipotence of God is a conceptual truth, then, of course, a proposition stating that He can perform a self-limiting act (for example, create the immovable stone) expresses a contradiction. Thus, the paradox is solved. He cannot perform such an action. The problem in this solution is that sentences implying (A3) do not *seem* contradictory. We ordinary mortals can perform self-limiting acts, and if God cannot do the same, it seems that there are things which are impossible for an omnipotent Being but not for mortals. And this proposition *does* seem contradictory.

The version presented above, the Paradox of Divine Omnipotence, or the Stone Paradox, is the oldest and best-known formulation of our problem. Clearly, we have two possible solutions to the Stone Paradox: either, the concept of omnipotence has to be re-defined so that (3A) is no more true (*C* does not include all possible states of affairs), or the premise (2A) should be rejected, so that omnipotence is no more a necessary attribute of God. Generally, all versions of the Paradox can be solved in these lines. In the instances of the Paradox discussed here, the modal expressions "can" and "necessarily" express *normative* modalities. My aim is to show the relevancy of the Paradox for normative theories. Here, special attention is paid to the version in which *a* is an individual exercising his or her personal autonomy.¹

¹ Of another version, see Lagerspetz 1995, ch. 6.

2. Maximal freedom

There are several other versions of the Paradox of Autonomy. The most influential of them is the Paradox of Voluntary Slavery. Almost all great modern writers of political philosophy, from Grotius to Nozick, have commented on it. As far as I know, no one has noticed that it actually is the old Stone Paradox in another guise.

This version of the problem is relevant to most political theories that take freedom as the central value and are based on some conception of natural rights. Here is a famous example:

Men are born and always continue, free, and equal in respect of their rights.

(...)

Political liberty consists in the power of doing whatever does not injure another. The exercise of natural rights of every man has no other limits than those, which are necessary to secure to every other man the free exercise of the same rights.

The cited passages are from the *Declaration of the Rights of Man and of Citizen*, the manifesto of the French Revolution. The quoted second sentence expresses the following principle:

(1B) For every state of affairs, s , a (= any citizen) can (= is morally permitted to) act so that either s or not- s holds, if and only if his/her act does not injure any other person.

This way to define the range of citizen's autonomy is typical for liberal and proto-liberal theories: it expresses the idea of a *general right to liberty*. Many people share this idea. For example, according to Kant, people have a right to the maximal freedom compatible with similar freedom of others (Kant 1965 [1797], 35, 43-44). J.S. Mill states that “the only purpose for which power can be rightfully exercised over any member of a civilized community, against his will, is to prevent harm to others” (Mill 1972 [1859], 73). According to Herbert Spencer, “every man is free to do which he wills, provided he infringes not the equal freedom of any other man” (Spencer 1978 [1889-93], 62). The First Principle of John Rawls requires that “each person is to have an equal right to the most extensive basic liberty

compatible with the similar liberty for others” (Rawls 1973, 60). For Jürgen Habermas, there exists “the right to the greatest possible equal measure of individual liberties” (Habermas 1996, 122). In these principles, the range of liberty of any citizen is maximal and restricted only by the equal liberty of their co-citizens. Like God, individuals are supposed to be sovereigns within their own spheres of autonomy. The very idea of *maximizing* freedom or liberty presupposes that there is a meaningful way to compare amounts of (negative) freedom. Otherwise, a maximum could not be defined. This idea of comparing amounts of freedom may be challenged, but even if we do accept it, a problem remains. For the idea of maximizing freedom ignores the possibility that there may be *several* different maxima which satisfy the requirement. Indeed, there could be infinitely many. The idea of maximization may be compatible with numerous radically different systems of liberties.

These problems are indirectly relevant to our present discussion. A theory based on natural rights, like that expressed in the *Declaration*, claims that people are born with their rights; the rights are their *essential* attributes as omnipotence is an essential attribute of God. The first part of the quotation expresses this principle. It says, in effect, that

(2B) (1B) is *necessarily* true.

Let us suppose that someone tries to alienate his or her right to liberty by giving it up completely and irrevocably, e.g. by selling him- or herself into slavery. By doing such an act, does he or she injure some other person? Is his or her act incompatible with the rights of others? Not in any obvious sense. *Prima facie*, it seems possible to infer from (1B) that

(3B) *a* can (= is morally permitted to) act so that the state of affairs expressed by (1B) is no more true of him or her.

Proposition (3B) amounts to the permissibility of voluntary slavery. Most modern political theorists, including Locke, Montesquieu, Kant, Mill, and Rousseau, have explicitly rejected this possibility. As Mill (1972 [1859], 157-8) says

the principle of freedom cannot require that he should be free not to be free. It is not freedom to be allowed to alienate his freedom.

Mill notices that this conclusion is an exception from his anti-paternalistic "harm principle" quoted above. Thus, there is at least one type of action justly prohibited even when these actions do not cause any harm to others.

The problem constituted by (1B)-(3B) is an instance of the Paradox of Autonomy, and Mill's conclusion is an example of the strategy of avoiding the Paradox by restricting the set *C* of possible actions. As in the original Stone version, another possibility would be to reject the possibility of self-limiting actions. According to the latter solution, God cannot limit His omnipotence in any way. In this version, the corresponding solution would mean that *all* our enforceable voluntary commitments should be declared invalid. This seems to constitute a radical restraint of individual liberty. Unlike (perhaps) in the case of God, one important aspect of our personal liberty is our ability to constrain our liberty - at least in certain ways and to a certain extent.

3. Voluntary slavery in political philosophy: an overview

Why did the great political theorists consider it worthwhile to discuss the problem of voluntary slavery? Is it not obvious that no decent community can allow such a practice? The answer is that many thinkers *did* accept (3B), and that it played an important role in influential political theories. For example, Francisco Suarez, Hugo Grotius and, most famously, Thomas Hobbes all accepted that an individual may sell him/herself into slavery, and they used this as a justification of political absolutism. As Grotius (2013 [1625], I.iii.8, 51) put it:

To every man it is permitted to enslave himself to any one he pleases for private ownership, as is evident both from the Hebraic and from the Roman Law. Why, then, would it not be permitted to a people having legal competence to submit itself to some one person, or to several persons, in such a way as plainly to transfer to him the legal right to govern, retaining no vestige of that right for himself.

Thus, the comments of Locke, Blackstone, Montesquieu, Rousseau, Kant and Mill on voluntary slavery are all attempts to answer an influential political argument.

As to show the continuous relevance of the problem constituted by the premises (1B)-(3B), we may compare two modern libertarian thinkers, Robert

Nozick and Robert P. Wolff. Both start with Kantian assumptions and take freedom as their central value. Both draw radically libertarian conclusions. However, their solutions to the Paradox of Autonomy are diametrically opposite. According to Nozick, his

nonpaternalistic position holds that someone may choose (or permit another) to do himself *anything*, unless he has acquired an obligation to some third party not to allow it. (Nozick 1974, 59)

The comparable question about an individual is whether a free system will allow him to sell himself into slavery. I believe it would. (ibid, 331)

By contrast, Wolff thinks that liberty or autonomy is not only a right but also a duty:

The moral condition demands that we acknowledge responsibility and achieve autonomy wherever and whenever possible. (Wolff 1976, 42)

It follows that any kind of submission to any authority is morally prohibited. Thus, for example, a commitment to the majority principle is "no more than voluntary slavery". For

neither majority rule nor any other method of making decisions in the absence of unanimity can be shown to preserve the autonomy of individual citizens. (Wolff 1976, 42).

According to the mainstream liberal consensus, people may legitimately restrict their own autonomy, but not abandon it completely and irrevocably. They may sell their labour, and thus give up a part of their freedom. Indeed, the possibility of performing such self-limiting acts is central in all versions of liberalism. It is the basis of all contracts, promises and other voluntary undertakings. They may sometimes limit individual freedom in a dramatic way. People may, for example, join a religious order and give up their rights to own property or to exercise freedom of speech. Some theorists believe that people are allowed to commit suicide. What people cannot do is to cease to be moral persons while remaining

persons in the empirical sense. The question is whether there is any way to justify this way to limit self-limitation that is not *ad hoc*.²

Let us look at the classic arguments for the inalienability of basic liberty. In his *L'esprit des lois*, Montesquieu presents almost all the standard arguments:

Neither is it true that a freeman can sell himself. Sale implies a price; now when a person sells himself, his whole substance immediately devolves to his master; the master therefore, in that case, gives nothing and the slave receives nothing (...). If it is not lawful for a man himself because he robs his country of his person, for the same reason he is not allowed to barter his freedom. The freedom of every citizen constitutes a part of the public liberty, and in a democratic state it is even a part of sovereignty. To sell one's freedom is so repugnant to all reason as can scarcely be supposed in any man. If liberty may be rated with respect to the buyer, it is beyond all price to the seller (...). The same law annuls all iniquitous contracts; surely then it affords redress in a contract where the grievance is most enormous. (Montesquieu 1949 [1748] X.v.2, 236)

We can distinguish four separate arguments from Montesquieu's text. First, an expression of consent for voluntary slavery is void for *conceptual* reasons. Second, such a consent is also *immoral*, because it violates our duties to third parties, the society and the State. Third, it is *irrational* to express consent to slavery, and contracts made by irrational persons should not be enforced. Finally, a slavery contract is always *unfair*.

Only the first, conceptual, argument is strictly non-paternalistic. It was used by William Blackstone, the influential 18th century English jurist:

Every sale implies a price, a *quid pro quo*; can an equivalent be given for life, and liberty, both of which, (in absolute slavery) are held to be in the master's disposal? His property also, the very price he seems to receive, devolves to his master, the instant he becomes a slave. In this case, therefore, the buyer gives nothing, and the seller receives nothing: of what

² There is some contemporary discussion of the problem. Of the contemporary commentators, Kershnar (2003) defends voluntary slavery as compatible with liberal principles; according to Freeman (2001), the problem shows that libertarianism is not a liberal view; Ellerman (2010) sees the possibility of voluntary slavery as a "litmus test" of theories of justice.

validity, then can a sale be which destroys the very principles upon which all sales are founded? (Blackstone 1876 [1765] I.xiv, 393)

Before analysing the argument in detail, a word should be said about “slavery”. The notion of slavery used in the arguments is that of a *complete and irrevocable* alienation of *all* rights, “absolute slavery” as Blackstone says. In absolute slavery, a slave has no rights against his or her master. (This does not exclude the possibility that slave-owners have duties *regarding* their slaves, as they may have duties regarding their domestic animals.) “Absolute” voluntary slavery is, then, an ideal type, not a description of any institution that has actually existed anywhere. Basically it means that (i) someone accepts that he or she has no rights, (ii) other people treat him or her as a creature who has no rights, and (iii) they justify (ii) by appealing to (i). A “voluntary slave” could be a person living under the reign of an elected dictator, or a married woman in an extremely patriarchal society.

4. Kantian arguments

In the form presented by Montesquieu and Blackstone, the conceptual argument does not convince. Persons who want to sell themselves into slavery may impose a condition that the price should be paid to a third party, or that they are allowed to use the price before the slavery contract comes into force. Or, then, they may give up their liberty just for nothing. Immanuel Kant formulated a more sophisticated version of the conceptual argument:

...no-one can voluntarily renounce his rights by a contract or legal transaction to the effect that he has no rights but only duties, for such a contract would deprive him of the right to make a contract, and would thus invalidate the one he had already made. (Kant 1977 [1793], 75)

In effect, Kant claims that the slavery contract produces a contradiction. A person who accepts such a contract says something like: “Hereby, I consent to being without any rights”. A valid expression of consent presupposes that the consenting person has a right (or power) to give his/her consent. In the self-enslavement case, this seems to produce a simple self-referential paradox analogous to the Liar. However, the argument is not satisfactory. A person saying, “All my *future statements* will be untrue” does not produce the Liar Paradox. Similarly, while a

slavery contract invalidates all future contracts made by the slave, the contract itself may remain valid. The terms of the contract should actually be read as “I consent *now* to being incompetent to give my consent *in any future occasion*”. Kant might reply that if slaves have no rights, they cannot have any moral duties. Then, the slavery contract would annul itself, for it would not bind the slave. For example, the slave would not be under duty not to escape. However, the contract might still give rights to the slave owner against *others*. The slave owner might have a right that the contract is enforced, that an escaping slave is returned in due course, and if the others tried to assist the slave in an attempted escape, they would be punished. It is true that the slave him/herself cannot be legally responsible for breaking the contract. Nevertheless, the slave owner may justify his position as the owner of the slave by referring to the slave’s earlier consent. Legally, the slave is like a dead person, but his or her consent, like that of a dead person, may still have legal effects.

Kant's argument against voluntary slavery has been interpreted as a direct consequence of the first version of his Categorical Imperative. If we *all* renounced our autonomy, there would be no masters, and, hence, no slaves either. The maxim behind a slave contract could not become a universal law (O’Neill 1989, 96). However, this argument is subject to the notorious problems related to Kant's first formulation of the Imperative. For example, we could not all become teachers, for then there would be no students left. When read in this way, the Categorical Imperative seems to exclude *all* asymmetric voluntary social relationships. This is implausible. A slave contract may violate Kant’s Categorical Imperative, but the reason cannot be just the asymmetrical nature of the master-slave relationship.

The *second* version of Kant's Imperative, which forbids us to treat people only as a means, seems to be a more promising starting point. A slave is used as a means in an obvious sense; indeed, this is more or less a part of the meaning of being a slave. However, it is not quite that simple. What exactly means that we should treat people as *ends* (rather than as a means)? A person cannot literally be an “end” of another person, for persons are not something that could be pursued or achieved. We treat somebody “as an end” by respecting those ends he or she has autonomously chosen. The Kantian prohibition to use other people *only* as a means does not mean that human beings cannot be used as a means at all. Suppose that my car is broken, and I have to use the services of a taxi driver. Quite literally, I use the driver as a replacement for a mechanical instrument: if I had a working car, I would have no need for a driver. The driver is valuable for me only as long

as she helps me to realize my own ends. However, as long as my using her as a means serves her own ends, too, I am not using her *merely* as a means to my ends. I do not deprive her status as an autonomous person who is capable of choosing her own aims, for among her autonomously chosen ends happens to be the end that she is temporarily sharing with me.

The last point is worth making, for some anarchists and radical socialists have claimed that *every* labour contract is unjust because it, by definition, implies that someone is using somebody else as a means to ends. For example, the anarchist political theorist L. Susan Brown (1993, 4) has argued how

at the moment an individual “sells“ labour power to another, he/she loses self-determination and instead is treated as a subjectless instrument for the fulfilment of another’s will.

According to Brown’s view, the prohibition to use other persons as means excludes *all* instrumental relations between human beings. This argument should be distinguished from the more conventional radical argument. According to the latter argument labour contracts may be invalid if the bargaining situations in which they are made are unequal and unfair. If Brown’s argument is a *conceptual* one, it holds even when the individual selling his or her power actually has an upper hand and can unilaterally dictate the terms of the contract. If the aim of the contract is to make him or her an “instrument” of his or her employer, he/she loses his/her self-determination even when the bargain is highly beneficial for him/her. This sounds implausible. Would we really like to say that a world-famous film star or a top executive is, for purely conceptual reasons, in the same position as a voluntary slave when making an employment contract? Are the bargaining positions of the contracting parties irrelevant for the evaluation of the moral status of a contract? (Paradoxically, Hobbes also thought so!) Actually, Brown’s version of anarchism is bound to limit our individual freedom in a very radical way. The mainstream liberals are quite correct in insisting that the possibility to perform *some* self-limiting acts is an essential element of our individual liberty. A person who is unable to make contracts or give promises of certain type is, *prima facie*, less free than a person who is able to make them.

None of the “Kantian” arguments discussed above are able to show why slave contracts are unacceptable. If my aim is to make myself a slave of somebody else, my prospective owner seems to respect my autonomous will by accepting the

contract. After making the contract my owner is no more bound to respect my later autonomous will. Supposing that I was autonomous when I signed the contract, my owner (or the society) would not respect my earlier autonomous will if he (or the society) refuses to treat me as a slave. It seems that we are in the same situation as before: both logically possible alternatives respect the freedom in one sense and violate it in another sense.

However, there is still one Kantian route left. Perhaps it is *I* rather than my prospective owner who violates the moral law by accepting the terms of the slavery contract. (Then, my prospective owner may be a partner in crime by helping me to execute my immoral project.) By signing an enslavement contract, do I violate the third version of the Categorical Imperative by treating a moral person as a means only? If I do, the person in question must be I myself. Presumably, I will get some benefit from the contract, and for the sake of that benefit, I am willing to deprive my own autonomy. Then, I am treating my *future self* only as a means to my present ends. For whatever ends I might form in the future, as a slave my attempts to realize any of them would be completely dependent on the will of my slave-master.

5. Mill and Sidgwick on voluntary slavery

The third argument of Montesquieu – based on the irrational nature of slave contracts – is probably the first one that comes into mind. Any proponent of liberty would accept the principle that contracts made by deeply irrational persons are null and void. In the slavery case, this was the line chosen by Rousseau and Mill. It is interesting to note that in the case of voluntary slavery, almost any liberal would be willing to apply Rousseau's notorious paternalistic principle that sometimes people "shall be forced to be free". If irrationality is the main (some would say the only) justification of paternalistic interventions, we need some criteria for irrationality. As Sidgwick (1908 [1891], VI§5) remarks, the liberal individualist has the burden of proof:

It is evident that if there are no such limitations, a man may exercise his freedom for its own destruction by contracting himself into slavery: and an individualistic system that allowed this would defeat its own fundamental aims. Still, the question is not free from perplexity. For any binding contract to be performed at a future date must diminish to some extent the

contractor's freedom of action when the time arrives. How then, it may be asked, can a line be drawn? If A may pledge himself to work for B for a year under conditions, why not for life unconditionally, if he judges this to be for his own interest. How can any limitation be justified except on the anti-individualistic assumption that the average man cannot estimate his own future interest as well as government can?

Without going deeper into the long and complicated discussion on paternalism and its justifications, we may adopt a useful distinction. A paternalistic intervention is based on a *person-mediating maxim* if it is seen as justified only when a person lacks certain mental characteristics (e.g. mental balance or adequate information), and when this fact can be ascertained in an independent way (e.g. by asking, making a test, or observing his or her overall behaviour). A paternalistic intervention is based on an *act-mediating maxim* if it is seen as justified whenever it is necessary to prevent acts of certain type (e.g. suicides). An act-mediating maxim can be justified by the irrationality of the subject of paternalistic treatment, if the prevented acts (say, attempts to commit suicide) are considered to be irrational by definition, so that no further evidence is needed. In the case of person-mediating maxims, the prevented act itself can be one sign of irrationality, but it is not sufficient. We need further evidence about the person's state of mind (Gray 1990). It is interesting that in *On Liberty*, voluntary slavery is the *only* case in which Mill applies an act-mediating maxim. All his other examples of justified paternalism – e.g. of the man who tries to cross a bridge without knowing that it is falling – are based on person-mediating maxims.

In his *Political Economy* (1848) Mill elaborates his anti-paternalist argument. As a utilitarian, Mill does not ground his principle of liberty on inalienable rights, but on the “presumption in favour of individual judgements“. Generally, individuals know their own interests better than any outside observer. Thus, paternalism is illegitimate. Acts can be in the interests of individuals even if they seem irrational to an outside observer. However, in the case of contracts which bound individuals “to continue doing something for ever or for a prolonged period, without any power of revoking the engagement“, the presumption no longer holds. In such cases, judgments cannot be “grounded on actual personal experience“ (Mill 1872 [1848] V.x.10, 579). For Mill, slave contracts are illegitimate because they are *ex hypothesis* irrevocable. However, it should be noticed that almost all of our important choices have *some* irreversible effects, and

some of them may be fatal. Mill might argue the slavery case has an exceptional status, for in this case the consequences of individual myopia are so grave that the state cannot allow individuals to make their own mistakes. In this particular case, the preferences of a prospective voluntary slave *have* to be based on insufficient information or on mistaken judgements; there is no need to prove it. But if we can show this in the slavery case, we may well be able to show it in other cases, too. For a liberal utilitarian, the problem is: how to justify paternalism in the slavery case without also justifying it elsewhere. Several contemporary theorists have commented on Mill's antislavery argument (Arneson 1980; Feinberg 1971; Fuchs 2001; La Selva 1987). Some – for example Joel Feinberg – have seen it simply as an *ad hoc* exception of Mill's general presumption: “voluntarily acceding to slavery is too much for Mill to stomach“ (Feinberg 1971, 118).

However, Mill was not just a utilitarian and a liberal individualist. In the present context, his view of human nature is essential. It was strongly influenced by Wilhelm von Humboldt's *Ideen zu einem Versuch, die Grenzen der Wirksamkeit des Staats zu bestimmen* (1792). von Humboldt believed in human perfectibility, but unlike most philosophical perfectionists, he thought that the role of legislation was mainly to remove obstacles to individual development. The law should make contracts possible, but it should not enforce contracts that prevent individuals' development towards perfection (von Humboldt 1969 [1792], 133-4). Mill subscribed to von Humboldt's perfectionism. Both society and individuals are evolving beings. *All arrangements intended to be immutable*, whether they be laws, religious dogmas, or private contracts, could become obstacles in the way of development. Like Mill's presumption for individual judgement, this view is based on a general but plausible empirical conjecture. Here, as in many other questions, Henry Sidgwick formulated the answer in the way that could have pleased Mill, too:

The answer is that the individualistic principle does not rest solely on the assumption of the actual superiority, generally speaking, of the individual's practical judgment where his own interests are concerned: it is based partly on a belief in the educative value of freedom as developing foresight, and not only foresight, but the moral qualities of self-reliance and energy. Hence it is not on the whole opposed to individualism to treat as invalid any contract resigning the rights to personal liberty and security and the right to acquire and dispose of property: it merely implies that the general

assumption that the individual can be trusted to look after his own interests has at a certain point to give way to the general assumption that freedom is good for him. (Sidgwick 1908 [1891], VI§5).

6. Conclusion

A natural-right libertarian moral system, like that of Nozick, includes at least the following principles: (1) every rational individual has a right to maximal (negative) liberty compatible with similar liberty of others. (2) All rights either belong to individuals by nature or follow from voluntary agreements between them. (3) For every duty there is a corresponding right possessed by some other individual(s). It seems that such a system has to permit total alienation of liberty. Most of us would see Nozick's conclusion as a *reductio ad absurdum* result: from a system based on maximal liberty, we can derive the possibility of complete annihilation of liberty (Ellerman 2010; Freeman 2001). What could be a more effective refutation of a moral theory than a demonstration that the theory may allow a total disregard of its most important basic value?

If we want to stick to the liberal prohibition of voluntary slavery, we are forced to reject at least one of the assumptions (1)-(3). Personally, I would reject all of them (I have already pointed out some problems in (2)), but the rejection of assumption (3) is enough for our present purposes. Even if rights invariably entail corresponding duties, duties need not entail rights. The liberty principle implies the right (and power) to limit liberty by consent. If there were a duty not to alienate one's liberty, according to (3) there should be *someone* who had the corresponding right. In the system of Locke, there was such a being: our liberty was granted by God, together with the duty to preserve ourselves and the rest of mankind, and the alienation of liberty was incompatible with that duty³. Similarly, for Montesquieu, we have duties towards the society in general. These duties are not correlated with the rights of any particular member of the society. Finally, according to Kant, we have duties to ourselves. His ultimate argument against self-enslavement is analogous to his argument against suicide. Because we cannot plausibly have

³ We may say that, for Locke, our natural rights are inalienable for the same reason as a policeman's right to make an arrest is inalienable. We have those rights because they are granted for us by a supreme power – God – in order to perform a specific task. We have rights because they allow us to fulfil our duties. Nozick does not include God into his system, so he is free to dispense the inalienability argument – but unlike Locke, he has no explanation why we do have any rights in the first place.

corresponding *rights* against ourselves, this argument entails a denial of the thesis that all our duties correspond with somebody's rights. Because modern libertarians like Nozick require that all our duties should be correlated with rights of other individual persons, arguments against voluntary slavery are not available for them. Thus, the rejection of inalienable rights in modern libertarianism is not just a minor revision within the liberal tradition. The fact that Locke, Kant, Blackstone, Montesquieu and Mill admitted the existence of duties that were not correlated with rights of other individuals (*contra* (3)) shows that their moral and political theories were actually *very* unlike those defended by the modern libertarians (and by some liberals). The change reflects the rejection of the republican influence within the liberal tradition. In the republican view, liberty invariably carries duties along it. Indeed, all the 17th and 18th century thinkers, and even many 19th century liberals like Herbert Spencer, thought that *we have rights because we have duties*, and not the other way round.

Moral theories are often classified as 'right-based', 'goal-based' and 'duty-based' (Dworkin 1977, 90-6). But this classification is far too simple. One might think that in a genuinely right-based theory, only other rights can override rights. However, the *general goal* to protect a system of rights may ground duties that are stronger than particular rights within that system. As individuals we may have a duty to defend a state which recognizes our rights, even if such a duty cannot be correlated with the rights of any particular individual. Such a view of political morality is still right-based in the sense that rights remain fundamental, but they do not work as 'trumps' (Dworkin) or 'side-constraints' (Nozick) against such collectively determined goals or duties which have the protection of the framework of rights as their purpose. There can be a trade-off between the *amount* of liberty enjoyed at a particular moment and the *security* of liberty in the future.

We see the point more clearly, when we consider what it would actually be like to live in a society where slave contracts were recognized. Although slavery would be voluntary, life would not be essentially different from a society in which it is non-voluntary. The courts and officials would have a duty to return runaway slaves back to their owners. The police would have to assist the slave-owners to put down slave rebellions. Non-slave-owning citizens would have to finance this by taxes. They also could be tried and punished for attempts to hide escaping slaves. Allowing voluntary slavery would change social relations in a way that would be, in the end, inimical to all liberty as well as to many other basic values. We would all become *potential slaves* who might be persuaded or bribed to

become merely instruments. Our duty not to allow ourselves to be enslaved and not to recognize the slavish status of others (voluntary or not) is a duty to our fellow citizens in general, not only to those who are actually treated as slaves. The argument is *not* that voluntary slavery cannot be allowed because we have to protect the potential slaves. It is rather that we have to protect *freedom in the entire society*. As Montesquieu says, “the freedom of every citizen constitutes a part of public liberty, and in a democratic state is even a part of the sovereignty“. According to the argument, a would-be slave has no right to realize his or her selfish (or non-selfish) desire to become a slave. Willing slaves are under duty to bear the burden of freedom. Note that in this view, freedom or liberty is still defined in the negative way: it consists of the liberties and powers of an individual citizen. However, these liberties can be safely enjoyed only in a free society. Because some forms of voluntary relinquishment of individual liberty have the tendency to undermine the free society, the precondition of these liberties and powers, such self-limiting acts should be prohibited. All this sounds rather obvious. However, sometimes it is useful to give reasons even for those truths we hold to be self-evident.

University of Turku

References

- Arneson, R. I. (1980), “Mill versus Paternalism”, *Ethics* 90, 470-489.
- Blackstone, W. (1876 [1765]), *The Commentaries of the Laws of England*, John Murray, London.
- Brown, L. S. (1993), *The Politics of Individualism. Liberalism, Liberal Feminism, and Anarchism*, Black Rose Books, Montreal.
- Dworkin, R. (1977), *Taking Rights Seriously*, Duckworth, London.
- Ellerman, D. (2010), “Inalienable Rights: A Litmus Test for Liberal Theories of Justice”, *Law and Philosophy* 29, 571-599.
- Feinberg, J. (1971), “Legal Paternalism”, *Canadian Journal of Philosophy* 1, 105-124.
- Freeman, S. (2001), “Illiberal Libertarians: Why Libertarianism is not a Liberal View”, *Philosophy and Public Affairs* 30, 105-151.
- Fuchs, A. E. (2001), “Autonomy, Slavery, and Mill’s Critique of Paternalism”, *Ethical Theory and Moral Practice* 4, 231-251.
- Geach, P. (1977), *Providence and Evil*, Cambridge University Press, Cambridge.
- Gray, J. (1990), “Freedom, Slavery and Contentment”, in J. Gray (1992), *Liberalisms: Essays in Political Philosophy*, Routledge, London, 69-88.
- Grotius, H. (2013 [1625]), *Hugo Grotius on the Law of War and Peace*, ed. S. C. Neff, Cambridge University Press, Cambridge.

- Habermas, J. (1996), *Between Facts and Norms*, transl. William Rehg, MIT Press, Cambridge (MA).
- von Humboldt, W. (1969 [1792]), *The Limits of State Action*, transl. J. W. Burrow, Cambridge University Press, Cambridge.
- Kant, I. (1977 [1793]), "On the Common Saying 'This may be True in Theory, but it does not Apply in Practice'", in *Kant's Political Writings*, transl. H. B. Nisbet, Cambridge University Press, Cambridge 61-92.
- Kant, I. (1965 [1797]), *The Metaphysical Elements of Justice (Part I of the Metaphysics of Morals)*, transl. J. Ladd, Bobbs-Merrill Co, Indianapolis.
- Kershner, S. (2003), "A Liberal Argument for Slavery", *Journal of Social Philosophy* 34, 510-536.
- La Selva, S. (1987), "Selling Oneself Into Slavery: Mill and Paternalism", *Political Studies* 35, 211-223.
- Lagerspetz, E. (1995), *The Opposite Mirrors. An Essay on the Conventionalist Theory of Institutions*, Kluwer, Dordrecht,
- Mill, J. S. (1972 [1859]), *On Liberty*, in J. S. Mill, *Utilitarianism, On Liberty and Considerations on Representative Government*, ed. H.B. Acton, Dent, London.
- Mill, J. S. (1872 [1848]), *The Principles of Political Economy*, Longmans, London.
- Montesquieu, (1949 [1748]), *The Spirit of Laws*, transl. T. Nugent, Hafner Press, New York.
- Nozick, R. (1974), *Anarchy, Sate, and Utopia*, Basil Blackwell, Oxford.
- O'Neill, O. (1989), *Constructions of Reason. Explorations of Kant's Practical Philosophy*, Cambridge University Press, Cambridge.
- Rawls, J. (1973), *A Theory of Justice*, Oxford University Press, Oxford.
- Sidgwick, H. (1908 [1891]), *The Elements of Politics*, MacMillan, London.
- Spencer, H. (1978 [1889-93]), *Principles of Ethics, vol. II*, Liberty Press, Indianapolis.
- Wolff, R. P. (1976), *In Defense of Anarchism*, Harper & Row, New York.

Frequently Asked Questions about Metaphilosophical Naturalism

HENRI PETTERSSON

A good philosophical conversation is like running a marathon, not a sprint. One needs patience, dedication and resilience to cope with adversity along the long miles. With that mindset, I here backtrack all the way to the starting line of one of the most divisive and far-reaching philosophical debates of our time, that is, the case of metaphilosophical naturalism. But before I can shuffle my jogging shoes even one step further, I must ask the reader to temper whatever expectations you might have. The present text does not hope to add any groundbreaking arguments or particularly deep insights into the already voluminous literature on this topic. Instead, my modest aim with this compact paper – if you can even call this scribble a paper – is to clarify the nature of naturalism (at least according to the broad sketch of it offered here) and to dispel certain persistent misunderstandings which seem to follow this metaphilosophical position closely.

I feel there is still a market for this kind of groundwork. There exists a lot of preconceptions, pigeonholing, loaded stereotypes and just plain toxic rhetoric in the metaphilosophical debates on naturalism – coming equally from representatives of both opposing parties. One way in which these attitudes rear their ugly heads is, for example, when philosophers at times cast doubt on the ulterior motives or philosophical acumen of their colleagues. It can be difficult to fathom why everyone else does not share our views on these issues when we ourselves are 100% rational and base our conclusions on nothing but cold logic and hard facts – surely there must then be something deeply flawed with the premises of any alternative lines of thinking. Maybe naturalists, possessing an academic inferiority complex about their background in philosophy, simply want to get on the bandwagon of the natural sciences and join the winning team of our current scientific zeitgeist? From the opposite direction maybe all critics of

naturalism are just extending their prior anti-scientific convictions, which might perhaps stem from a religious worldview, into the realm of philosophy? Alas, such rhetorical moves are nothing new, since raising accusations of irrationality and opportunism are recurring tropes in the philosophical playbook, all the way from the polemics between philosophy and sophistry in ancient times to the more recent sociological divide between analytic and continental forms of philosophy (Glendinning, 2002).

Secondly, a metaphilosophical big picture such as naturalism is not necessarily felt when one is busy engaging in discussions on the grassroots level of so-called first-order philosophy. Instead, metaphilosophical visions loom somewhere in the background on a higher theoretical plane of their own, and many philosophers might not even be aware of these foundational concerns. Consequently, the actual commitments and implications of metaphilosophical views can remain distant and vague. I hope to raise awareness of these issues. This will not, however, be a neutral and unopinionated primer to the topic, as my own sympathies have been on the side of (moderate) naturalism since the earliest days of my philosophy studies at university. As a result, the text may appear like a covert attempt to vindicate naturalism.

One more thing: I have framed this text in the form of questions and answers. The joyous occasion that gave rise to the collection where this paper appears allows some leeway for a somewhat more unconventional and playful approach, and I have taken advantage of that opportunity. I make no claims about who actually asks these questions and how frequently that might happen. It is an *ad hoc* narrative device, so please play along and do not take this exercise too seriously.

After this brief warm-up lap, it is time to get our feet really moving with the first question which may already be on the minds of some readers. (Will the “funny” allusions to running end at some point? No, they will follow us all the way to the finish line.)

1. What exactly is metaphilosophical naturalism?

Metaphilosophical naturalism¹ is a conception of philosophy. As such, it offers a comprehensive, philosophically argued outlook on the proper subject matter,

¹ A terminological note: “naturalism” is one of those devilishly slippery terms in philosophy, since it is employed in different meanings in various branches of philosophy – metaphysics, epistemology, ethics and many numerous other instances. For the purposes of this short paper,

methodology and aim of philosophy itself (Overgaard, Gilbert & Burwood, 2013, 11). What are philosophical theories about? How can we gain knowledge of these matters? What exactly is our prime directive as philosophers? Among the established academic disciplines, philosophy is somewhat unusual in the sense that, even more than two thousand years after its initial formation, such fundamental questions are still the subject of fierce disputes among philosophers (Glock, 2008, 7). Here different conceptions of philosophy come in as an explanatory tool to make sense of this colorful internal diversity within the philosophical world. It can be said that philosophizing is always practiced from within the point of view of one conception of philosophy or another, so every philosopher has already some sort of an answer to the issues of what philosophy is, how it is practiced and to what end – even if this answer might be only unarticulated in their minds and when spelled out turn out to be philosophically incoherent. When we work explicitly on theoretical questions constituting conceptions of philosophy, we are doing metaphilosophy. Metaphilosophy might alternatively be called “philosophy of philosophy” in line with the established sub-disciplines of philosophy of history, philosophy of physics, philosophy of biology and so on – the only thing different now is that we turn philosophy’s gaze on philosophy itself (Williamson, 2008).

The distinctive take naturalism offers on the core metaphilosophical issues is to highlight the qualitative kinship between philosophical theorizing on the one hand and scientific research on the other. For naturalists, these two modes of inquiry are seamlessly continuous with each other at their edges and they work in reciprocal collaboration furthering shared goals. There are no sharp borders between different special sciences, and philosophy is no exception here. Although naturalists agree that typical philosophical problems differ in many significant respects from the questions studied by the empirical special sciences, naturalists also think that the empirical data provided by the non-philosophical modes of inquiry can have an instructive, corroborative, corrective and also falsifying relevance for philosophical considerations.

In metaphilosophical naturalism philosophy thus exists in a symbiotic relationship with the special sciences. This alliance means that philosophers can assist scientists with their contributions but also that philosophy is itself an equally

naturalism is a metaphilosophical position, and while it has natural connections with many other sub-forms of naturalism across philosophy (especially metaphysical and epistemological naturalism), I must exclude them from this overview.

receptive member in this partnership without any extraordinary privileges. Indeed, this receptiveness can even be a vital condition for progress in philosophy, so philosophers must be vigilant and keep their eyes and ears open to what is going on in the neighboring sciences. Accordingly, philosophical research will often be most fruitful in situations where specialists hailing from different disciplines combine their forces to study a certain phenomenon and cross-pollination between various approaches can occur.

I feel that at this point an illustration might be helpful. To use a completely randomly chosen case example from the philosophical literature, consider the philosophical reflections on the paradoxical nature of colors (Arstila, 2005). For centuries, philosophers have tried to understand just what colors are. Ripe bananas are yellow, grass is green and the clear summer sky is blue. That much is self-evident from our everyday experiences, yet the true nature of colors remains elusive to us. Are they properties of physical objects in the mind-independent external world or rather projections created by our cognitive systems? It would seem that neither philosophy nor empirical science alone can fully explain what colors are, but we need input from both in our final account.

This initial characterization of naturalism sounds uncomplicated and unproblematic – almost too good to be true. It can be hard to see who would want to oppose this idea and why. This segues us into the next section.

2. Who opposes naturalism?

On the basis of its core metaphilosophical doctrines, naturalism contrasts with a number of long-established and influential conceptions of philosophy which, for various reasons, argue that philosophy should be recognized as a *sui generis* and autonomous activity among the academic disciplines. Philosophy is the best and, in fact, the only expert on its own questions. I use the collective label “anti-naturalism” for these views, although it designates a broad and heterogenous category – the conceptions of philosophy grouped together here do not necessarily share any other metaphilosophical theses beyond their general opposition of naturalism (this sentiment is already present in the very name “anti-naturalism” itself, which is clearly introduced as a reactionary antonym for the name “naturalism” without a positive message of its own). Indeed, many forms of anti-naturalism are actually contradictory with each other on other important

metaphilosophical issues if we take a closer look at their visions of what philosophy should be.

It is crucial to note at the outset that the motivation for advocating anti-naturalism does not (necessarily) stem from any deep or pre-philosophical conviction against the natural sciences. In fact, many anti-naturalists may well hold sciences in high regard and conceive of philosophy as working in active collaboration with the special sciences (a good historical example of this approach are the logical positivists of the Vienna Circle). Nevertheless, even in such collaborative endeavors anti-naturalists want to maintain that different fields of study must remain in their own lanes on the road. Philosophical questions differ qualitatively from those of the special sciences. For some anti-naturalists, autonomous philosophy can also claim the position of some kind of “first philosophy”, so that philosophy possesses at least some amount of authority over the sciences (Bealer, 1998). Such an authoritative ambition, however, does not need to be an essential constituent thesis of anti-naturalistic conception of philosophy.

The philosophical reasons and arguments used to motivate metaphilosophical anti-naturalism are related to an idiosyncratic philosophical method, object or aim (there are also natural links between these three dimensions, so often an anti-naturalistic conception of philosophy has interrelated takes on these issues). Maybe the nature of philosophical knowledge is *a priori* while the empirical sciences deal with *a posteriori* way of knowing? Maybe normative, transcendental or conceptual issues are the proper province of philosophy whereas the empirical sciences focus on strictly factual matters? Maybe philosophy, unlike the truth-oriented sciences, is not a cognitive enterprise in the sense that it does not actually pursue truths or build theories? These are just a few rough illustrations and, as was already stated, there is no single big family tree of anti-naturalism of which all such conceptions of philosophy are further branches, originating from the same historical roots and sharing the same philosophical DNA. Thus, the merits and weaknesses of anti-naturalist conceptions of philosophy must always be considered on a case-by-case basis without sweeping generalizations. However, I will not delve into these discussions here as they are not within the scope of this overview. Instead, I will now consider how important this whole metaphilosophical debate really is, and whether naturalism and anti-naturalism exhaust all possible options we have here.

3. What is at stake?

The question of metaphilosophical naturalism has become an increasingly more popular topic during the past decades. Overgaard, Gilbert and Burwood (2013) regard it as *the* metaphilosophical question for our times, as do Cappelen, Gendler and Hawthorne (2016, v). This development has been partly driven by the rise of multidisciplinary ventures such as cognitive science, which has stimulated philosophers to rethink and justify their roles as contributing members in these scientific group projects. Moreover, modern science policy, which encourages researchers to cross the borders separating disciplines and to seek multidisciplinary collaboration, has also for its part encouraged this metaphilosophical self-reflection among philosophers.

In recent projections (see Glock, 2008, 257–9) concerning the future megatrends of philosophy, certain analysts have raised the possibility of such a dramatic turn of events where the deep metaphilosophical differences between the naturalists and the anti-naturalists could with time escalate to split philosophers into two opposite camps – not unlike the way in which the analytic/continental divide did earlier for the better part of the twentieth century. In this way the contrast between naturalistic and anti-naturalistic conceptions of philosophy could provide us with a new metaphilosophical frame of reference, which reflects more accurately the sociological converging of the “post-analytic age” we are currently undergoing in philosophy. Hereafter naturalists and anti-naturalists retreat to their own isolated social circles, in which they have their separate philosophical societies, associations, conferences, technical lingos, peer-reviewed journals, periodicals, mail-lists, bulletins, heroes, philosophical emphases, textbooks, writing styles, professorial chairs and so forth. Are such visions realistic?

I find it pointless and premature at this point to take a definite stand on these visions as a sociological narrative about what is currently happening – or going to happen – within the profession of philosophers worldwide. Only time will tell. If I were a betting man, I would find it more plausible that the total plurality of conceptions of philosophy will continue to expand in the future driven by the process of specialization and the snowballing number of new philosophical research topics (the problems and topics targeted by philosophers are becoming more and more meticulously defined, which partitions the field of philosophy into increasingly smaller ecological niches). It is much more conceivable that both analytic and continental philosophy will continue to evolve on their current paths

and thus become more and more internally diverse, so that the bridge-building between the sects and sub-groups they contain becomes more and more arduous. With this in mind it is doubtful that any strong *esprit de corps* can form among philosophers regarding metaphilosophical naturalism or anti-naturalism – their attention is focused on smaller issues.

There is also a deeper question of whether each and every conception of philosophy is really classifiable as an instance of either naturalism or anti-naturalism, so that this juxtaposition could really be used to organize the philosophical landscape into two opposing halves. Is every philosopher, individual article, book, conference presentation and so forth really a representative of either naturalism or anti-naturalism as if this topic is a defining either-or touch stone for all philosophy? It might be reasonable to leave room for a third category called *non-naturalism*. This option collects under its umbrella all those conceptions of philosophy for whom the fate of naturalism is practically insignificant – it does not affect these philosophical activities one way or another. This might happen when the aims of philosophy – such as igniting societal change, authoring aphoristic literature assessed on aesthetical criteria or striving for the personal growth and wellbeing of the practitioner of philosophy herself – do not really come in touch with the empirical special sciences. It would also seem that different areas of philosophy differ in how typically they trigger metaphilosophical reflection on naturalism. For example, epistemology and philosophy of mind, which are naturally related to the functioning of human cognition, are fertile ground for discussions on how psychological research might inform our philosophical views. In contrast, my own current area of specialization, philosophy of education, for the most part skips this topic entirely (Pettersson, 2022; see also Mumford, 2008). It is thus not a widely shared experience among philosophers that the debate on naturalism is somehow inevitable for them. It is of course understandable that in the writings of those philosophers, who have themselves taken passionate stances against or in favor of naturalism, this is one of the most pressing metaphilosophical topics of our age. But it is quite another issue to what extent this debate really is present in contemporary philosophy and what its actual take away messages for actual philosophizing are.

Marathon runners often report the feeling of “hitting the wall” after about 3/4 of the way into the race (this sudden feeling of fatigue is caused by the depletion of glycogen stores in the liver and muscles). I hope the reader is not experiencing similar struggles as we now enter the final stretch of this paper. If it

helps, the finish line is already in sight. Next, I move to consider the question of whether either naturalism or anti-naturalism can be considered as a default or traditional view of philosophy's nature.

4. Is anti-naturalism a default or traditional view of the nature of philosophy?

At the level of superficial impressions, philosophy appears to be a non-empirical endeavor. Instead of relying on careful experimental designs and sets of empirical data, philosophizing typically has the outward appearance of deeply cerebral craft – sometimes the humorous description “armchair philosophy” is used to paint the picture (Rorty, 1979). Within academia philosophy seems to best resemble in this respect fields such as mathematics, rather than the experimental special sciences. Moreover, philosophizing today is not that different from the activities carried out by figures such as Plato, Descartes and Kant – which tells us something about the fact that the nature of philosophy has not radically changed over the centuries with breakthroughs in research technology in the same way as such innovations have driven progress in the natural sciences (for example, the invention of the microscope). Indeed, the non-empirical methodology is often highlighted as one of the distinctive traits of philosophy so that deviating from this feature would amount to already stepping beyond the bounds of philosophy and thus practicing some other kind of an inquiry (Glock, 2008, 96). The idea of “empirical philosophy” even sounds like an oxymoron (what could such a thing really be?).²

Against this backdrop, metaphilosophical anti-naturalism – in one form or another – can now appear like some kind of a default or orthodox conception of philosophy, which captures philosophy's true soul and distinctive point of view. Occasionally, one can even hear talk of a traditional conception of philosophy as well, which the “anti-traditional” and “post-philosophical” naturalists have newly challenged in revolutionary spirit only from the mid-twentieth century onwards (Bealer, 1996, 1). In this narrative, embracing naturalism is tantamount to eliminating philosophy as we have come to know it from the academic map.

It must be granted that there is some intuitive pull in this line of thinking. Any revolutionary metaphilosophical position, which radically redefines the

² As a candidate we can suggest the movement of experimental philosophy, which has gained popularity during the past few decades and introduced methods such as questionnaires and statistical data-analyses to philosophy (see Knobe & Nichols, 2007).

nature and purpose of philosophizing thoroughly, can be seen as already changing the subject altogether: what connection does this brand-new activity really have to the philosophy of old to warrant the continued use of this name? Does the acceptance of this kind of conception of philosophy add up to killing and replacing philosophy with some other kind of new activity?

There are several possible retorts to the argument above. One is to say that naturalism is not in fact as revolutionary a view in the history of philosophy as it is often made to be. In their own view, naturalists want to preserve continuity with certain pre-existing metaphilosophical strands present in the history of Western philosophy, and not instigate a radical or revolutionary movement, where the identity of philosophy is completely re-thought in contrast to the practices of earlier lovers of wisdom. Naturalists see their metaphilosophical view in a manner where it corresponds to the best and lasting ideals of good philosophy, which have been behind philosophy long before the constituent theses of metaphilosophical naturalism were first explicated in the philosophical literature or before the specific name of “naturalism” was determinedly used for these views. Kornblith (1994, 49) writes on this issue: “The idea that philosophy must somehow be grounded in the sciences is not new, and indeed, has given rise to an extraordinarily set of philosophical ideas. Descartes, Locke, Leibniz, Kant, Marx, Reichenbach, and numerous others sought to show that their ideas comported well with the best available sciences of their times.”

Secondly, naturalism is not (necessarily) a very revolutionary position in terms of philosophical methodology either. In the images painted by its discontents, naturalism is sometimes portrayed as a crude form of *scientific imperialism*, against which it would be wisest for loyalist philosophers to rise in a unified front to secure the future existence of their discipline and to defend the millennia-old autocracy philosophy has (allegedly) traditionally enjoyed. It must be admitted, however, that when we look for possible origins of these misunderstandings, naturalists themselves are at least partly to blame here, since they often use needlessly provocative rhetoric in expressing their views, which then motivates the image of the alleged revolutionary spirit of naturalism (although it is also true that certain naturalists do harbor these kinds of strong reformist ideas but, in this regard, they constitute the more radical wing of naturalism, so that we should not make any straightforward generalizations regarding all naturalistic conceptions of philosophy based on their specific metaphilosophical views).

Naturalistic philosophy can also trigger a related suspicion, that this kind of conception of philosophy demotes philosophy only to a secondary role of “synthesizing science journalism”, where philosophers create simple summaries of the most recent achievements of contemporary science and explain how they might pertain to philosophical matters. Blackburn (2004, xv), for example, writes provokingly that philosophers who support naturalism are merely “cheerleaders” in the scientific world without having their own truly constructive effort to give to these scientific undertakings.

These threats of scientific imperialism, however, are highly spurious. The central idea of naturalists is not “anti-philosophically” to simply delegate philosophical questions hereafter to the desks of scientists, while philosophers themselves become out of work as they have nothing more to contribute to the academic world. Instead, philosophy can still have its own fruitful role on a more ambitious level beyond mere “science journalism” and “cheerleading”. Moreover, naturalists do not mean even in a somewhat less radical mindset that the ideal philosophizing should from now on simply imitate – poorly – in its methodological procedures and in other practices the functioning of some empirical discipline or another (such as using experimental designs in direct testing of philosophical hypotheses). Naturalists, too, can think that empirical methods are not always a miraculous one-size-fits-all solution to every problem we have. The notion of the unity of philosophy and the sciences, to which naturalists adhere, is based on the qualitative similarity there exists deep down at the theoretical level between the cornerstones of these enterprises (as opposed to mere unsophisticated methodological monism or crude scientism): the crucial thing here is that the fundamental answers we give to the questions regarding the epistemological and ontological issues pertaining to the nature and subject matter of philosophical knowledge do not refer to any *sui generis* phenomena, viewpoints, sources of evidence or ultimate aims, which are not already present (or, could potentially be present) in the philosophical explanations given of the inner mechanisms of the sciences. Therefore, even in the cases where philosophers ruminate on philosophical topics with their philosophical methods, naturalists take these investigations to be conducted in the “overall spirit” of the empirical sciences and within a jointly shared “meta-scientific” framework with the sciences. Outside of this framework there is no alternative path to knowledge or enlightened source of “higher wisdom”, which philosophers could employ to approach their topics. According to one crystallization, naturalism means simply committing to a rather

moderate thesis, which states that when the natural sciences speak in unison in favor of some views, philosophers should either accept these conclusions or present scientifically compelling reasoning for their rejection (Burgess & Rosen 1997, 65). In this regard then, naturalism is not necessarily a particularly revolutionary metaphilosophical position, which would require us to undertake massive changes to the established practices of philosophy (naturalists can, however, have constructive suggestions how philosophy could be done even better from now on, but even then the initial template for these modifications is the present way of practicing philosophy, and not something completely different).

It is possible to regard philosophy as a unique discipline also in the metaphilosophical model of naturalists, so that when we compare it with typical empirical sciences, philosophy is a general endeavor in terms of its viewpoints and methods. Philosophy does not have a particular subject matter of its own, but rather operates on a higher level of abstraction compared to the empirical special sciences. If we return here to the above naturalistic metaphor of the web formed by all scientific knowledge, we can note that philosophy's designated sector in this web of knowledge is to concentrate on the central nexus-points interwoven close to the heart of the web, as it is these nexus-points which govern the way in which the web is built outwards from the center, whereas special sciences work at the outer edges of the web dealing with more hands-on issues pertaining to the way in which the web is connected to empirical reality. Of course, on this issue we can recognize more nuanced variation between diverse philosophical topics, so that a certain part of philosophy can be more "theoretical" in its viewpoints and thus lie closer to center of the web, whereas some other philosophical problems are conversely in closer vicinity to the outer edges of the web (similar variance in the level of abstractness exists between the research questions of empirical sciences too, since not all scientific research is about systematically testing the empirical hypotheses suggested by theories in a laboratory environment surrounded by test tubes and expensive pieces of research equipment – there is actually much propaedeutic conceptual and speculative investigation done also within the various empirical sciences). In any case, the crucial thing for all philosophers, no matter what they are working on, is to maintain a constant impression on how the research problems they investigate connect in the larger picture to the other research in proximity.

5. Concluding thought

In this text I have covered some introductory issues concerning metaphilosophical naturalism. In the light of these rough answers, naturalism may seem an unproblematic view. However, there are many issues outside my scope that could better expose the big open questions concerning this position. How do naturalists ultimately understand the scientific method? How do formal sciences such as mathematics and logic fit into this picture? What about the humanities, such as history? How is normativity explained in the naturalistic picture? These are just some of the big questions that I have excluded from this overview. Interested readers can continue to explore the relevant literature.

I began this essay by comparing philosophy with running a marathon. In truth, it takes a very special kind of person to master both sports and to do it with style. With these words, I wish Professor Valtteri Arstila a happy birthday and happiness for the many more years (and miles) ahead!³

University of Oulu

References

- Arstila, V. (2005), *The Paradox of Colors*, Reports from the Department of Philosophy, Vol. 15, University of Turku, Turku.
- Baggini, J. & Stangroom, J. (eds.) (2002), *New British Philosophy – the interviews*, Routledge, London.
- Bealer, G. (1996), "On the Possibility of Philosophical Knowledge", *Philosophical Perspectives*, Vol. 10, 1–34.
- Bealer, G. (1998), "Intuition and the Autonomy of Philosophy", in M. DePaul and W. Ramsey (1998), 201–240.
- Blackburn, S. (2004), "Foreword", in H. Carel and D. Gamez (2004), xiii–xviii.
- Burgess, J. & Rosen, G. (1997), *A Subject with No Object*, Oxford University Press, Oxford.
- Cappelen, H., Gendler, T. & Hawthorne, J. (2016), "Preface", in H. Cappelen, T. Gendler and J. Hawthorne, (eds.) (2016), v–vi.
- Cappelen, H., Gendler, T. & Hawthorne, J. (eds.) (2016), *The Oxford Handbook of Philosophical Methodology*, Oxford University Press, Oxford.
- Carel, H. & Gamez, D. (eds.) (2004), *What is Philosophy?*, Continuum Publishing King's Lynn, UK.

³ I thank the editors of this volume for the possibility to be included here. Moreover, I thank Alexander Carruth for revising my English.

- DePaul, M. & Ramsey, W. (eds.) (1998), *Rethinking Intuition – the Psychology of Intuition and Its Role in Philosophical Inquiry*, Rowman & Littlefield, Lanham.
- Glendinning, S. (2002), "The Analytic and the Continental", in J. Baggini and J. Stangroom (2002), 201–218.
- Glock, H-J. (2008), *What is Analytic Philosophy?*, Oxford University Press, Oxford.
- Knobe, J. & Nichols, S. (2007), "An Experimental Philosophy Manifesto", in J. Knobe and S. Nichols (2007), 3–14.
- Knobe, J. & Nichols, S. (eds.) (2007), *Experimental Philosophy*, Oxford University Press, Oxford.
- Kornblith, H. (1994), "Naturalism: Both Metaphysical and Epistemological", *Midwest Studies in Philosophy*, Vol. 19, No. 1, 39–52.
- Mumford, S. (2008), "Metaphysics", in S. Psillos & M. Card (2008), 26–35.
- Overgaard, S., Gilbert, P. & Burwood, S. (2013), *An Introduction to Metaphilosophy*, Cambridge University Press, Cambridge.
- Pettersson, H. (2002), "De-Idealising the Educational Ideal of Critical Thinking", *Theory and Research in Education*, Vol. 18, No. 3, 322–338.
- Psillos, S. & Card, M. (eds.) (2008), *The Routledge Companion to the Philosophy of Science*, Routledge, London.
- Rorty, R. (1979), *Philosophy and the Mirror of Nature*, Blackwell, Oxford.
- Williamson, T. (2008), *The Philosophy of Philosophy*, Blackwell, Malden MA.

Understanding, Charity, and Interpretation ¹

DAVID ROSENTHAL

To determine whether one disagrees with somebody, one must have a sound understanding of what that person is saying. But it is sometimes difficult to distinguish cases of misunderstanding others from cases of substantive disagreement. One can best resolve that quandary by making every effort to understand accurately, and that requires construing what the other person says as charitably as possible. But such charity is more complicated than often recognized, since it applies in several ways: in respect of how others use individual words, what assertions they make, and what inferences they regard as acceptable. And applying charity must rely on how one sees the relevant things; charity is unavoidably by one's own lights. An additional type of charity applies to the background explanatory pictures that sometimes figure in what people say. Construing in respect of one factor can affect how we construe in respect of another. So tradeoffs in how one construes the several factors are sometimes necessary, which can even lead to equally good alternative construals. Charitable construal of our own thoughts also occurs, often unconsciously, in resolving difficulties and, more generally, in guiding and regulating our thought processes.

1. Misunderstanding and Disagreement

We typically think about verbal communication in terms of the everyday conversations we have with one another. Such conversations typically flow smoothly and effortlessly, and rarely require or involve conscious reflection and

¹ Extensively revised from a presentation to the Hamilton College Humanities Forum on Translation and Cultural Exchange, October 17, 2011.

what to say. So we often pay relatively little attention to the difficulties or failures in understanding one another that can occur in everyday conversation.

Still, it is not all that unusual for mutual misunderstanding to occur in daily conversation. Such misunderstanding may well just pass unnoticed, since the main desideratum in such conversations is often successful social interaction, and that encourages the parties simply to ignore misunderstandings that do not matter in some special way. It can be easier and more comfortable to let minor misunderstandings pass and continue as though mutual comprehension is unimpaired. The parties will often just move on when they find what somebody said to be unclear or possibly mistaken, again unless getting things right matters for some reason. In these ways we tend to treat conversing as largely effortless and transparent.

Getting things right does often matter when people are discussing something that they especially care about, and then potential misunderstanding or disagreement can compel attention. This often happens with political issues or with matters of moral or aesthetic judgment. But it may also happen just with everyday decisions, for example, about what to do. And disagreements occur strikingly often in academic exchanges, especially in the humanities and social sciences, where there may seem to be fewer accepted ways to settle an issue.

But it can sometimes be difficult, in any of these cases, to tell whether people are disagreeing about something or just failing to understand one another. It might seem to one that another person would readily agree with one if only they understood what one is saying. But it can also be tempting to see somebody as disagreeing with one in a substantive way, even when a more careful and sympathetic exploration would reveal that the other person is simply using words in a relevantly different way.

And no matter how thoughtful one is about such things, it is not always all that easy or straightforward to distinguish accurately whether somebody disagrees with one or instead misunderstands what one is saying. That is so not only with relatively abstract matters of politics, aesthetics, or morality, but also with relatively mundane matters. It can even happen that one person sees misunderstanding where the other sees substantive disagreement.

The difficulty in distinguishing disagreement from misunderstanding in interpreting somebody else's written work, say, in a political or legal or academic context, can be especially intractable because then there is no back-and-forth exchange that could help resolve things. The frustration that arises from reading

something that seems wrong can impede any effort to understand the other person's remarks as generously as one might when one actively interacts with somebody. So there is a tendency to see substantive disagreement in the written work of others. It is somewhat unusual that commentary on written work explores the possibility that something that seems wrong might be due simply to the author's way of putting things is different from one's own.

Looked at from the perspective of somebody not engaged in a commentary or other exchange, this can seem unfortunate. When viewed by somebody who is independent and reasonably knowledgeable about the matter at hand, it may well seem that avoiding misunderstanding would not be all that difficult if the people involved were not so dug in on construing any apparent difference as due to a failure on the part of the other person.

2. Identifying Misunderstanding

How should we go about distinguishing misunderstanding and disagreement? When two people seem to differ in what they are saying, how can we determine, in both mundane and more controversial contexts, whether that difference is due to their not understanding one another or, instead, to their actually disagreeing with one another? These questions arise both for one looking at an exchange from the outside and for those who are party to the exchange. And the difficulty can be challenging. Indeed, the interplay of misunderstanding and disagreement may even make it seem that people agree when they do not.

It is especially important to resist jumping to any conclusion about disagreement vs. misunderstanding, since irrelevant factors can bias one in one direction or the other. One natural approach would be to see if one of the two possibilities can be eliminated, thereby establishing the other. This can happen with patient continued conversation, which sometimes shows that the appearance of disagreement was illusory, and that the apparent difference was solely due to misunderstanding. But if success in dispelling disagreement remains elusive despite sustained effort, it may be natural to conclude that the difference does reflect genuine disagreement.

But any such conclusions will be reliable only to the extent that there is a good-faith effort to avoid misunderstanding. And that can be more complicated and challenging than might sometimes appear. A serious source of difficulty is that misunderstanding can result from several different factors. Perhaps the most

straightforward and obvious factor is that people sometimes use words in ways that differ, and it is not always transparent when that happens. The ways the parties use the relevant words may overlap enough to conceal the operative differences.

But though differences in the way words are used may well be the most salient source of misunderstanding, it is by no means the only potential source. People may use the individual words in the same way, but still assign different truth conditions to various sentences that contain those words. The two parties differ instead about what would make various sentences true. And that would lead each party to misconstrue the assertions that the other party makes in using such sentences.

The assignment of truth conditions not only affects speech acts of assertion, but also affects all the other types of speech act so brilliantly studied by J. L. Austin (1961/1979), including but not limited to questions, requests, and expressions of doubt or wonder. All these require interpretation that rely on the truth conditions that a speaker associates with the corresponding assertoric sentences.

It may seem that assigning different truth conditions to the same assertoric sentences would be somewhat unusual. But it very likely happens a lot more than we recognize. Many cases are doubtless simply written off impatiently as substantive disagreements, rather than as a semantic divergence about how to evaluate assertoric sentences for truth. Differences in the way words are used are a lot easier to detect than semantic differences in the assignment of truth values. But they can be crucial in revealing and disarming cases of genuine misunderstanding

Semantic differences in the use of individual words and in connection with assigning truth conditions to sentences do not exhaust the potential for causing misunderstanding. Assertions are often linked together inferentially, and people can differ about which inferential connections they accept. And those differences can be altogether independent of how the parties use individual words or assign truth conditions to sentences.

Differences about which inferential connections the parties accept can be more difficult to detect than the other two, because it typically emerges only when one person offers reasons for some belief, relying on some inference, and the other person rejects that inference and so denies the cogency of those reasons. But however difficult to detect, differences about whether to regard an inference as acceptable can readily result in mutual misunderstanding.

So misunderstanding can result from differences of three distinct types, about the way the parties use individual words, about how they assign truth conditions to sentences, and about how they evaluate the acceptability of inferences. And these three ways in which people's use of language can differ will often interact in ways that make identifying each more difficult and less accurate, thereby amplifying the risk of mutual misunderstanding.

3. Construing Others

I'll return to these three sources of misunderstanding in section IV, and an additional source in section V. But it will be useful in the meantime to highlight the crucial tie between understanding, on the one hand, and interpretation and translation, on the other. When I say something and you grasp the meaning of what I say, you in effect make what I say your own in a particular way.

It is not that you thereby endorse what I say or even think it is sensible. But you cast what I say in your own terms, using words in ways that you would in talking about the matters under consideration. You in effect translate what I say into words that are natural to you. Sometimes the words you use to render what I say will be exactly the same as the words I used. But they need not be; indirect discourse often uses words different from those that would figure in a direct-discourse quotation.

All understanding of what somebody else says, even in the most basic and mundane cases of communication, is in this way unavoidably a matter of translation of what that person says into one's own terms. One cannot grasp what somebody else says except in terms that are natural and make reasonable sense to one. Understanding what somebody else says is interpreting what that person says, by putting it into words one could oneself use to talk about the matter at hand. It is a type of translation, even if only natural language is involved. We can think of it as translation from one idiolect into another.

Because we typically think of translation as being from one natural language to another, it typically goes unnoticed that we must translate from one idiolect to another when both belong to a single natural language. But all mutual communication and mutual comprehension consists in translation between idiolects, whatever the natural language. And all failures of understanding are failures of such translation between idiolects.

The only anchor one has about what words mean, what the truth conditions of sentences are, and what inferences are acceptable is our own usage in respect of words, assertions, and inferences. When things go smoothly and there are no differences to be resolved, we each tacitly take for granted that others operate with word meaning, truth conditions, and inference validity exactly as we ourselves do.

But that assumption is reasonable and appropriate only so long as there are no differences in what the parties say. When differences do arise, so that we seem to disagree or misunderstand the other person, we must do our best to render what that person says in terms we ourselves would comfortably use and readily understand. There is no other way. We cannot interpret what others say except by casting what they say in our own terms. The only way to understand what others have in mind by what they say is to find the most sympathetic and successful translation of what they say into terms we find natural enough to be comprehensible.

Our own use of words, assignments of truth conditions, and acceptance of inferences are the inevitable anchors for grasping what others say. That does not undermine at all the thoroughgoing social nature of our use of language. The way we use words, assign truth conditions, and assess inferences are all deeply influenced by how others use language. Our linguistic interactions with others form the dispositions we have in respect of those three aspects of how we ourselves use language. And the way that interacting with others effects our linguistic dispositions will interact with the way one construes others and the way other seem to construe us. Because of all that, we end up using words, assigning truth conditions, and assessing inferences in ways that overlap very heavily with the way others do.

Still, those ways of using language will rarely if ever be exactly the same from one individual to another. The linguistic dispositions each of us forms by interacting with others will always be to some degree idiosyncratic, leading to the differences that obtain among idiolects. And if the resulting idiolectal idiosyncrasies are not recognized and adjusted for, they can readily result in failures of mutual understanding.

When people have fluent command of the same natural language and engage in casual conversation, it will typically seem that they do not differ in their use of words and the way they treat sentences and inferences. They easily and naturally react to one another's remarks, saying what comes to mind.

But when differences that matter in some way do occur, so that there is reason to assess whether the difference is due to misunderstanding or substantive disagreement, each party must interpret the other into their own words. And all anybody in the end has to go on in is one's degree of success in rendering others' speech into words we would use ourselves and they will take to be accurate.

4. Charity of Translation

To maximize understanding what another person says, one must construe how that person uses words, assigns truth conditions, and evaluates inferences. These three factors reflect three units of meaning that operate our thinking: term-sized units, sentence-sized units, and inference-sized units. Maximizing success in understanding somebody requires construing that person's remarks so that those three factors differ as little as possible, ideally so little as not to be noticeable by either party.

Any such construal is inevitably rendered in one's own terms. How then does one seek to minimize divergence? To understand the other person, we want to do maximum justice to what that person is saying and has in mind. But since one cannot get directly inside the other person's mind, what can guide the way one construes the other's remarks? The considerations discussed in the last two sections will help.

Consider first the term-sized factor of now the other person is using individual words. We all tend to use words in ways we believe conform to standard norms for those words. So it is natural to construe another person's use of words as reflecting such norms. But it can sometimes be that we can make the best sense of what somebody says by seeing that person as using words in a somewhat different way. And we can seek to maximize successful understanding by ascribing to the other person the use of words that seems to make the best sense of what that person says. We can in this way do our best to put ourselves in the other person's communicative shoes in respect of the way that person uses words.

But how does one determine what makes the best sense of the other person's assertions? The only way is what seems to one to make the best sense. One must rely on what makes the best sense by one's own lights. There is nothing else one can go on. We must assume that the other person intends to use words in a way that makes the best sense. So it is charitable to construe that person's use of words in that way. But because we cannot get into the other person's thought processes,

our only access to what makes the best sense is in the end how we see the relevant matters.

Often the best way to construe another person's words will be the same as the way one would oneself use those words, what W. V. Quine describes as "the normal tacit method of homophonic translation" (1960, p. 59). But when it seems that what the person says makes better sense if some relevant words are construed differently, we should go with that construal. This constitutes charity for the construing of the other person's words.

Much the same points hold for how we can best assign truth conditions to the sentences another person uses in making assertions. We should assume that people intend to use words in a way that makes good sense. So we can best construe somebody's words in whatever way makes the best sense of what they say. Similarly, we should assume that people also intend that their assertions are true. Charity of construal will require that we take the assertoric sentences others use to be true as often as we reasonably can.

Again, that is truth as one sees the relevant matters. One should, as much as possible, construe others' assertions as being true by one's own lights. One has in the end no access to what is true apart from one's own best estimate about what is true. There is no other standard by means of which one can construe as true the sentences others use to make assertions.

The tendency to construe as true the sentences others use to make assertions is reasonably natural and automatic. When somebody says something that strikes us as wrong, we often think, "What that person must really mean is such-and-such." We in effect slot in truth conditions for the other person's sentences that would make the relevant assertions true by our lights.

This natural urge to maximize the truth of what other say is often referred to as the principle of charity (Wilson 1959; Quine 1960; Davidson 1973-1974). But it is only one of charitable construal. Construing another person's use of words as making the best sense of statements that contain those words is another, term-sized form of charitable construal.

And things can get complicated with two distinct ways of construing charitably, one pertaining to term-sized units of meaning and the other to sentence-sized units. When another person's assertions do not seem true to one, one can adjust the meanings of some constituent words or the truth conditions one assigns to that person's sentences, or even both. Sometimes it will be obvious what the most charitable adjustment is, but in many cases it may not be. One has then simply

to do one's best with the two available types of adjustment, and be sensitive to the tradeoffs that can arise between adjusting for word meaning and for truth conditions.

The complications in implementing an overall charitable construal increase dramatically when we take into consideration the reasons people give for saying things. Such reasons rely on inferences from reasons to the statements for which they are reasons. So we must then add a principle of charity for the construing of others' inferences. We must not only construe words so as to make the best sense of what others say and truth conditions so as to maximize the truth of others' assertions; we must also construe the inferences they make as being acceptable as often as possible. And again, our only measure of acceptability for inferences will rely on the way we see things.

We must accordingly figure out how to balance all three types of charitable construal. Inferences can be good even if not all their constituent sentences are true. So maximizing acceptability of inferences and maximizing truth of assertions are independent. And it may be easiest to see a person's inference as acceptable if one takes some constituent sentence not to be true, or the sentences as all true if one sees the inference as faulty. Indeed, that is often what happens when others seem to agree with us for what we think are the wrong reasons.

Charitable construal is often taken to be a matter of maximizing simply the truth of another's assertions or simply the acceptability of the other person's inferences. Apart from leaving out the need to construe meanings of the other's words, any one-factor version of charitable translation inevitably overlooks the need to balance the three types of charity so as to produce most charitable overall version of what the other person says.

Indeed, the need to balance those three factors has to my knowledge not been explicitly noted in the literature. But once one notes that charity must apply in these three independent ways, it is obvious that tradeoffs between the three will sometimes be necessary. This makes it crucial also to recognize that charity does apply in all those ways.

When it does seem necessary to strike balance among the three, moreover, there may be alternative ways to do so that are equally charitable, so that there is no single best construal. It can happen that we get a good overall construal by taking all the assertions to be true but sacrificing the acceptability of some inferences, but that we get an equally charitable overall construal by saving validity and giving up some truth.

If there are alternative equally charitable construals, continued conversation or written exchange may point to one overall construal that is preferable to all the parties concerned. But it can also turn out that such additional interaction does not suggest a single best construal.

The occurrence of alternative best construals may sometimes explain how there can be equally compelling but incompatible readings of texts, something familiar to anybody who has worked extensively in the history of philosophy (Rosenthal 1989, §4). And it can likely also explain incompatible readings of arguments and positions in academic, political, or legal contexts that one cannot decisively adjudicate among.

Relying on charitable construal is at bottom seeking to do the greatest justice possible to the other person's remarks. This idea is by no means new, and arguably has strong roots in antiquity. John Stuart Mill's famous remark that "[h]e who knows only his own side of the case, knows little of that" (1859/1991, 42) is in effect an injunction to maximize charity in construing the views of those with whom one disagrees. And Mill traces this approach to Cicero, presumably because Cicero, like other Academic skeptics, held that one should be able to argue either side of any issue, doing justice to each (Remer 2022, esp. p. 212; Brittain and Osorio 2022). Mill does not note, nor to my knowledge does Cicero, that this requires balancing the charitable construal of words, sentences, and inferences. But the foregoing considerations make it plain that it must.

Because charitable construal must rely on how the person doing the construing sees things, the interpretation of historical texts, whether in philosophy, law, politics, or other matters, will always in that way be "Whiggish," to use Richard Rorty's useful term (1979; see also Rosenthal 1989, §§3-4). Such Whiggish interpretation in no way suggests or warrants simply reading one's own views into historical texts. It is only that the way we understand historical texts unavoidably relies on what sentences we take to be true and what inferences valid, and our use of individual words.

It is sometimes urged that interpreting historical texts requires extensive immersion in the writings of the relevant period and tradition. Such immersion can indeed be useful, and perhaps sometimes indispensable. But such immersion requires one to interpret everything one reads in the relevant corpus. And it unavoidably relies on charity in respect of word meaning, truth, and inference. Immersion in a historical period must itself proceed by charitable interpretation.

5. Charity and Background Theories

The three forms of charity described so far are characterized by the semantic unit they apply to, words, sentences, and inferences, respectively. But there is a fourth factor that can figure in the way we understand what other say, and which calls for a type of charitable interpretation. This fourth way pertains to the general beliefs people have, often somewhat theoretical in nature, which sometimes inform and underlie what people say and think. People's assertions and inferences often reflect some theoretical picture about the phenomena under discussion, and then one may not accurately understand what they say without taking into consideration that theoretical picture.

This requires a fourth type of charity in interpreting what others say. We must determine what background theoretical picture is most likely operative, relying on which background picture makes the best sense of what people say. Charity here may be more exploratory and speculative than with the use of words, truth conditions of sentences, and validity of inferences. In those three cases, people explicitly use particular words and sentences, and typically make their inferences explicit as well.

By contrast, the background picture that informs their statements and inferences often is not explicitly stated. In that case one must be sensitive to whether some theoretical picture is operative, and if so what it is. It will seem sometimes that the most charitable construals of words, sentences, and inferences do not by themselves make good sense of what somebody says, and that can suggest that the person has some tacit background picture in mind.

One can then charitably consider likely hypotheses about what picture would make the best sense of what the other person is saying. A person's remarks will sometimes clearly suggest some underlying theoretical picture, and sometimes an explanatory picture will be front and center in what the person says. But often it will be only tacit. And in those cases, charity can provide a strong and useful guide about what picture is operative. One ascribes the tacit background picture that makes the best sense of what is said.

Charity here as always inevitably rests on how the person doing the interpretation sees things. One must ascribe whatever background theoretical picture seems to one to make the best sense of the other person's remarks. One might not endorse the background picture one ascribes to the other person; indeed, it may well be a picture that one roundly rejects. But it must seem to one to make

the best sense of what the other person says. And once again, there may be tradeoffs between the background picture one ascribes and the charitable construals of words, truth conditions, and inferences.

If one sees such background theoretical pictures as simply sets of assertions related by inferential connections, they would not require a distinct type of charitable interpretation. One could then simply be charitable about the other person's assertions and inferences. But there is more to such background pictures than simply the truth of a number of assertions that are bound by inferential ties. Such pictures also play a distinctive explanatory role in connection with the relevant phenomena. It is that explanatory role that unifies and individuates such background pictures. And it is such explanatory roles that figures in the charitable detection of what background picture makes the best sense of what another person is saying. It is in that way that appeal to such background picture calls for a distinct type of charitable interpretation.

An example may help. In describing others in folk-psychological terms, we take for granted the posits, assumptions, and explanatory framework of whatever folk-psychological picture one has. And the specifics of what is assumed in somebody's folk-psychological picture can vary from one person to another.

Some today might tacitly fold in neo-Freudian ideas that others would roundly reject. And some religious convictions might also affect one's folk-psychological picture, thereby making a difference to how one describes others folk-psychologically. In talking with somebody, one will sometimes recognize that they simply see how people operate folk-psychologically in ways different from our own take on those things. Charity in determining another person's take on the explanatory framework of folk psychology may then be needed for a correct construal of what that person is saying.

There are other examples. Some pertain to how we see the workings of social groups, and to how we think social groups should operate. Some people think things work best politically and socially when government has a minimal role in people's lives, whereas others believe that government and other organizations should provide support to people when needed. Some see political and social matters in a largely conservative way, interpreting things in distinctively traditional ways, whereas others interpret social and political matters in a more open-ended way to accommodate and even encourage change. Some tend to picture social interactions on the model of life in small towns, whereas others picture them in a large and diverse metropolitan environment. Failing to take

account charitably of such differences can impede an accurate understanding of what others say.

Yet other examples figure in academic contexts. Theoretical considerations can underlie what somebody says in an academic dispute, and understanding those background considerations can then be pivotal to an accurate grasp of what that person says. Doubtless that sometimes also occurs with the aesthetic and moral judgments people make. And it is said by many that people who grew up in different generations sometimes have a background picture distinctive of their generation, and which affects many things they say and think. In all such cases it can be important in fostering mutual understanding to take charitable account of differences in background pictures.

Interpreting what somebody says is inextricably bound up with interpreting whatever theoretical approach they have about the things under discussion. One often cannot do one without doing the other. This observation recalls the argument Quine calls pressing from above for the indeterminacy of translation, which involves interpreting some relevant theory held by the person being interpreted (1970, 183). And it is of a piece more generally with Quine's insistence that no serious line can be drawn between semantic and theoretical considerations (Quine 1951). Meaning and belief interact in ways that require that one construe the two together as a package.

6. Some Observations

As noted in section IV, the first three types of charity operate on distinct units of meaning, term sized, sentence sized, and inference sized. But the fourth can also be seen as pertaining to a distinctive unit of meaning, that which can sustain a fabric of theoretical explanations. Theorizing can be formulated inferentially, with postulates that lead to the phenomena to be expressed, but it need not be. It can instead simply reflect theoretical claims with explanatory connections to the relevant phenomena.

One can see the contrast between inferences and a more general set of explanatory connections in the way Plato distinguishes in Book VI of Republic between the type of intellectual activity he sees as occurring in connection with the third level of his divided line and that which is characteristic of the fourth. The third level involves inferences from hypotheses that are treated there as beginning points (511a-b). The fourth level, by contrast, embeds hypotheses in an

explanatory fabric that treats them not as beginning points, but as genuine hypotheses to be explored and evaluated: “truly as hypotheses—stepping stones to take off from” (Plato 1992, 511b5, p. 195). If we put aside Plato’s ontological commitments, we have a clear contrast between inference and a fabric of explanatory connections.

The structural analogy between types of charity and Plato’s four levels also holds for the first two forms of charity. We apprehend the things that occur at Plato’s second level by way of beliefs, which represent in a sentence-sized way. By contrast, the first level is populated by images (εἰκόνας; 510a1), which represent in a term-sized way (509d-510a).

The most common way to represent how words are used is with the lexical entries in a dictionary. But those entries are in effect abbreviations of sentences that say that the target word means what the lexical entry specifies. And any other way of specifying the use of individual words is likely also to involve sentences. It can be explicit, as in saying ‘The word ‘dog’ refers to domestic canines’ or even ‘Dogs are domestic canines’, or tacit, as in uttering ‘dog’ while pointing at a dog, which is a nonverbal way of saying that ‘dog’ refers to that type of thing.

But that does not affect the distinction between word-level charity and sentential charity. The sentences that specify how an individual word is used are readily distinguishable from sentences in which that word figures in saying something else. So we can apply charity about individual words independently of charity about truth conditions.

Charity of construal is a tool for maximizing mutual understanding. But mutual understanding may not be always one’s overriding goal in how one construes what somebody else says. In competitive exchanges, such as debates or legal proceedings, for example, the desire to score points can override the wise injunction of Mill and Cicero to arrive at an accurate understanding of what one’s opponent says.

There are, however, other more interesting and intellectually challenging ways in which one’s primary goal may not involve having an accurate understanding of others. Harold Bloom (1997) has advanced the striking and provocative idea that writers often tacitly, though also actively and creatively, misread those of the predecessors who most influence them. Bloom has mainly the literary world in mind, and in particular poetry. And he sees this process as driven in large measure by one’s feeling that those who most influence one have

already scaled the relevant artistic heights, leaving one unclear how one might make one's own mark. Hence Bloom's idea of the anxiety of influence.

Bloom sees literary figures as often in effect engaging in a kind of intellectual struggle with those who most influence them, which leads them to misread those predecessors, but in ways that can be highly compelling. Such agonistic but creative misreading of others in seeking to make one's own mark likely occurs in other intellectual enterprises as well. Academics in the humanities and social sciences often focus on aspects they reject in the writing of those who have most strongly influenced them, sometimes resulting in new and revealing approaches to the relevant topics.

As noted at the outset, when somebody says something that strikes one as wrong, it can be unclear whether to see oneself as misunderstanding what is said or as simply disagreeing in a substantive way. Bloom casts readers as simply settling agonistically for disagreement with what others write, especially those who most strongly influence one, and work off those disagreements to develop their own views.

But even in their writings of those who most deeply influence one, there is typically more than enough for one to disagree with. So even if one interprets those works charitably, resulting in an accurate understanding of them, there should be ample opportunity, as well as significant motivation, for one to elaborate one's own views. Bloom's creative misreading should not be necessary. And working off of real disagreements, as against those that one agonistically misreads, would likely result in developing views that are more stable and compelling. It may even be that doing so is a lot more common than Bloom imagines.

In ordinary conversation as well as in more formal interactions, we expect that others will construe us charitably in the ways described above. So if somebody were to respond to what one says in a way that suggests total lack of concern with charitable construal, it would seem at best strange, and likely leave one with a sense that something is significantly amiss.

But that is actually the way it often seems when in the exchanges one has with current artificial intelligence programs (AIs). The large-language models (LLMs) on which AIs run do generally produce responses that address the topic at hand, but those responses often do not seem specifically directed to what one has said. However relevant the machine's response may be to the subject matter, there is often an odd sense of disconnectedness. The machine does not seem actually to understand what one says, as against simply picking up on what topic one is talking

about. Even when the machine replies to a direct question, there is often something off about the response.

Given the way current AIs operate, that is to be expected. The gigantic libraries of texts that constitute LLMs are structured to generate new text that pertains to the topic raised by whatever input it receives. Things are moving fast, and there is now some hope that AIs can simulate some cases of human reasoning (e.g., Metz 2024).

But generating a sense of human communicative interaction is another matter. A machine could appear to have a genuine understanding of what a person says to it only if it can perform charitable assessments of the inputs it receives. That would require evaluating inputs for both truth and validity, and current LLMs do not have anything like the resources needed to do so. Without an ability to interpret inputs charitably, the machine's responses will seem detectably artificial in a significant way.

Alan Turing's classic article (1950) urged testing a computer program for intelligence by seeing whether people would take its responses to be those of another person. Turing was careful to specify that this test, now generally known as the Turing test, should be thought of as only testing whether a machine exhibits intelligence, and not also be whether it can think. Turing held that the question whether machines can think is "too meaningless to deserve discussion" (1950, p. 42; see Mitchell 2024).

But perhaps we can reduce skepticism somewhat about that latter question. Charitable construal may not be necessary in testing for intelligence, but it likely is pivotal in testing for thinking. So if a machine did generate responses to human inputs that led humans to see those responses as reflecting a charitable construal of those inputs, that might encourage a sense that the machine is thinking, or at least doing something very close.

Developing programs that could engage in charitable construal is by no means impossible. But it would not be trivial. The program would begin with need a database that enabled it to identify things commonly taken to be true and false. But it would also have to be able to evaluate commonsense inferences for validity. And doing so would require access to the connections among words that sustain commonsense inferences, which are relatively holistic in nature. Constructing a database for that would be no mean feat. So the appearance of thinking in machines, as against mere intelligence, is likely not around the corner.

7. Self-construal and Thought Processes

Charitable construal maximizes success in understanding what another person says. And since what somebody says expresses their thoughts, charitable construal enables one to grasp the other person's thoughts. What about one's own statements and thoughts? Does charitable construal have any application there?

We rarely have any need to construe our own individual words. If I say 'The bank is nearby' I know which meaning of 'bank' I have in mind. And there is typically no need to construe one's own speech acts. Every thought that is expressed in a speech act is conscious (Rosenthal 2005, ch. 10), and being conscious means that the thought appears in one's stream of consciousness (Rosenthal 2022). And knowing in that way what thought is expressed by what one says typically obviates any need for one to construe one's own speech acts.

But things are not always so straightforward with one's conscious thoughts. One can be confused about what one consciously thinks. And one may sometimes have a sense that one's conscious thoughts about a particular matter do not all fit together smoothly or that what one thinks about something is not quite right, even if one cannot see quite what the difficulty is. One's conscious thinking may also sometimes lead to a problem or dead end, which one cannot see how to resolve or get around.

It sometimes happens in such cases that taking a break or getting a good night's sleep results in resolving such difficulties even without one's giving the matter any further conscious thought. One will then not be aware of what mental process led to one's seeing clearly how to think about the relevant things; evidently unconscious thought processes enabled one to see how things fit together or to solve the problem. Often this leads to one's having thoughts that are different, even if only slightly, from the thoughts one had earlier.

How does an unconscious thought process work to resolve such difficulties? The unconscious thought process plainly must take into consideration the main thrust of one's thinking. One's general approach to the matter at hand does not change; it is simply that difficulties with affected that approach get resolved. So the unconscious thought process must be sensitive to most of the things one holds true in that connection, and most of the relevant inferences. And taking those things into consideration in effect constitutes operating charitably with the relevant thoughts.

Such charitable assessment of one's own earlier thinking is necessary if the unconscious thought processes are to resolve a difficulty with that earlier thinking, as against simply coming up with new ways of thinking about the matter at hand. To come up with a salutary resolution of whatever the difficulty was, the unconscious thought process must charitably assess how best to make those thoughts fit well together.

Such unconscious charitable construal of one's own thoughts would operate much as one charitably construes remarks made by somebody else. The charitable construal of one's own thinking operates on most of the relevant thoughts, but charitably modifies one or more beliefs or inferences in the light of the overall body of beliefs and inferences that one is committed to about the matter under consideration. Such charitable self-construal can in that way dispel confusion, resolve difficulties, and solve problems. And it results in one's in effect finding out, as one might put it, what one had really thought all along about the matter at hand.

Charitably construing others is unavoidably anchored in what assertions one sees as true and what inferences as acceptable. Similarly, imposing rationality on our own thinking, by dispelling confusion and resolving difficulties, must also rely on how we ourselves think about the issues under consideration. New thoughts can dispel confusion or resolve difficulties only if they fit comfortably with one's overall approach to those issues.

Charitable assessment of one's own thinking can also occur consciously. Sometimes that will simply be to get one's conscious thoughts in order and make their inferential connections more transparent. Other times conscious self-assessment may lead one to modify some of one's thoughts, to eliminate some belief or inference as mistaken, or to come up with new thoughts altogether.

Thinking always involves adjusting one's thoughts and coming up with new thoughts. So the charitable assessment of one's own thoughts figures not only in dispelling confusion and resolving difficulties, but more generally in guiding and regulating our thought processes. And our ability to regulate our own thinking by such charitable self-assessment likely develops hand in hand with our ability to interpret others. This is an important way in which our thinking is responsive to the interactions we have with others.

Self-construal of one's thoughts also occurs in connection with consciousness. For a thought to occur in one's stream of consciousness is for it to appear subjectively that one has that thought. So one must subjectively represent

oneself as having that thought, which is in effect to construe oneself as having it. But the charitable self-assessment of one's thoughts that occurs in resolving difficulties need not be conscious. For a thought to be conscious, one must explicitly represent oneself as having that thought, even if that representing does not itself occur consciously (Rosenthal 2005, chs 2-4).

But the self-interpretation that does occur when a thought is conscious introduces a second type of confusion that can occur in our thinking. That self-interpretation can itself go amiss, so that one is confused about what it is that one actually thinks. In such cases charitable self-construal can again help induce clarity about what one's thoughts actually are.

We take mutual understanding in conversation so much for granted that we seldom reflect on what such understanding consists in and what it requires. Without a solid grasp of how mutual understanding does operate, we can often be at a loss when misunderstanding occurs. The various mechanisms of charitable construal provide the needed theoretical framework. And those charitable operations are also fundamental to guiding the way we ourselves think.

City University of New York, USA

References

- Austin, J. L., *Philosophical Papers*, Oxford University Press, 1961/1979.
- Bloom, Harold (1997), *The Anxiety of Influence: A Theory of Poetry*, 2nd edn., Oxford: Oxford University Press.
- Brittain, Charles, and Peter Osorio (2022), "The Ciceronian Dialogue," in *The Cambridge Companion to Cicero's Philosophy*, ed. Jed W. Atkins and Thomas Bénatouïl, Cambridge: Cambridge University Press, 25-42.
- Davidson, Donald (1973-1974), "On the Very Idea of a Conceptual Scheme," *Proceedings and Addresses of the American Philosophical Association*, 47: 5-20.
- Metz, Cade (2024), "OpenAI Unveils New ChatGPT That Can Reason through Math and Science," *New York Times* (September 12), <https://www.nytimes.com/2024/09/12/technology/openai-chatgpt-math.html>
- Mill, John Stuart (1859/1991), *On Liberty*, in *On Liberty and Other Essays*, ed. John Gray, Oxford: Oxford University Press.
- Mitchell, Melanie (2024), "The Turing Test and Our Shifting Conceptions of Intelligence," *Science*, 385, 6710 (August 15): DOI: 10.1126/science.adq9356
- Plato (1992), *Republic*, tr. G. M. A. Grube, rev. C. D. C. Reeve, 2nd edn., Indianapolis: Hackett Publishing Company.
- Quine, W. V. (1951), "Two Dogmas of Empiricism," *The Philosophical Review*, 60, 1 (January): 20-43.
- Quine, W. V. (1960), *Word and Object*, Cambridge, Massachusetts: MIT Press.

- Quine, W. V. (1970), "On the Reasons for the Indeterminacy of Translation," *The Journal of Philosophy*, 67, 6 (March 26): 178-183.
- Remer, Garry (2022), "Philosophy, Rhetoric, and Politics," in *The Cambridge Companion to Cicero's Philosophy*, ed. Jed W. Atkins and Thomas Bénatouil, Cambridge: Cambridge University Press, 200-214.
- Rorty, Richard M. (1979), *Philosophy and the Mirror of Nature*, Princeton: Princeton University Press.
- Rosenthal, David (1989), "Philosophy and Its History," in *The Institution of Philosophy*, ed. Avner Cohen and Marcelo Dascal, Peru, Illinois: Open Court, pp. 141-176.
- Rosenthal, David (2005), *Consciousness and Mind*, Oxford: Clarendon Press.
- Rosenthal, David (2022), "Mental Appearance and Mental Reality," in *Qualitative Consciousness: Themes from the Philosophy of David Rosenthal*, ed. Josh Weisberg, Cambridge: Cambridge University Press, pp. 243-271.
- Turing, A. M. (1950), "Computing Machinery and Intelligence," *Mind*, 59, 236 (October): 433-460.
- Wilson, N. L. (1959), "Substances without Substrata," *The Review of Metaphysics*, 12, 4 (June): 521-539.

Can Computers Reason Like Medievals? Building ‘Formal Understanding’ into the Chinese Room¹

LASSI SAARIO-RAMSAY

What do the Middle Ages and modern computers have in common? Nothing, you may say: the Middle Ages are synonymous with darkness, stagnation, and ignorance, whereas computers are the crown jewels of enlightenment, progress, and innovation. I would reply that they actually have many things in common, one of which is their intimate relationship with logic, the study of valid reasoning—and yet the medieval conception of valid reasoning turns out to be a tough nut to crack for a computer, as we shall see.

The question of whether computers can distinguish valid from invalid arguments was first asked before there were any actual computers. The idea of implementing an algebra of logic on a computing machine had been envisioned as early as the 17th century by G. W. Leibniz (Davis, 2018, 5). In 1928, David Hilbert and Wilhelm Ackermann posed the so-called Decision Problem (*Entscheidungsproblem*), leading to the development of various models of computation that were then incarnated by actual computers as we know them. We can paraphrase this problem, given certain now obvious conventions and results, as follows:²

¹ This paper has undergone peer review.

² The problem, as it was originally formulated, was to find “a procedure that allows one to decide through finitely many operations whether a given logical expression is universally valid or, alternatively, satisfiable” (Hilbert and Ackermann, 1928, 73). We may equate such a procedure with an actual computer, given the so-called Church–Turing thesis (see Copeland, 2024) and the fact that all prominent models of computation coincide with what actual computers can do. A formula A is “universally valid” (i.e., a logical truth) if and only if the argument from the empty set of premises to A is formally valid, and an argument from A_1, A_2, \dots, A_n to C is formally valid (i.e., C is a logical consequence of A_1, A_2, \dots, A_n) if and only if the formula $A_1 \wedge A_2 \wedge \dots \wedge A_n \rightarrow C$ is universally valid.

(The Decision Problem) Is it possible for a computer to *decide* for any given argument in first-order predicate logic whether it is formally valid?

The problem was to find a ‘decision procedure’, or an ‘effective method’, that would take an argument of first-order predicate logic as its input, terminate after a finite number of steps, and produce the output ‘Yes’ if the argument is valid and ‘No’ if it is not. The corresponding problem about propositional logic had already been solved by the truth-table method, which allows one to decide the validity of any given argument in finitely many steps. It had also been proven that decision procedures exist for certain fragments of first-order predicate logic. However, in 1936, Alonzo Church and Alan Turing proved that such a procedure could not be found for full predicate logic and that the Decision Problem was thus unsolvable (Church, 1936; Turing, 1936). In spite of the many amazing things that computers are capable of, the full depth of valid reasoning is something that evades them.

This, of course, is not to say that computers would not be useful for inferential purposes. Even if the Decision Problem is unsolvable, we may find consolation in the existence of various complete proof systems that can be straightforwardly implemented on a computer. Although not all invalid arguments are refutable in these systems in a finite number of steps, all valid arguments *are* provable in these systems in a finite number of steps. We can simply write a computer program that takes an argument as its input and looks for a proof of the conclusion from the premises by systematically applying the rules of the chosen proof system. If the argument is invalid, there is no proof to be found and the execution of the program will never end, but if the argument is valid, there is a proof to be found, and the computer is guaranteed to find it eventually. We have thus solved a weakened version of the Decision Problem:

(The Recognition Problem) Is it possible for a computer to *recognize* any given formally valid argument in first-order predicate logic?

The solution of the Recognition Problem depends on *formalization*, namely, the fact that the conditions of valid inference can be formulated as formal rules that only consist in manipulating strings of symbols (the *syntax*) without regard to their meanings (the *semantics*). The computer can, for instance, be told that the rule of universal instantiation,

$$\frac{\forall x\varphi(x)}{\varphi(c)}$$

is valid regardless of what the open formula $\varphi(x)$ is about or what object the constant term c denotes. Thus a computer can, in principle, confirm that the argument:

- (1) Valtteri is a human being; every human being is an animal; therefore, Valtteri is an animal

is valid; it only has to be formalized into, e.g.,:

- (1') $H(v), \forall x(H(x) \rightarrow A(x));$ therefore, $A(v),$

and then the computer can search for a proof using the rules of the system. Nothing in this process requires that the computer ‘understand’ what the premises and the conclusion *say* or what the terms that occur in them *mean*.

In our everyday life, however, we are interested in all sorts of arguments, only a part of which can be validated in a formal system of the kind described above. Consider the following arguments, for example:

- (2) Valtteri is a human being; therefore, Valtteri is an animal.
(3) Turku is west of Helsinki; therefore, Helsinki is east of Turku.
(4) Daffy is black; therefore, Daffy is not white.

These arguments, as natural as they may seem, do not have a valid form in standard first-order predicate logic. Nowadays we are accustomed to classifying such arguments as merely *materially* valid and leaving them outside the scope of logic proper, but perhaps that is just a prejudice due to our received philosophy of logic (cf. Read, 1994; Halbach, 2020, n. 1). If only we could travel back to the Middle Ages, we would find that most logicians at that time would classify these kinds of arguments as *formally* valid because their conclusions are contained in the ‘formal understanding’ of their premises, or because the negations of their conclusions are ‘formally incompatible’ with their premises. Let us call arguments of this kind ‘conceptual-formally’ valid (to be specified shortly).

To be sure, I am not saying that we have to travel as far as the Middle Ages to find anything like conceptual-formal validity. It lives on in the notion of *analyticity*, traditionally attributed to Kant (cf. Sundholm, 2013), famously refuted by Quine (1951), and recently reconsidered by Boghossian and others (see e.g.,

Boghossian, 1996; Russell, 2008; Boghossian and Williamson, 2020). The related notion of ‘analytic validity’ or ‘analytic consequence’ is sporadically mentioned in the literature (e.g., Kment, 2006; Etchemendy, 2008; Read, 2015); a mechanism called “Analytic Consequence (Ana Con)” has actually been implemented on a computer, in a restricted and incomplete form, as part of the software accompanying the textbook *Language, Proof and Logic* (Barker-Plummer, Barwise and Etchemendy, 2011).³ Even so, it goes without saying that the medieval account of these issues is in many ways different from ours: our concern with artificially constructed languages, for one thing, is even more alien to the medievals than their Aristotelian worldview of definitions, accidents, and other ‘predicables’ is to us. Not only is a systematic study of the medieval account an interesting task in its own right, but it also casts a new light on our contemporary considerations on formal validity.

The problem that I pose and try to solve in this paper, then, is this:

(The Medieval Recognition Problem) Is it possible for a computer to recognize any given *conceptual*-formally valid argument?

As it turns out, this problem is considerably harder than the ordinary Recognition Problem. The notion of conceptual-formal validity, unlike that of ordinary validity, is inexact to begin with. In order to solve the Medieval Recognition Problem, we must first rephrase the problem in more precise terms, or rather, whether conceptual-formal validity can be precisely defined is part of the problem itself. I will begin by sketching a preliminary account of conceptual-formal validity within the framework of modern logic.

1. The Medieval Account of Conceptual-Formal Validity

14th-century theories of formal consequence (*consequentia formalis*) are by far the most remarkable precedent of our contemporary idea of logical consequence or formal validity (cf. Dutilh Novaes, 2020). These theories dethroned the Aristotelian syllogism, which had held pride of place among valid arguments for over a thousand years, in favour of a general account of valid arguments of all sorts. In this section, I will formulate a schematic definition of formal validity for

³ Something reminiscent of ‘conceptual-formal’ validity may also be found in contemporary information-theoretic logic (Corcoran, 1998), intensional logic (Fitting, 2022), and varieties of necessity (e.g., Plantinga, 1978; Fine, 2005).

an arbitrary first-order language, inspired by what I take to be some of the key tenets of the medieval theories.

The medievals typically first define what it is for an argument to be valid in the first place and then distinguish between formally and materially valid arguments. In our terminology, their definition of validity (consequence) may be stated as follows:

(Validity) An argument is valid if and only if it is necessarily the case that, if the premises are true, then the conclusion is also true.⁴

The definition clearly hinges on the sense of “necessarily”. In contrast to contemporary philosophers of logic who try to neutralize the appeal to necessity in one way or another, medieval authors typically presuppose a primitive notion of necessity and understand it in a very broad sense, including merely temporarily or ‘as of now’ (*ut nunc*) necessary cases that we would regard as contingent (see Read, 2012; Strobino, 2017). In the above definition, I presuppose a narrower notion of necessity that does not include such cases; in so doing, I am deliberately glossing over irrelevant aspects in the medieval sources to facilitate the systematization of the relevant ones.⁵

In order to define *formal* validity, we need an account of what the form of an argument amounts to. There were roughly two competing notions of form in the 14th century, which I call *structural form* and *conceptual form*.⁶ For our purposes, the structural form of an argument is simply what we get when we translate the argument into the language of predicate logic and abstract away from the ‘intended’ interpretation of the non-logical constants. Structural form is

⁴ Equivalently: an argument is valid if and only if it is *impossible* that the premises be true and the conclusion false. This definition is ultimately based on Aristotle’s definition of a syllogism (Aristotle, 2009, Ch. 1).

⁵ In other words, I restrict the class of valid consequences to what many medieval authors call ‘absolute’ or ‘simple’ (*simplex*, as opposed to *ut nunc*) consequences. This restriction is of no importance to us, given that we are solely concerned with formal consequences. Some medieval authors (e.g., Peter of Mantua) subsumed the formal vs. material distinction under simple consequences; others (e.g., John Buridan) subsumed the simple vs. as of now distinction under material consequences; either way, formal consequences turn out to be necessary in the narrow (simple) sense.

⁶ My distinction between structural and conceptual form is a systematization of passages found in Walter Burley (2000, 173), Adam Wodeham (1990, 82–83), and several subsequent authors in the 14th century. I have accounted for this distinction elsewhere (Saario-Ramsay, revised and resubmitted). Similar distinctions have previously been suggested by Klima (2016, 335) and Strobino (2017, 184).

located, as it were, at the level of the object language and involves no explanation of the non-logical constants in the metalanguage. For instance, the structural form of the argument (2) is:

(2') $H(v)$; therefore, $A(v)$,

where the predicate symbols H and A and the individual constant v may receive any interpretations of the appropriate type. The conceptual form, by contrast, is given in the metalanguage: in addition to the structural form, it imposes some constraint on the mutual relationships of the interpretations of the non-logical constants. In the minds of the medievals, the argument (2) is underwritten by the understanding that, in the Aristotelian hierarchy of universal terms, the concept 'animal' is part of the essence of the concept of 'human being'. Correspondingly, the conceptual form of (2) consists of (2') *and* the requirement that the extension of H be included in that of A :

(2'') ' $H(v)$; therefore, $A(v)$ ' and $ext(H) \subseteq ext(A)$.

Thus, for instance, 'Bucephalus is a horse; therefore, Bucephalus is a mammal' and 'Aristotle is Greek; therefore, Aristotle is Egyptian' are both of the structural form (2'), but only the first is of the conceptual form (2'').

With these notions of form at hand, we can proceed to define formal validity based on the following scheme:

(___-formal validity) An argument is ___-formally valid if and only if every instance of its ___ form is valid.

Each of the two notions of form, when applied to the scheme above, results in a corresponding notion of formal validity. An argument is *structural-formally valid* if every instance of its structural form is valid, and *conceptual-formally valid* if every instance of its conceptual form is valid.⁷ Structural-formal validity roughly

⁷ The two kinds of validity correspond to the "FO Con" and "Ana Con" procedures of the *Language, Proof and Logic* package mentioned above. "In theory", Ana Con should be able to verify any argument that is valid in virtue of the meanings of logical constants *and the non-logical constants* of the software's specific object language, while in practice, its implementation disregards the meanings of two predicates as too complex and even then fails to find some proofs (Barker-Plummer, Barwise and Etchemendy, 2011, 60–61, 577–578; see also Cohen, 2004). I take

corresponds to the so-called Parisian account of formal consequence, represented by John Buridan and a few others (see Archambault, 2017; Ciola, 2018). Conceptual-formal validity, in turn, renders a line of thought gradually developed in the so-called British tradition of formal consequence (see Saario-Ramsay, revised and resubmitted). My version of it is mostly inspired by Peter of Mantua (Strobino, 2017) and, to some extent, by Paul of Venice (1990). Structural-formal validity is analogous to our modern notion of logical consequence and has in fact been formalized using the resources of modern formal logic (Dutilh Novaes, 2007), but it is conceptual-formal validity that more closely corresponds to the dominant views in the 14th century (cf. Saario-Ramsay, revised and resubmitted). On this account of formal validity, arguments like (2), (3), and (4) are not merely materially valid, as they are on the structural account, but formally valid. According to a common formulation in the British school, such arguments are formally valid because the conclusion is *contained* in the signification or ‘in the formal understanding’ (*de formali intellectu*) of the premises, so that anyone who understands that the premises are true also understands that the conclusion is true (see e.g., Seaton, 1973; Green-Pedersen, 1981), or because the contradictory opposite of the conclusion ‘is formally incompatible with’ (*formaliter repugnare*) the premises (Paul of Venice, 1990).

Let us now return to the Medieval Recognition Problem. Are computers capable of recognizing conceptual-formally valid arguments? To get things going, let us adopt for now the set-up of John Searle’s Chinese Room Argument (Searle, 1980, 1984). Imagine that a computer is like a person—let us call her Siri—sitting in a closed room with two letterboxes, one for input and the other for output. She has a book of formal rules that tell her what conclusions she is allowed to derive from what premises—the rules are “formal” in the sense that they only concern the shapes of symbols—and apart from these rules, she knows absolutely nothing about the symbols. She is given a finite set of premises and a conclusion as an input and she is supposed to produce an output “Yes” if the conclusion is derivable from the premises by the rules. Accept, for the sake of argument, that this set-up is an adequate allegory of a computer. The question is, can you write a book of rules for conceptual-formal validity so that Siri would be able to correctly recognize all and only conceptual-formally valid arguments?

it that a full implementation for the particular language in question would be possible along the lines I discuss in Sections 0 and 0.

If Siri is given (1) as an input and asked if it is structural-formally valid, she can easily confirm that it is: all she needs to do is derive $H(s) \rightarrow A(s)$ from the second premise by the rule of universal instantiation and then derive $A(s)$ from that formula and the first premise by the rule of Modus Ponens. By contrast, if she is given (2) as an input and asked if it is conceptual-formally valid, it is far from obvious what kind of rules could have been written to allow her to handle cases like this. At face value, it seems that she would need to *understand* what the terms H (human being) and A (animal) are about, but it was assumed at the outset that she does not understand the underlying language beyond what is written in the rulebook. It might be thought that encoding all the complex connections between concepts understood by the human intellect into formal rules to be mechanically followed by a non-intellect is very challenging if not impossible. Even if they could be so encoded, we might wonder whether that would amount to “formal understanding” in the medieval sense: as Searle himself says, “whatever purely formal principles you put into the computer, they will not be sufficient for understanding” (Searle, 1980, 238).⁸

I will say that a notion of validity is *formalizable* if it can be implemented by a Chinese Room in the way outlined above. Importantly, validity must be reduced to formal rules that Siri can follow without understanding their content, and Siri will follow these rules in a completely mechanic manner: she is not allowed to make any exceptions or change the rules along the way. As noted, there are *prima facie* reasons to suspect that conceptual-formal validity cannot be reduced to formal rules of the kind described. These considerations give rise to the following argument:

Validity can be recognized by a computer only if it is formalizable;
Conceptual-formal validity is not formalizable;
Therefore, conceptual-formal validity cannot be recognized by a computer.

⁸ One of my referees objected that Searle’s Chinese Room Argument is not really relevant for this paper. I agree that I could very well study the question of modelling conceptual-formal validity without any reference to the Chinese Room. I just think that the Chinese Room provides us with a nice illustration of the idea of formalization: in this sense, it is nothing but a rhetorical device used to enliven the discussion. That said, I do think that there is an interesting analogy between the Searlean juxtaposition of syntax vs. semantics and the medieval juxtaposition of structure vs. conceptual containment. At the very least, the medieval way of combining “form” and “understanding” clearly challenges the Searlean way of contrasting these two notions.

That is to say, the Medieval Recognition Problem is unsolvable. Add the premise that conceptual-formal validity is essential to everyday reasoning, and you get the corollary that computers are unable to accommodate everyday reasoning. The crucial question is, do we need to accept this argument? In the following sections, I will attack the second premise—the claim that conceptual-formal validity is not formalizable—from various directions, drawing on different sources of inspiration in modern logic.⁹ My aim is not so much to argue for any particular view as to bring together alternative approaches, relate them to the present problem, and map out paths for further inquiry. Having identified problems with each of the considered attacks on the second premise, I devote the final section to questioning the first.

2. Enriching the Object Language

I am certainly not the first to think about ways of formalizing the medieval account of ‘conceptual-formal’ validity within modern logic. Peter King does so while discussing William of Ockham’s rule that a consequence “from a superior distributed term to an inferior distributed term”, such as “Every animal is running; therefore, every human being is running”, is valid (Ockham, 1974, 591). King notes that this rule:

treats the relation between the terms ‘animal’ and [‘human being’] as a formal feature. Modern first-order logic does not normally respect such relationships, but could do so in a number of ways: indexing or sorting the term-variables; adding semantic rules along the lines of meaning-postulates; and the like. (King, 2001, n. 35)

I will first consider the suggestion of indexing or sorting the term variables and postpone the discussion of meaning postulates. Given that term variables like ‘animal’ and ‘human being’ are usually formalized in first-order predicate logic by predicate symbols, we might follow King’s lead by enriching the object language with subscripted predicate symbols, where the subscripts are used to impose certain limitations on the interpretation of the symbols in question. Let us say that

⁹ I am thus taking up the challenge formulated by Stephen Read at the end of his paper (Read, 1993, 259) as follows: “What I have not tackled is the open question whether a non-classical account of logical consequence, perhaps based on the recurring idea of containment or inclusion of consequent in antecedent, can be developed, inspired by the medieval phrase *de formali intellectu*. But that will have to wait for another occasion.”

for any predicate symbols P and Q , there is a subscripted predicate symbol $P_{\subseteq Q}$, the extension of which must be included in that of Q . Ockham's example can then be formalized as:

$$\forall x(A(x) \rightarrow R(x)); \text{ therefore, } \forall x(H_{\subseteq A}(x) \rightarrow R(x)),$$

which comes out as valid in the semantics described.

A weakness of this strategy is revealed by considering another example:

(5) Valtteri is a human being; therefore, Valtteri is a living thing,

where the formalization of 'human being' by $H_{\subseteq A}$ and 'living thing' by L does not produce a valid form. We have two options: either we introduce an alternative symbol for 'human being', such as $H_{\subseteq L}$, and use that symbol rather than $H_{\subseteq A}$ in the formalization of this argument, or we conflate the two subscripts within one predicate symbol, as in $H_{\subseteq A, L}$, and use that symbol in the formalization of both arguments.

Further examples along these lines make the notation even more complicated. The problem is not just that all relations of mutual inclusion have to be integrated into the syntax; mutual inclusion is but one of many different types of term relationships, each of which needs its own notation. In argument (3), for instance, the relationship between the terms 'is west of' and 'is east of' is not that one is superior to the other; rather, they are each other's inverses in the sense that x is west of y if and only if y is east of x . Following the common practice in mathematical logic, this relationship would be formalized by something like:

$$W(t, h); \text{ therefore, } W^{-1}(h, t).^{10}$$

The more terms are included in the object language, the more complex is the syntax required to encode all their mutual relationships. And yet, as long as there is only a finite number of terms and only a finite number of different types of relationships among terms, all the possible relationships between the terms are in principle encodable by subscripted symbols. Once we have set down all the different types of term relationships that we want to be able to encode, we can tell the computer how to decode them. The question is, will this strategy provide us with a solution to the Medieval Recognition Problem?

¹⁰ The terms of argument (4) exhibit yet another kind of relationship, namely, mutual incompatibility. They could be formalized by, say, B and $W_{\cap B=\emptyset}$, where the subscript requires that the extensions of the predicates be disjoint; and so on for other types of term relationships.

I argue that it does not. A core intuition underlying the problem is that the computer should already ‘know’ how particular terms relate. It should, for instance, ‘know’ that the predicate ‘animal’ applies to every individual to which the predicate ‘human being’ applies. By the ‘enriched language’ strategy, however, the computer would not ‘know’ any such relationships; it would merely ‘know’ abstract patterns of the sort: ‘if there are any predicates of the form Q and $P_{\subseteq Q}$, then Q applies to every individual to which $P_{\subseteq Q}$ applies’. It would be the task of the user of the computer to fill in all relevant information on term relationships when feeding an argument to the computer, and this would have to be done separately for every argument ever fed to the computer. For instance, in order to check the validity of argument (5), the user should formalize not only the predicates as H and L , but also the inclusion of H in L by adding the appropriate subscript. The user would, in effect, do all the interesting work on behalf of the computer. We must therefore look for a solution elsewhere.

3. The Suppressed Premise Strategy

Perhaps the most obvious way to try to formalize conceptual-formal validity is what Read (1994) calls “the suppressed premise strategy”. This strategy is based on the idea that the condition on conceptual connections expressed in the metalanguage could be written out in a sentence of the object language and added as a premise to the argument to make it structural-formally valid. In the conceptual form of (2), for instance, the metalanguage condition $ext(H) \subseteq ext(A)$ could be translated into the object language sentence $\forall x(H(x) \rightarrow A(x))$. When we add this sentence as a premise to (2), we get the argument (1), which is structural-formally valid if and only if (2) is conceptual-formally valid.

Something like this can be already found in medieval authors. John Buridan, while defining formal validity in terms of structural form and thereby classifying consequences like (2) as merely materially valid, remarks that they can nevertheless be reduced to formally valid ones by adding a necessarily true premise (Buridan, 2015, 68). For instance, (2) could be reduced to a formal consequence by adding the premise ‘Every human being is an animal’, which expresses a necessary truth. Similarly, (3) and (4) could be reduced to formal consequences by adding, e.g., the necessary truths ‘Turku is west of Helsinki if and only if Helsinki is east of Turku’ and ‘Daffy is not both black and white’,

respectively. As it turns out, all conceptual-formally valid arguments can be reduced to structural-formally valid arguments in this way (cf. Read, 1994).

The Achilles heel of the strategy lies in the requirement that the added premise be *necessarily true*. The mere fact that an argument can be made structural-formally valid by adding a premise is insufficient for its conceptual-formal validity: indeed, *any* argument ' φ ; therefore, ψ ' can be trivially made structural-formally valid by adding the premise $\varphi \rightarrow \psi$. This leaves us with a dilemma. Either we require that the user supplies the suppressed premise, if there is one, for every argument that they feed to the computer, or we devise a computer that can supply the suppressed premise by itself. The first option would be cheating, for the user would then be doing the work of identifying conceptual-formally valid arguments while the computer would only recognize structural-formally valid ones.¹¹ The other option would require that the computer not only formulates the missing premise, but also recognizes that it is necessarily true, and that is easier said than done. Even if we found a way to teach the computer to recognize necessary truths, the act of adding a premise is not the critical step; on the contrary, the whole weight of the problem is shifted to recognizing the necessary truth of that premise (cf. Read, 1994, 258). Plausibly, a computer that can recognize necessary truth can just as well recognize conceptual-formal validity directly without recourse to structural-formal validity. We are back at square one.

4. The Shorthand Strategy

Another obvious way to formalize conceptual-formal validity is based on the idea that defined terms be short for their definitions (cf. Quine, 1951; Fine, 2005, 236). This insight would allow us to reduce conceptual-formally valid arguments to structural-formally valid ones by replacing the defined terms with their definitions. For instance, given that a 'human being' is defined as a 'rational animal', we could replace 'human being' with 'rational animal' in the argument (2) and thereby reduce it to:

Valtteri is a rational animal; therefore, Valtteri is an animal.

¹¹ Recall that we are looking for a formalization that is 'static' like that of structural-formal validity. The computer is not allowed to store suppressed premises from the user's input into its memory and use them for processing future input. We will give up this requirement in the final section.

In formalized language, this corresponds to applying the definition $H(x) \equiv_{df} R(x) \wedge A(x)$ to the argument (2'), resulting in:

$R(v) \wedge A(v)$; therefore, $A(v)$.

This argument is clearly structural-formally valid, and hence we would conclude that the original argument (2) is conceptual-formally valid.

Certain qualifications have to be made to circumvent certain difficulties analogical to the problems of the suppressed premise strategy. First, the fact that a given argument can be made structural-formally valid by expanding suitable 'definitions' of the primitive terms is insufficient for conceptual-formal validity. For instance:

Plato is a philosopher; therefore, Plato is a potato

reduces to a structural-formally valid argument under the substitution of 'mashed potato' for 'philosopher', but this 'definition' is of course incorrect. It is an obvious requirement that only correct definitions be used. Moreover, we cannot expand the correct definitions before we feed the argument to the computer, for it would, again, be cheating.

It remains to be shown how to encode all the definitions to a computer. This does not seem too difficult: cannot we just list all the definitions and append them to Siri's rulebook, like a dictionary where she can look up the definitions of primitive predicate symbols? So long as there is only a finite number of non-circular definitions formalizable in the object language, it is fairly easy to design an algorithm for expanding all of them in any given argument. Suppose, for instance, that the dictionary included the following definitions (here given in the vernacular for the sake of illustration):

'bachelor' \equiv_{df} 'unmarried man'
'man' \equiv_{df} 'male human'
'human' \equiv_{df} 'rational animal'.

We could then reduce the argument:

Plato is a bachelor; therefore, Plato is rational

to a structural-formally valid one, applying each of the three definitions in turn:

Plato is an unmarried man; therefore, Plato is rational

Plato is an unmarried male human; therefore, Plato is rational

Plato is an unmarried male rational animal; therefore, Plato is rational.

This approach applies to all cases of conceptual-formal validity that are grounded on definitions of terms. Argument (2) is another example of this kind. Arguments (3) and (4), however, are different.¹² In fact, the examples (2), (3), and (4) have been deliberately modelled upon three medieval ‘degrees of formality’ (Paul of Venice, 1990, 90–93; see also Strobino, 2017), only the first of which is based on definitions. It is impossible that Valtteri would be a human being but not an animal, because the definition of the term ‘human being’ includes the term ‘animal’.

The two other degrees of formality are grounded on necessary truths of a weaker kind, where the meanings of the terms are necessarily connected but not through their definitions. To accommodate these degrees of formality, we should include in Siri’s dictionary all such necessary truths, which are typically more complicated than definitions. Take (4), for instance. Black and white are not defined in terms of each other but rather, say, in terms of wavelengths of light or primitive phenomenal qualities—neither is more primitive than the other (Read, 1994, 250–251)—and yet it is a necessary truth that nothing is both black and white. The conceptual form of (4) is:

‘ $B(d)$; therefore, $\neg W(d)$ ’ and $ext(B) \cap ext(W) = \emptyset$,

where $ext(B) \cap ext(W) = \emptyset$ expresses a non-definitional necessary truth. Such necessary truths cannot be reduced to dictionary definitions, but perhaps they can be encoded to a computer in some alternative way.

5. Meaning Postulates

Rudolf Carnap (1952) defends the notion of analyticity against Quine’s (1951) famous attack by introducing semantic rules known as *meaning postulates*.¹³ In

¹² Boghossian would say that they are examples of “a priori statements that are not Frege-analytic” (Boghossian, 1996, 368).

¹³ Cf. Boghossian’s (1996) idea of “implicit definition”.

the framework of the present paper, Carnap's meaning postulates can be cashed out as the object language translations of the metalanguage conditions pertaining to conceptual form. The difference with respect to the suppressed premise strategy is that these sentences are not added as premises to any particular arguments but rather constitute a body of 'presuppositional knowledge' that is common to all arguments and belongs to the underlying system as a whole. The difference with respect to the shorthand strategy is that these sentences include not only definitions but also weaker necessary truths that are supposed to constrain the interpretation of the predicate symbols. To continue with our previous metaphor, we could say that meaning postulates are added as a special appendix to Siri's rulebook. This appendix is not made up of definitions like a traditional dictionary, but rather of non-logical axioms that enable Siri to deal with non-logical symbols according to their intended meanings.

To guarantee the validity of arguments (2), (3), and (4), we may introduce the following meaning postulates:

$$(MP2) \quad \forall x(H(x) \rightarrow A(x))$$

$$(MP3) \quad \forall x\forall y(W(x, y) \leftrightarrow E(y, x))$$

$$(MP4) \quad \neg\exists x(B(x) \wedge W(x)).$$

Such postulates are arguably similar to 'intrinsic middles' or 'intrinsic topics' that were appealed to by medieval logicians like Walter Burley and William of Ockham (cf. King, 2001, 130; Archambault, 2018; Crimi, 2018). For instance, the formal consequence 'A human being is running; therefore, an animal is running' was apparently thought to hold through the intrinsic middle 'Every human being is an animal' (Crimi, 2018, 261).

While this strategy is in many ways superior to the previous ones, it should be noted that the available meaning postulates are limited by the expressive power of the object language. There are cases in which the requisite meaning postulate simply cannot be expressed in the object language (cf. Etchemendy, 1999, 71). Consider, for instance, arguments such as:

Ada is a parent of Bob; therefore, Bob is not a parent of Ada.

Ada is a parent of Bob; Bob is a parent of Chris; therefore, Ada is an ancestor of Chris.

Ada and Bob have no ancestor in common; Chris is a parent of Ada; therefore, Chris is not a parent of Bob.

In order to be able to deal with all possible arguments involving the binary predicates ‘is a parent of’ and ‘is an ancestor of’, we would need particularly strong meaning postulates. Let us assume, for simplicity, that there are no other (non-logical) predicates in our language besides these two. The predicate ‘is a parent of’ might be defined by axioms of irreflexivity (‘no one is a parent of oneself’), asymmetry (‘for all x and y , if x is a parent of y , then y is not a parent of x ’), and so on. However, to establish the semantic connection between ‘is a parent of’ and ‘is an ancestor of’, we would need to add a postulate along the following lines:

For all x and y , x is an ancestor of y if and only if there are z_1, z_2, \dots, z_n such that x is a parent of z_1 , z_1 is a parent of z_2 , ..., and z_n is a parent of y .

This postulate can only be expressed in higher-order logic. While we can choose our object language to be of any order we like, limitations of expressive power are bound to reappear at the higher level. In general, the Carnapian approach limits the conceptual form to what can be expressed in the object language, but plausibly, the realm of meaning exceeds any such technical limitations: the medievals, in particular, knew no distinction between object and metalanguage. In the next section, I will consider the possibility of transferring the Carnapian approach to the level of metalanguage.

6. From Rules to Models

The three previous strategies have all tried to embed conceptual-formal validity to structural-formal validity through some restrictive conditions expressed at the object level, be it in the form of premises, definitions, or semantic rules. In this section, I show how the notion of conceptual-formal validity can be directly formalized at the metalevel, taking up the informal characterization that we started with in the first section and making it more precise by modifying the standard model-theoretic semantics.

Suppose that our object language contains the predicate symbols H (‘human being’) and A (‘animal’). In ordinary metalogic, we would characterize these symbols as non-logical constants and vary their interpretation quite freely from model to model, while keeping the interpretation of logical constants essentially

fixed across models. I say “quite freely” and “essentially fixed” because, on one hand, the predicate symbols cannot be assigned just any sets as extensions: they have to be subsets of the domain. The domain, on the other hand, is the set that the quantifiers range over, and this set varies with the model. In Etchemendy’s (1999) terminology, there is a “cross-term restriction” that fixes a certain connection between the predicate symbols and the quantifiers, namely, that the extensions of the predicate symbols are subsets of the domain of quantification. However, “there is no significant difference between this sort of restriction and the use of so-called meaning postulates” (Etchemendy, 1999, 71), such as the requirement that the extension of H be included in that of A . Why, then, is the former restriction accepted in the standard semantics while the latter is rejected? Etchemendy’s response is to reject the former along with the latter; our response, by contrast, is to incorporate the latter in the metalogic along with the former.

Gillian Russell (2024) makes a similar point by distinguishing between fixed (‘conditional’) and variable (‘environmental’) meanings of expressions. The key insight, as I see it, is this. The classical demarcation of expressions into logical and non-logical is, at best, a rough approximation; to make it more accurate, we must formalize each expression as consisting of a fixed (logical) and a variable (non-logical) part. As we saw above, not even in standard metalogic is there really a sharp boundary between logical and non-logical constants. Take the interpretation of the universal quantifier, for instance. Its intuitive meaning ‘everything’ can literally be split up into the fixed part ‘every’ and the variable part ‘thing’ (cf. Etchemendy, 1999, 67). The meaning of the quantifier requires satisfiability by *every* member of the domain, and this requirement invariably applies in any model, while the domain itself—the extension of ‘thing’—varies with the model. But if quantifiers are already cut in half like this, why not let the cut extend to constants traditionally held as non-logical? Indeed, we may similarly divide the interpretation of our predicate symbols H and A into two parts: the requirement that the extension of H be a subset of that of A is the fixed part, and the elements of the extensions, which can be any elements of the domain as long as the required relation of subethood holds, is the variable part. Logicality is thus not a binary classification but a matter of degree: some constants are more logical than others, depending on the relative ‘sizes’ of their fixed parts (cf. Sagi, 2018).

With this demarcation of fixed and variable parts in mind, we get a precise idea of what an “instance” of a conceptual form is: it is a model, the definition of which has been amended by imposing certain restrictions on the interpretations of

the predicate symbols. Just like with ordinary structural form, formal validity is truth-preservation in all models, only now the set of relevant models has been narrowed down. We have thus incorporated the Carnapian meaning postulates into the metalogic in a way that is a natural continuation of the usual semantics. On this account, as Russell (2024, 233) puts it, “there is no principled discontinuity between logical and analytic consequence”, or as we would say, *between structural-formal and conceptual-formal validity*. Argument (2) is conceptual-formally valid because, by the refined definition of a model, all the models of its premise are also models of its conclusion.

To make things a bit more concrete, consider models $M_0 = \langle D_0, I_0 \rangle$ and $M_1 = \langle D_1, I_1 \rangle$ with the following domains:

$$D_0 = \{\text{Valtteri, Socrates, Daffy, Helsinki, Turku, Big Ben, } \sqrt{2}\}$$

$$D_1 = \{\text{Plato, Aristotle, Alexander, Brownny, Bucephalus, Marathon}\}.$$

The predicates H and A and the individual constant v are interpreted as follows:

$$I_0(H) = \{\text{Valtteri, Socrates}\}$$

$$I_0(A) = \{\text{Valtteri, Socrates, Daffy}\}$$

$$I_0(v) = \text{Valtteri}$$

$$I_1(H) = \{\text{Plato, Aristotle, Alexander}\}$$

$$I_1(A) = \{\text{Plato, Aristotle, Alexander, Brownny, Bucephalus}\}$$

$$I_1(v) = \text{Marathon}.$$

In both domains, the extension of H is a subset of that of A ; otherwise the construction would not qualify as a model. Model M_0 is a model of both $H(v)$ and $A(v)$, whereas the model M_1 is a model of neither $H(v)$ nor $A(v)$. There could just as well be a model M_2 which is not a model of $H(v)$ but is a model of $A(v)$. There cannot, however, be a model which is a model of $H(v)$ but not of $A(v)$, and hence the argument (2) is conceptual-formally valid.

Similar adjustments to the definition of a model could be made to accommodate the conceptual-formal validity of arguments (3) and (4). An example of an argument that is merely materially valid might be something like ‘There are no irrational numbers; therefore, Valtteri is running’, formalized as:

$$\neg \exists x I(x); \text{ therefore, } R(v).$$

Here, the premise is thought to be necessarily false but only ‘as a matter of fact’, because of its particular content.¹⁴ In our formalization, this is to say that not all models of the premise are also models of the conclusion.

Let me use this opportunity to make a slight digression and sketch a systematic hierarchy of different types of models. For the sake of illustration, I will conceive of these models as possible worlds as in Etchemendy’s (1999) ‘representational semantics’.¹⁵ In place of the scheme of ‘___-formal validity’ proposed in the first section, let us now adopt the following, which is closer to the contemporary (Tarskian) way of doing things:¹⁶

(___ validity) An argument is ___ly valid if and only if in all ___ly possible worlds (models) where its premises are true, its conclusion is also true.

At the first level of the hierarchy, we have the standard definition of a model. The corresponding models are identified as the *structural-formally* (or *logically*) possible worlds, and the resulting set of valid arguments is that of structural-formally valid ones. At the second level, we have the definition of a model amended in the way described above, resulting in *conceptual-formally* (or *semantically*) possible worlds and conceptual-formally valid arguments. Conceptual-formally possible worlds are a subset of structural-formally possible worlds, and conversely, conceptual-formally valid arguments are a superset of structural-formally valid arguments. At the third level, the definition of a model is further restricted so that it excludes metaphysical impossibilities, such as there being no irrational numbers, but still includes physical impossibilities, such as Valtteri running faster than the speed of light. Such *metaphysically* possible worlds are, again, a subset of conceptual-formally possible worlds, and the resulting set of metaphysically valid arguments is a superset of the set of conceptual-formally valid arguments. In this way, the medieval idea of conceptual form can be seen as introducing an intermediate sphere of necessity between the familiar categories of

¹⁴ The medievals’ favourite example of such a falsehood was ‘God does not exist’ (see e.g., Strobino, 2017).

¹⁵ Here I diverge from Russell’s (2024) account, wherein models are cashed out not as possible worlds but as combinations of worlds and languages. In this respect, I also diverge from other ‘hybrid’ accounts of logical consequence that have been developed in response to Etchemendy, such as those of Sher (1996) and Shapiro (2007).

¹⁶ Note the analogy with “Tarski’s thesis” (McGee, 1992) and the “Generalised Tarski Thesis” (Beall and Restall, 2005).

logical necessity and *metaphysical* (or broadly logical) necessity (cf. Plantinga, 1978; Fine, 2005). The three spheres of necessity are illustrated in Fig. 1.¹⁷

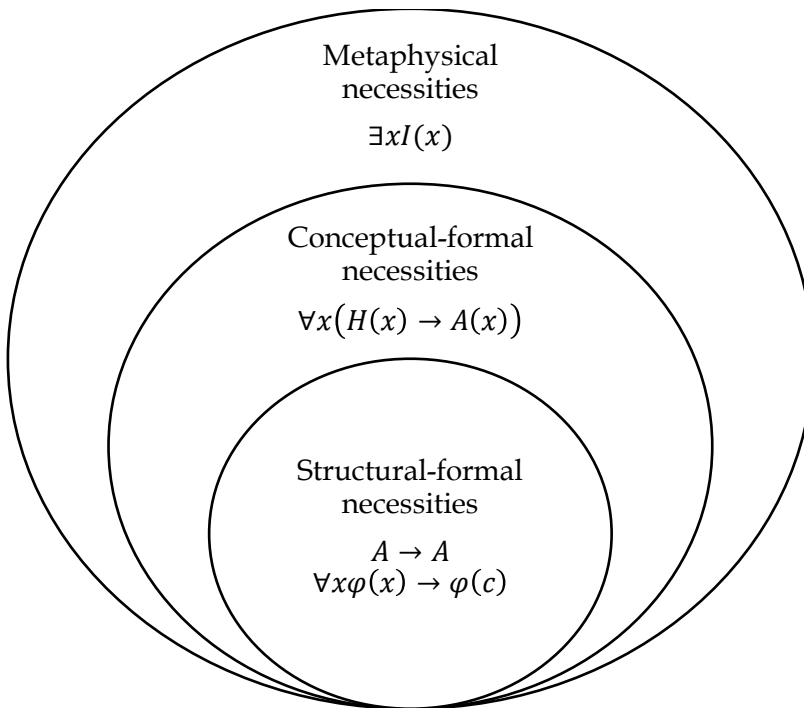


Fig. 1. Spheres of metaphysical, conceptual-formal (semantic), and structural-formal (logical) necessity.

This formalization of conceptual-formal validity not only connects to the contemporary discussion on the nature of necessity; it is also, I think, more faithful to the medieval account of degrees of formality than the alternative formalizations. This formalization also fares better than Carnap's meaning postulates in making room for conceptual relationships that cannot be expressed in first-order logic. Yet the problem was whether this formalization makes it possible for Siri to recognize conceptual-formally valid arguments, and that, unfortunately, seems somewhat

¹⁷ The hierarchy could be extended in the obvious way by recognizing *physically* (causally, naturally) possible worlds as a subset of the metaphysically possible ones (cf. Plantinga, 1978, 14). In the opposite end, we could likewise recognize 'paralogically' possible worlds as a superset of logically possible ones. Such worlds would attach no fixed meanings to any constants beyond the minimal requirement that they be of the appropriate type: for instance, the logical connectives could be interpreted as any truth-functions of the appropriate arity, so that contradictions like $p \wedge \neg p$ would come out true in some possible worlds.

doubtful. The problem is that Siri could no longer restrict herself to the first-order rules of inference; instead, she would have to work with the definition of a model, and that definition is given in the higher-order metatheory. As is well-known, there is no complete proof system even for second-order logic with standard semantics. The details are too complicated to be discussed here, but the general worry is clear: basically, the metalogic may be too thick for Siri to swallow.

7. Beyond Formalization

All the strategies for solving the Medieval Recognition Problem discussed above ended up in difficulties of one kind or another. Should we give up and conclude that the problem is unsolvable? Perhaps it is indeed the case that conceptual-formal validity is not fully formalizable; perhaps there simply is no place for conceptual form in the Chinese Room. Even so, that does not yet show that conceptual-formal validity cannot be recognized by a computer. The claim that conceptual-formal validity is not formalizable was merely the second premise of our argument to that conclusion. We are still left with the first premise, namely, the claim that conceptual-formal validity is recognizable by a computer *only if* it is formalizable. Let me conclude by briefly outlining ways of debunking this first premise.

The first premise was originally based on the assumption that the Chinese Room is an adequate allegory of a computer. If a computer is nothing but an automaton manipulating symbols according to pre-given instructions, then it seems indeed that the only way for a computer to recognize conceptual-formally valid arguments would require a list of instructions for recognizing such arguments by their syntactic appearance. We do not, however, need to accept that the Chinese Room is an adequate allegory of a computer. Developments in artificial intelligence and related fields (see e.g., Bringsjord and Govindarajulu, 2024) suggest that a machine-learning model could be trained to recognize conceptual-formally valid arguments and outperform any traditional formalization we could possibly think of. As opposed to our original ‘top-down’ Chinese Room, where Siri was applying ready-made rules to given cases, our renovated ‘bottom-up’ Chinese Room would have Siri learn from concrete examples of conceptual-formally valid/invalid arguments and write the rules by herself, as it were, on that basis. The rules would amount to probabilistic calculations rather than, say, complete axiomatizations, and they would occasionally get it wrong, but perhaps we should give up on the requirement of complete accuracy. After all, human beings are fallible, too, and the class of conceptual-formally valid arguments may

be too vague anyway to ever be given a fully definitive formalization. The accuracy of the model might never reach 100%, but at least we could approximate the accuracy by testing the model with different data than it was trained with, and we could always increase the accuracy by retraining the model with more data or by fine-tuning its inner architecture.

Arguably, conceptual-formal validity has already been realized in the said sense by the recent rise of large language models. Take ChatGPT, for instance, which is a chatbot based on such a model. When I enter the prompt “Valtteri is a human being. Does it follow that Valtteri is an animal?” into ChatGPT, I get the following response:

Yes, it follows that Valtteri is an animal because humans are biologically classified as animals. Specifically, humans belong to the kingdom Animalia. Therefore, if Valtteri is a human being, Valtteri is also an animal. (OpenAI, 2024)¹⁸

Indeed, it seems that the Chinese Room has already undergone a thorough renovation. Thanks to the renovation, conceptual form appears more welcome in the Chinese Room, but we can go even further than that. The rise of externalist ‘4E’ theories of mind and the concomitant advances in robotics provoke the thought that one day we could build a full-blown artificial mind (Telakivi and Arstila, 2021). Imagine, for instance, an artificial neural network that is embodied in a robot, embedded in and evolving in continuous interaction with its environment through sensors and actuators, pretty much like the human brain. Such a system would then be expected to *make sense* of its environment and come to learn *meanings* of terms like ‘human being’ and ‘animal’ based on its experience, emulating the way human beings learn the meanings of those terms (cf. Telakivi and Arstila, 2021). Just as we can give a conceptual-formally valid argument to a human being who then confirms that it is conceptual-formally valid, we could give such an argument to the artificial neural network, which would likewise confirm that it is conceptual-formally valid. The inner workings of the network might appear as a black box—we might not, for instance, be able to identify any part of the network as a representation of the term ‘human being’ or of the rule by which its meaning is governed—and again, there is no 100% guarantee of accuracy, but perhaps that is just the price we have to pay for

¹⁸ This response was created using the preselected model, GPT-4o.

conceptual-formal validity. Perhaps conceptual form is too savage a beast to be tamed by any rules of our clean-cut modern logic; too widespread a jungle to be contained by any room.¹⁹

University of Helsinki

References

- Archambault, J. (2017), *The Development of the Medieval Parisian Account of Formal Consequence*, PhD thesis, Fordham University. Available at: <https://research.library.fordham.edu/dissertations/AAI10617109> (Accessed: 26 January 2023).
- Archambault, J. (2018), “Consequence and Formality in the Logic of Walter Burley”, *Vivarium* 56(3–4), 292–319. Available at: <https://doi.org/10.1163/15685349-12341355>.
- Aristotle (2009), *Prior Analytics: Book I*, translated with an introduction and commentary by G. Striker, Oxford University Press, Oxford. Available at: <https://doi.org/10.1093/actrade/9780199250400.book.1>.
- Barker-Plummer, D., J. Barwise and J. Etchemendy (2011), *Language, Proof and Logic*, 2nd edn., CSLI Publications, Stanford, CA.
- Beall, J. C. and G. Restall (2005), *Logical Pluralism*, Oxford University Press, Oxford. Available at: <https://doi.org/10.1093/acprof:oso/9780199288403.001.0001>.
- Boghossian, P. A. (1996), “Analyticity Reconsidered”, *Noûs* 30(3), 360–391. Available at: <https://doi.org/10.2307/2216275>.
- Boghossian, P. and T. Williamson (2020), *Debating the A Priori*, Oxford University Press, Oxford. Available at: <https://doi.org/10.1093/oso/9780198851707.001.0001>.
- Bringsjord, S. and N. S. Govindarajulu (2024), “Artificial Intelligence”, in E. N. Zalta and U. Nodelman (eds.) *The Stanford Encyclopedia of Philosophy*, Summer 2024, Metaphysics Research Lab, Stanford University. Available at: <https://plato.stanford.edu/archives/sum2024/entries/artificial-intelligence/> (Accessed: 19 June 2024).
- Buridan, J. (2015), *Treatise on Consequences*, translated with an introduction by S. Read, Fordham University Press, New York, NY. Available at: <https://doi.org/10.5422/fordham/9780823257188.001.0001>.

¹⁹ I am grateful to Gabriel Sandu, José Filipe Silva, and two anonymous referees for their helpful comments on the manuscript. The ideas developed in Section 0 have originally been inspired by Gillian Russell’s presentation “Rethinking Logical Consequence” and the subsequent discussion in the Metaphysics and Logic Seminar at the Arché Research Centre, University of St Andrews, in May 2023.

- Burley, W. (2000), *On the Purity of the Art of Logic: The Shorter and the Longer Treatises*, translated by P. V. Spade, Yale University Press, New Haven, CT. Available at: <https://doi.org/10.12987/yale/9780300082005.001.0001>.
- Carnap, R. (1952), “Meaning Postulates”, *Philosophical Studies* 3(5), 65–73. Available at: <https://doi.org/10.1007/BF02350366>.
- Church, A. (1936), “A Note on the Entscheidungsproblem”, *Journal of Symbolic Logic* 1(1), 40–41. Available at: <https://doi.org/10.2307/2269326>.
- Ciola, G. (2018), “Marsilius of Inghen on the Definition of *Consequentia*”, *Vivarium* 56(3–4), 272–291. Available at: <https://doi.org/10.1163/15685349-12341359>.
- Cohen, S. M. (2004), “The Logic of Boolean Connectives”, in *Introduction to Logic: Lecture Notes*, University of Washington. Available at: <https://faculty.washington.edu/smcohen/120/Chapter4.pdf> (Accessed: 19 June 2024).
- Copeland, B. J. (2024), “The Church-Turing Thesis”, in E. N. Zalta and U. Nodelman (eds.) *The Stanford Encyclopedia of Philosophy*, Spring 2024, Metaphysics Research Lab, Stanford University. Available at: <https://plato.stanford.edu/archives/spr2024/entries/church-turing/> (Accessed: 4 April 2024).
- Corcoran, J. (1998), “Information-Theoretic Logic”, in C. Martínez, U. Rivas, and L. Villegas-Forero (eds.) *Truth in Perspective*, Ashgate, Aldershot, 113–135. Available at: <https://doi.org/10.4324/9780429433450>.
- Crimi, M. (2018), “Formal and Material Consequences in Ockham and Buridan”, *Vivarium* 56(3–4), 241–271. Available at: <https://doi.org/10.1163/15685349-12341360>.
- Davis, M. (2018), *The Universal Computer: The Road from Leibniz to Turing*, 3rd edn., CRC Press, Boca Raton, FL. Available at: <https://doi.org/10.1201/9781315144726>.
- Dutilh Novaes, C. (2007), *Formalizing Medieval Logical Theories: Suppositio, Consequentiae and Obligationes*, Springer, Dordrecht. Available at: <https://doi.org/10.1007/978-1-4020-5853-0>.
- Dutilh Novaes, C. (2020), “Medieval Theories of Consequence”, in E. N. Zalta (ed.) *The Stanford Encyclopedia of Philosophy*, Fall 2020, Metaphysics Research Lab, Stanford University. Available at: <https://plato.stanford.edu/archives/fall2020/entries/consequence-medieval/> (Accessed: 19 May 2023).
- Etchemendy, J. (1999), *The Concept of Logical Consequence*, CSLI Publications, Stanford, CA.
- Etchemendy, J. (2008), “Reflections on Consequence”, in D. Patterson (ed.) *New Essays on Tarski and Philosophy*, Oxford University Press, Oxford, 263–299. Available at: <https://doi.org/10.1093/acprof:oso/9780199296309.003.0011>.

- Fine, K. (2005), “The Varieties of Necessity”, in K. Fine (ed.) *Modality and Tense: Philosophical Papers*, Oxford University Press, Oxford, 235–260. Available at: <https://doi.org/10.1093/0199278709.003.0008>.
- Fitting, M. (2022), “Intensional Logic”, in E. N. Zalta and U. Nodelman (eds.) *The Stanford Encyclopedia of Philosophy*, Winter 2022, Metaphysics Research Lab, Stanford University. Available at: <https://plato.stanford.edu/archives/win2022/entries/logic-intensional/> (Accessed: 14 June 2024).
- Green-Pedersen, N. J. (1981), “Nicolaus Drukken de Dacia’s Commentary on the *Prior Analytics* – with Special Regard to the Theory of Consequences”, *Cahiers de l’Institut Du Moyen-Âge Grec et Latin* 37, 42–69.
- Halbach, V. (2020), “The Substitutional Analysis of Logical Consequence”, *Nous* 54(2), 431–450. Available at: <https://doi.org/10.1111/nous.12256>.
- Hilbert, D. and W. Ackermann (1928), *Grundzüge der theoretischen Logik*, Springer, Berlin.
- King, P. (2001), “Consequence as Inference: Mediaeval Proof Theory 1300–1350”, in M. Yrjönsuuri (ed.) *Medieval Formal Logic: Obligations, Insolubles and Consequences*, Springer, Dordrecht, 117–145. Available at: https://doi.org/10.1007/978-94-015-9713-5_5.
- Klima, G. (2016), “Consequence”, in C. Dutilh Novaes and S. Read (eds.) *The Cambridge Companion to Medieval Logic*, Cambridge University Press, Cambridge, 316–341. Available at: <https://doi.org/10.1017/CBO9781107449862.014>.
- Kment, B. (2006), “Counterfactuals and the Analysis of Necessity”, *Philosophical Perspectives* 20(1), 237–302. Available at: <https://doi.org/10.1111/j.1520-8583.2006.00108.x>.
- McGee, V. (1992), “Two Problems with Tarski’s Theory of Consequence”, *Proceedings of the Aristotelian Society* 92(1), 273–292. Available at: <https://doi.org/10.1093/aristotelian/92.1.273>.
- Ockham, W. of (1974), *Opera Philosophica. Vol. I: Summa Logicae*, edited by P. Boehner, G. Gál, and S. Brown, The Franciscan Institute, St. Bonaventure, NY.
- OpenAI (2024), *ChatGPT*. Available at: <https://chatgpt.com/share/670b99b8-42d4-8005-825a-f8203151af08> (Accessed: 13 October 2024).
- Paul of Venice (1990), *Logica Magna: Part II Fascicule 4*, edited by G. E. Hughes, Oxford University Press, Oxford.
- Plantinga, A. (1978), *The Nature of Necessity*, Oxford University Press, Oxford. Available at: <https://doi.org/10.1093/0198244142.001.0001>.
- Quine, W. V. (1951), “Two Dogmas of Empiricism”, *The Philosophical Review* 60(1), 20–43. Available at: <https://doi.org/10.2307/2181906>.

- Read, S. (1993), “Formal and Material Consequence, Disjunctive Syllogism and Gamma”, in K. Jacobi (ed.) *Argumentationstheorie*, Brill, Leiden, 233–259. Available at: https://doi.org/10.1163/9789004450448_019.
- Read, S. (1994), “Formal and Material Consequence”, *Journal of Philosophical Logic* 23(3), 247–265. Available at: <https://doi.org/10.1007/BF01048482>.
- Read, S. (2012), “The Medieval Theory of Consequence”, *Synthese* 187(3), 899–912. Available at: <https://doi.org/10.1007/s11229-011-9908-6>.
- Read, S. (2015), “Proof-Theoretic Validity”, in C. R. Caret and O. T. Hjortland (eds.) *Foundations of Logical Consequence*, Oxford University Press, Oxford, 136–158. Available at: <https://doi.org/10.1093/acprof:oso/9780198715696.003.0005>.
- Russell, G. (2008), *Truth in Virtue of Meaning: A Defence of the Analytic/Synthetic Distinction*, Oxford University Press, Oxford. Available at: <https://doi.org/10.1093/acprof:oso/9780199232192.001.0001>.
- Russell, G. (2024), “Logical Consequence (Slight Return)”, *Aristotelian Society Supplementary Volume* 98(1), 233–254. Available at: <https://doi.org/10.1093/arisup/akae012>.
- Saario-Ramsay, L. (revised and resubmitted), “The Two Schools of Formal Consequence in the 14th Century: Redrawing the Distinction”.
- Sagi, G. (2018), “Logicality and Meaning”, *The Review of Symbolic Logic* 11(1), 133–159. Available at: <https://doi.org/10.1017/S1755020317000247>.
- Searle, J. R. (1980), “Minds, Brains, and Programs”, *Behavioral and Brain Sciences* 3(3), 417–424. Available at: <https://doi.org/10.1017/S0140525X00005756>.
- Searle, J. R. (1984), *Minds, Brains and Science*, Harvard University Press, Cambridge, MA.
- Seaton, W. K. (1973), *An Edition and Translation of the Tractatus de Consequentibus by Ralph Strode, Fourteenth Century Logician and Friend of Geoffrey Chaucer*, PhD thesis, University of California, Berkeley.
- Shapiro, S. (2007), “Logical Consequence, Proof Theory, and Model Theory”, in S. Shapiro (ed.) *The Oxford Handbook of Philosophy of Mathematics and Logic*, Oxford University Press, Oxford, 651–670. Available at: <https://doi.org/10.1093/oxfordhb/9780195325928.003.0021>.
- Sher, G. Y. (1996), “Did Tarski Commit ‘Tarski’s Fallacy’?”, *Journal of Symbolic Logic* 61(2), 653–686. Available at: <https://doi.org/10.2307/2275681>.
- Strobino, R. (2017), “What Is Form All About? A 14th-Century Discussion of Logical Consequence”, in L. Cesalli, F. Goubier, and A. De Libera (eds.) *Formal Approaches and Natural Language in Medieval Logic*, Brepols, Turnhout, 173–194. Available at: <https://doi.org/10.1484/M.TEMA-EB.4.2017095>.

- Sundholm, G. (2013), “Containment and Variation; Two Strands in the Development of Analyticity from Aristotle to Martin-Löf”, in M. van der Schaar (ed.) *Judgement and the Epistemic Foundation of Logic*, Springer, Dordrecht, 23–35. Available at: https://doi.org/10.1007/978-94-007-5137-8_3.
- Telakivi, P. and V. Arstila (2021), “Onko mahdollista rakentaa keinotekoinen mieli?”, in P. Raatikainen (ed.) *Tekoäly, ihminen ja yhteiskunta: Filosofisia näkökulmia*, Gaudeamus, Helsinki, 87–101.
- Turing, A. M. (1936), “On Computable Numbers, with an Application to the Entscheidungsproblem”, *Proceedings of the London Mathematical Society* s2-42(1), 230–265. Available at: <https://doi.org/10.1112/plms/s2-42.1.230>.
- Wodeham, A. de (1990), *Lectura secunda in librum primum sententiarum: Distinctiones II–VII*, edited by G. Gál and R. Wood, St. Bonaventure University, St. Bonaventure, NY.

Some Reflections on the Notion of Arbitrary Function

GABRIEL SANDU

1. The background

In *Frege's Philosophy of Language* (1973), Dummett claims that Frege's notion of a function coincides with the notion of an arbitrary correspondence:

[...] Frege had not the slightest qualm about the legitimacy or intelligibility of higher-order quantification: he used it from the first, in *Begriffsschrift*, freely and without apology, and did not even see first-order logic as constituting a fragment having any special significance (Dummett, 1973, 223)

[...] it is true enough, in a sense, that, once we know what objects there are, then we also know what functions there are, at least, so long as we are prepared, as Frege was, to admit all "arbitrary" functions defined over all objects. (*ibid.*, 177)

In Hintikka and Sandu (1992), I claimed with Hintikka, against Dummett, that Frege's notion of a function and of a class cannot be that of an arbitrary correspondence or arbitrary collection of objects. We argued that Frege favoured, instead, some variety of *non-standard interpretation*, for which the domain of the function variables is something less than the characteristic functions of all subsets of the domain over which the individual variables range. Let me also mention that when we wrote our paper, we were unaware that Dummett himself changed his mind in his later work, *Frege's Philosophy of Mathematics* (1991), where he wrote:

[...] Frege fails to pay due attention to the fact that the introduction of the [class] abstraction operator brings with it, not only new singular terms, but an extension of the domain. [...] may be seen as making an inconsistent demand on the size of the domain D , namely that, where D comprises n objects, we should have $n^n \leq n$, which holds only when $n = 1$, whereas we must have $n \geq 2$, since the two truth-values are distinct: for there must be n^n extensionally non-equivalent functions of one argument and hence n^n distinct value-ranges. But this assumes that the function-variables range over the entire classical totality of functions from D into D , and there is meagre evidence for attributing such a conception to Frege. His formulations make it more likely that he thought of his function-variables as ranging over only those functions that could be referred to by functional expressions of his symbolism (and thus over a denumerable totality of functions), and of the domain D of objects as comprising value-ranges only of such functions. (Dummett, 1991, 219–220)

As we see, in the last two sentences of this passage Dummett is attributing to Frege a so-called *non-standard interpretation* of the function-variables, which is alike in spirit to some of the considerations I put forward with Hintikka:

Sometimes, a non-standard interpretation is guided by the idea that only such properties, relations, and functions can be assumed to exist as can be defined or otherwise captured by a suitable expression of one's language. In the case of theories with infinite models, this leads inevitably to a non-standard interpretation, for there can be only a countable number of such definitions or characterisations available for this purpose. Hence they cannot capture all the subsets of $\text{do}(M)$, for there is an uncountable number of them. (Hintikka & Sandu, 1992, 292)

For me and Hintikka, the fact that the connection between an argument and the corresponding value of a function is captured by a suitable expression in one's language shows that the correlation between arguments and values is not arbitrary but is determined by a formal law, which usually takes the form of an equation. In the same vein, the definability of a class by an expression of the relevant language shows that membership in the class is determined by a norm, which usually takes the form of a concept. For such reasons, the main argument of our paper focused

on the distinction between, on one side, the idea of arbitrary variation between values and arguments, and the idea of correlation as determined by a formal law, on the other. We argued that Frege could not have had a standard interpretation of his function-variables given that the notion of a law was important for him when characterising functions. We partly supported our claim by invoking Frege's discussion of the inadequacies of the definition of a function proposed by Czuber. Once again, we noticed that Czuber's notion of correlation involves no assertion as to the law of correlation, but can be set up in the most various ways, whereas Frege's conception of correlation focuses on the idea of a law:

Correlation, then, takes place according to a law, and different laws of this sort can be thought of. In that case, the expression 'y is a function of x' has no sense, unless it is completed by the law of correlation. (Frege 1904a, 662; Frege, 1904b, 112)

In addition, Frege tells us that such a law is usually specified by an equation:

Our general way of expressing such a law of correlation is an equation in which the letter 'y' stands on the left side whereas on the right there appears a mathematical expression consisting of numerals, mathematical signs, and the letter 'x', e.g. ' $y = x^2 + 3x$ '. (ibid.)

Frege also remarks that with the specification of a law by an equation, "variability has dropped out of sight, and instead generality comes into view, for that is what the word 'law' indicates" (ibid.). One of Frege's conclusions is that the notion of a function has nothing to do with variation, that 'x' does not denote an "indefinite" or "variable" number, but serves to express *generality*.

In Sandu (2005), I considered the notion of an arbitrary correlation in the context of Ramsey's criticism of Russell's notion of a class in *Principia* and his moving away from a predicative notion of a function towards the notion of a *function-in-extension*, that is, an arbitrary correlation between arguments and propositions. The idea was to bring another, indirect evidence to my earlier claim with Hintikka, against Dummett, to the effect that Frege could not have defended the idea of arbitrary correlation, for that would have placed him in the same camp with Ramsey, against Russell. In fact, I thought that Russell's notion of a propositional function stands on the same side, viz., that of a non-standard

interpretation, with Frege's notion of a concept, and in opposition to Ramsey's notion of a function in extension in his *Foundations of Mathematics* (Ramsey, 1925).

All in all, the debate around Frege's and Russell's foundational program (logicism) brought to the forefront a philosophical discussion about the nature of functions and sets. As we see, two opposite conceptions emerged from the debate. According to the so-called *non-standard*, intensional conception, functions and sets are logically secondary, that is, they do not exist independently of the properties, norms, laws (equations), concepts, etc. which glue arguments and values, or members, together, and are logically primary. According to the *standard*, extensional conception, extensions, arbitrary correlations in extension, etc. are primary, or at least make perfectly good sense independently of the laws, concepts, etc. which may introduce them. In the case of functions, we saw that in Frege's case, the above distinction can be spelled out in terms of the more basic distinction between generality and variability.

The present paper contains some further reflections on these matters. I will argue that, depending on the purpose, each of the two notions is needed and makes perfectly good sense. I will show this using two case studies. The first one deals with Ramsey's notion of propositional function in extension needed to show that mathematical truths are *necessities* (*in Wittgenstein sense*.) Ramsey argued that, in the absence of such functions, one can only show, as Russell did, that mathematical truths are *general* but not *necessary*. My presentation will be based extensively on Trueman (2011). The second example will be about the application of mathematics to causal reasoning in empirical sciences. Roughly, the argument will be that for explanatory, counterfactual and predictive reasoning, one needs not only knowledge of the functional dependencies (variability), but also the specification of the underlying laws (equations). My argument will be largely based on Judea Pearl's work on causation.

2. Russell's logicism program

Ramsey criticises Russell's *Principia*'s notion of a propositional function arguing for the need of its extension in the context of the logicist reduction of mathematics to logic. In his logicism program in *Principia Mathematica* (1910-1911), Russell proposed a reduction of mathematics to logic in two steps:

- One first defines mathematical notions in terms of classes
- One then eliminates classes in terms of logical notions

One of the main logical notions Russell makes use of in his program is the notion of *predicative propositional function*:

(*) Only such classes, relations, and functions can be assumed to exist that can be defined by a suitable expression of one's language.

In his criticism of Russell, Ramsey (1925) makes a distinction between two kinds of (logician) reductions of mathematics to logic:

1. The definition of all concepts of a mathematical theory in terms of logical notions; and
2. (1.) plus the derivation of the axioms of the resulting theory from purely logical axioms.

Ramsey points out that a reduction of type (1.) would only show the *generality of mathematics*:

(+) Mathematical truths are the most general truths of the universe.

Let me point out right away that this position is, indeed, what one expects, given Russell's conception of the truths of logic as the most general truths of the universe. However, Ramsey continues, only a reduction of type 2. achieves a proper reduction, that is, a reduction which also shows the *necessity of mathematics: mathematical truths as tautologies in Wittgenstein's sense*. The sense of necessity Ramsey is concerned with here, viz., tautology in Wittgenstein's sense, is that of a sentence true *in every* universe of discourse. It is perhaps worth pointing out that Ramsey anticipates here Carnap's distinction between two kinds of logicist reductions. In Carnap (1931), Carnap points out that Russell operated a reduction of type 1. Most importantly from the vantage point of the present paper, Ramsey observes that in order to perform a reduction of type 2., that is, what he calls a *proper* reduction, one would have to give up the predicative notion of propositional function Russell uses in *Principia*.

I will introduce now Ramsey's notion of a predicative function, Ramsey (1925), which is a reformulation of Russell's corresponding notion. I will rely heavily on Trueman's exposition in Trueman (2011).

The starting point of Ramsey's system is a class of propositions built from a stock of atomic propositions, e.g., 'John is tall', through (possibly an infinite application of) truth-functional connectives. *Atomic predicative functions of individuals* are the result of replacing by variables any of the names of individuals in an atomic proposition, e.g., ' \hat{x} is tall'. This notion is then extended to cover truth-functions of propositional functions and propositions. Ramsey's definition is as follows:

Suppose we have functions $\varphi_1(\hat{x}, \hat{y})$, $\varphi_2(\hat{x}, \hat{y})$, etc., then saying that a function $\psi(\hat{x}, \hat{y})$ is a certain truth-function [...] of the functions $\varphi_1(\hat{x}, \hat{y})$, $\varphi_2(\hat{x}, \hat{y})$, etc. and the proposition p , q , etc., we mean that any value of $\psi(\hat{x}, \hat{y})$, say $\psi(a, b)$, is that truth-function of the corresponding values of $\varphi_1(\hat{x}, \hat{y})$, $\varphi_2(\hat{x}, \hat{y})$, ... etc. i.e. $\varphi_1(a, b)$, $\varphi_2(a, b)$, etc. and the propositions p , q , etc. (Ramsey, 1925, *ibid.* p. 38)

Here are two examples of predicative propositional functions in Ramsey's style: The atomic predicative propositional function ' \hat{x} is tall' maps:

Socrates to the proposition: Socrates *is tall*

Plato to the proposition: Plato *is tall*

Aristotle to the proposition: Aristotle *is tall*

Etc.

The second example is the formula

$$[G(\widehat{x}_1, \widehat{x}_2) \vee p] \vee [H(\widehat{x}_1, \widehat{x}_2) \vee q]$$

which is a complex predicative propositional function. It is a truth-function of $G(\widehat{x}_1, \widehat{x}_2)$, $H(\widehat{x}_1, \widehat{x}_2)$, p and q .

Ramsey emphasizes that the *predicativity of a propositional function*, say ' $F(\hat{x})$ ', consists in the fact that the proposition ' $F(a)$ ' which the propositional function ' $F(\hat{x})$ ' assigns to ' a ', says or predicates the same thing of a as the proposition ' $F(b)$ ' which ' $F(\hat{x})$ ' assigns to ' b ', does of b . I regard Ramsey's

definition of a predicative function as a manifestation of the same phenomenon I discussed earlier in connection with Frege and his specification of a function by a formal law. In Russell's case, as construed by Ramsey, the "law" which keeps the arguments and values together is a predicative propositional function. A good example is our earlier propositional function ' \hat{x} is tall', which maps 'Socrates' to the proposition 'Socrates is tall', 'Plato' to the proposition 'Plato is tall', i.e., 'F(Socrates)' is 'Socrates is tall', 'F(Plato)' is 'Plato is tall', etc.

Ramsey uses his notion of predicative propositional function to give an account of quantification in the *Tractatus*. The proposition ' $\forall xF(x)$ ' is conceived of as the conjunction of all the values of ' $F(\hat{x})$ ', and the proposition ' $\exists xF(x)$ ' as the disjunction of all these propositions. Similarly, the higher-order predicative propositional function ' $\forall \varphi f(\varphi, \hat{x})$ ' is the conjunction of all the values of ' $f(\varphi, \hat{x})$ ' (Ramsey, 1925, p. 40). In addition, the use of quantifiers is governed by what is known as the exclusive interpretation of quantifiers, e.g., ' $\exists xR(x, a)$ ' is conceived of as a disjunction of all the values of ' $R(\hat{x}, a)$ ' except for ' $R(a, a)$ ', and similarly for ' $\forall xR(x, a)$ ', which is conceived of as the conjunction of these values.

Ramsey's reduction 2. of mathematics to logic is achieved in two steps, following Whitehead and Russell. In the first step mathematics is reduced to the theory of classes, e.g., each natural number n is defined as the class of all n -membered classes of individuals. For instance, 1 is defined as the class of all singletons, 2 as the class of all doubletons, etc. In the second step, the theory of classes is reduced to logic. It is here that propositional functions are needed, as class-terms are partially eliminated in favour of propositional functions. As a result of this process, every class is presented as the extension of a propositional function. At this stage Ramsey realised that if the only admissible propositional functions are predicative functions, then there can be no proper reduction of mathematics to logic. As logical truths are tautologies, then the failure of this reduction would also be a failure to show that mathematical truths are tautologies in Wittgenstein's sense.

Here is Ramsey's main argument as presented by Trueman (2011):

If 1 is defined as the class of singletons of individuals and 2 as the class of doubletons of individuals, then the mathematical truth that $1 \neq 2$ requires that there be a singleton or a doubleton: otherwise, 1 and 2 would both be empty and hence identical. If, in turn, the existence of a singleton or doubleton of individuals is to be reduced to logic then, assuming [...][that all

logical truths are tautologies, and vice versa], it must be a tautology that some propositional function is true of exactly one or exactly two individuals. But, if every propositional function is predicati[...][ve] then this is not a tautology, and this is because predicati[...][ve] functions do not logically discriminate between individuals, meaning that it is not contradictory for every individual to satisfy exactly the same predicati[...][ve] functions as every other individual. (Trueman, 2011, p. 296)

The considerations above illustrate the kind of challenge Ramsey faced. If e.g., it is a tautology that some propositional function is true of exactly two individuals, then it must be a contradiction that that predicative function is satisfied by every individual. But it is known that predicative functions do not discriminate between individuals, that is, it is not a contradiction that every individual satisfies the same predicative function as every other individual. This follows from the logical independence of atomic predicative functions, which propagates through the inductive clauses to all predicative functions. The conclusion is that it cannot be a tautology that some predicative function is true of exactly one individual, or exactly two individuals, etc.

3. Logical versus analytical necessity

This section contains a digression on a different notion of necessity than the *Tractarian* one. It has been noticed that the argument establishing that predicative functions do not logically discriminate between individuals at no point appeals to the *Tractarian* assumption that all necessity is logical necessity. We could, for instance, introduce a different notion of necessity, call it ‘analytic necessity’, which is necessity in virtue of meaning. Such a notion of necessity has been considered, among others, by the Finnish logician Erik Stenius (1972). According to it, a statement is analytic if it is true in virtue of the semantic conventions for some of its symbols. Alternatively, a statement is analytic if, according to the semantic conventions for some of its expressions, no state of affairs is a truth restriction for it (that is, no state of affairs makes it false).

The statement ‘If a is red, then a is not green’ symbolised by ‘ $R(a) \rightarrow \neg G(a)$ ’ is an example of an analytic statement in Stenius’s sense. Its truth-table is:

$R(a)$	$G(a)$	$\neg G(a)$	$R(a) \rightarrow \neg G(a)$
T	T	F	F
T	F	T	T
F	T	F	T
F	F	F	T

As we see, this truth-table seems to contain in the first row a truth-restriction in Stenius' sense, that is, ' $R(a)$ ' is true but ' $\neg G(a)$ ' is false (' $G(a)$ ' is true). But for Stenius this truth restriction is not a possible state of affairs because the colours green and red are logically incompatible. Therefore, the first line must be erased as a result of which the table contains no truth-restriction.¹

We witness here a violation of the logical independence of the atomic propositions which is of a different kind than the one we have considered so far: the truth of ' $R(a)$ ' is incompatible with that of ' $G(a)$ ', that is, one and the same individual cannot simultaneously satisfy the propositional functions ' $R(\hat{x})$ ' and ' $G(\hat{x})$ '. Given that ' $R(a) \rightarrow \neg G(a)$ ' is analytic for each a , then so is the statement 'No red object is green' symbolised by ' $\forall x(R(x) \rightarrow \neg G(x))$ ' (which, as we recall, is interpreted as an infinite conjunction).

The failure of logical independence in Stenius' sense leads to a new notion of necessity, *analytical necessity*. Ramsey wanted to show that mathematical statements are logically necessary not in the analytical sense but in the Tractarian sense of truth in all situations (models). To achieve this goal, he needed to give up the kind of logical independence of atomic propositions we considered in the previous section and formulate a notion of propositional function which could discriminate between individuals and would not be grounded in the notion of analytical necessity illustrated in this section. He did that by introducing the notion of *propositional function in extension*. Before discussing it, let me point out that Frege went a different way: although he wanted to show that mathematical (arithmetical) statements are reducible to logic, he did not conceive of logical statements as necessary in the Tractarian sense. For Frege, like for Russell, true logical statements are the most general statements about the universe of discourse.

¹ The method of truncated truth-tables that Stenius uses here originates actually with G. H. von Wright who, in his *Essays in Modal Logic* (1951), uses it to give a semantical interpretation of modal logic.

4. Ramsey's propositional functions in extension

Ramsey needed, then, a new notion of a propositional function which would allow him to distinguish between individuals. In order to do this, he needed to extend the notion of a function to cover also non-predicative propositional functions, where predicativity is understood as above. Here is how Ramsey expresses himself:

The only practicable way to do it as radically and drastically as possible; to drop altogether the notion that $\phi(a)$ says about a what $\phi(b)$ says about b , to treat propositional functions like mathematical functions, that is, to extensionalise them completely. Indeed it is clear that, mathematical functions being derived from propositional, we shall get an adequate extensional account of the former only by taking a completely extensional view of the latter. (Ramsey, 1925, p. 52)

Ramsey could not see a way to define explicitly a function in extension, and for this reason he contented himself to explain this notion rather than define it. His explanation is given in terms of the notion of arbitrary correlation, that is, a relation in extension between propositions and individuals, which associates to each individual a unique proposition. In specifying the nature of this correlation, he remarks that it may be “practicable or impracticable” (*ibid.*). I take this to be another way for Ramsey to say that the correlation is not determined by a formal law, but is just an arbitrary association between individuals and propositions.

For instance, the predicative propositional function, ‘ \hat{x} is wise’ makes the following correlations between arguments and values

Socrates \mapsto *Socrates is tall*
Plato \mapsto *Plato is tall*
Etc.

whereas a non-predicative propositional function would be just an arbitrary correlation, e.g.,

Socrates \mapsto *Socrates is tall*
Plato \mapsto *Plato is a Greek philosopher*
Aristotle \mapsto *Socrates is not tall*
Etc.

Ramsey uses propositional functions in extension to define identity:

$$x = y =_{df} \forall \Phi_e [\Phi_e(x) \equiv \Phi_e(y)]$$

where Φ_e ranges over propositional functions in extension (recall the substitutional definition of the universal quantifier).

We notice that when a and b are the same individual, then ' $a = b$ ' is the conjunction of ' $p \equiv p$ ', ' $q \equiv q$ ', ..., which is a tautology. In other words, the extension of the above propositional function is the class of pairs of individuals (a, b) such that a is identical with b . On the other side, when a is distinct from b , then there is a propositional function ' Φ_e ' such that $\Phi_e(a)$ is p while $\Phi_e(b)$ is $\neg p$ (see our example above). In this case $a = b$ is a conjunction of propositions which includes $p \equiv \neg p$, that is, a contradiction. In other words, none of the pairs (a, b) such that a is distinct from b satisfies the above propositional function. Notice that in this case, the proposition p that the propositional function ' Φ_e ' assigns to a does not say of a what the proposition $\neg p$ that the propositional function ' Φ_e ' assigns to b , says of b . Here we may take a and b to stand for Socrates and Aristotle, respectively, and p and $\neg p$ to stand for 'Socrates is tall' and 'Socrates is not tall', respectively.

After having defined identity, Ramsey introduces set-theoretical notions, using the above non-predicative notion of identity. For singletons he uses the propositional function ' $\hat{x} = a$ ', where a is an arbitrary individual and identity is understood as above. Then ' $a = a$ ' is a tautology, and for any distinct b , ' $b = a$ ' is a contradiction. In other words, the extension of the propositional function ' $\hat{x} = a$ ', is the singleton $\{a\}$. Hence it is a tautology that *some* propositional function is true of exactly one individual. Similarly, the propositional function ' $\hat{x} = a \vee \hat{x} = b$ ' defines the doubleton $\{a, b\}$, i.e., its extension is the class $\{a, b\}$.

Ramsey's notion of propositional function in extension contributed substantially to the extensionalist conception of sets/functions which in turn led to the notion of *arbitrary set* and *arbitrary function* and to the so-called *standard interpretation* of a set/function variable. On the other side, Frege's and Russell's notion of a function, which require a law (equation) to "glue" arguments and values together, was a typical example of an intensional interpretation, which gradually led to a *non-standard interpretation* of set/function variables. Here is a more recent illustration from the model-theoretical tradition.

Recall Ramsey's definition of identity using functions in extension. Here is the same definition in second-order logic (set-theory) with a standard interpretation:

$$x = y =_{df} \forall X [X(x) \leftrightarrow X(y)]$$
$$x = y =_{df} \forall f [f(x) = f(y)]$$

where $X(f)$ is a (second order) variable ranging over sets (functions). Notice that the only pairs which are in the extension of $x = y$ are (a, b) with a the same individual as b . For if there were a pair (a, b) which satisfied $x = y$ such that a is distinct from b , then the singleton $\{a\}$ would falsify $\forall X[X(x) \leftrightarrow X(y)]$. Given the standard interpretation, the set $\{a\}$ exists. An analogue argument, which strongly resembles Ramsey's argument which used functions in extension, holds for the second-order definition of identity using characteristic functions: if there was (a, b) which satisfied $x = y$ such that a is distinct from b , then the function f such that $f(a) = 0$ and $f(b) = 1$ would falsify $\forall f [f(x) = f(y)]$.

5. Arbitrary functions: How does one individuate them?

In the present section, I will consider two objections against the notion of propositional functions in extension. They are discussed in Trueman (2011). One of them is due to Sullivan, the other to Trueman himself based on some remarks made by Wittgenstein. Both are reactions to the failure to give a satisfactory answer to the question 'How does one individuate (apprehend) propositional functions in extension?' The answer to this question obviously has an impact on the size of the class of functions mathematicians should be concerned with. It is known that Weierstrass considered the idea of arbitrary function but rejected it as being so vague and so general as not to enable us to find any interesting things to say about it; in the same spirit, Kronecker rejected the idea of an arbitrary function altogether.

Sullivan's (1995), objection to propositional functions in extension is that, unlike their predicative relatives, which need only one value for their individuation, they need all their values to be individuated. It is not difficult, intuitively, to see why this is so. Take any argument and consider the proposition which is the value of the propositional function for that argument. By deleting the argument, you can recover the propositional function. To take an example, if

' $F(\hat{x})$ ' is a propositional function and you know that ' $F(\text{John})$ ' is 'John is tall', then you also know that ' $F(\text{Peter})$ ' is 'Peter is tall', etc. On the other side, if φ_e is a propositional function in extension, then if you know that ' $\varphi_e(\text{John})$ ' is 'Paris is beautiful', you cannot infer anything about ' $\varphi_e(\text{Peter})$ '. This property which Sullivan calls *containment*, makes predicative propositional functions much more accessible ("practicable") than their extensional relatives.

Sullivan is not entirely correct and Trueman (2011, Sect. 4) gives an example of a propositional function which shows Sullivan's argument to be invalid, that is, the function needs all its arguments to be individuated. I will ignore it and take Sullivan's argument to be approximatively correct (it actually holds, as Trueman points out, only for atomic predicative propositional functions.) Even if the property of containment holds only for a proper subclass of functions, it still explains why these functions are more accessible than their extensional relatives and how our conceptual system can somehow integrate and manipulate potentially infinite correlations of arguments and values. And in the absence of properties like containment or other mechanisms which perform a similar function, the questions still remains: How does one individuate functions in extension, especially the infinitary ones?

Trueman discusses another property that predicative functions have and extensional functions lack: *substitution*. This is an expected move once we recall that we needed substitution to explain the notion of atomic predicative propositional function in the first place. Recall that such a function is the result of replacing by variables any of the names of individuals in an atomic proposition. For example, the atomic predicative propositional function ' \hat{x} is wise', is obtained from the proposition 'Socrates is wise' after substituting the variable ' \hat{x} ' for the name 'Socrates'. Thus, one can understand the notion of mapping in terms of the operation of substitution, e.g., the mapping of 'Socrates' to 'Socrates is wise', is explained in terms of the substitution of 'Socrates' for ' \hat{x} ' in ' \hat{x} is wise', etc. The sense in which non-atomic predicative functions "map" names to propositions is explained analogously. It is quite clear that substitution, as an effective operation on expressions in an underlying language, explains the property of containment and provides a criterion for the individuation of predicative functions. And it should be equally clear that substitution cannot ground the notion of mapping that underlies propositional functions in extension. One has to try something else.

Here is one possibility which comes more easily to one's mind. Recall our last example, where we noticed that the operation of substitution generates a table:

$F(\hat{x})$	
'Socrates'	'Socrates is wise'
'Plato'	'Plato is wise'
....

One could follow the same idea and introduce propositional functions in extension by tables, e.g.,

$F_e(\hat{x})$	
'Socrates'	'Socrates is wise'
'Plato'	'Socrates is not wise'
...	...

As perhaps expected, this suggestion does not solve the problem we are concerned with. For the whole matter of dispute is how to understand the relation between the names on the left side, and the corresponding propositions on the right side. In this connection, Trueman refers to an argument by Wittgenstein (1969, part II, Chap. 16) who points out that the name 'Socrates' appears on the left side of the table only to *direct* us to a line in a table. We could have actually used any signs we liked: numerals, letters, or squares of colour, etc. The fact that we chose to mark each line of this table with strings which look like the names of Socrates and Plato, Wittgenstein continues, should not mislead us into thinking that they are those names. This introduces some further complications when we try to combine in one and the same expression predicative propositional functions and propositional functions in extension. For instance, in the conjunction ' $F(\text{Socrates}) \wedge F_e(\text{Socrates})$ ', there is a difference between the first and second occurrence of the string 'Socrates', the first is an occurrence of the name of Socrates but the latter is not.

Wittgenstein's remarks sound surprisingly similar to Frege's criticism of the extensional notion of a set and of the individuation of sets through their members. Here is what Frege writes in "Booles rechnende Logik und die Begriffsschrift" (1983, vol. I, p. 38; 1979, p. 34):

But it is surely a highly arbitrary procedure to form concepts merely by assembling individuals, and one devoid of significance for actual thinking unless the objects are held together by having characteristics in common. It

is precisely these which constitute the essence of the concept. Indeed one can form concepts under which no object form, where it might perhaps require lengthy investigation to discover that this was so. Moreover, a concept, such as that of number, can apply to infinitely many individuals. Such a concept would never be attained by logical addition. Nor finally may we presuppose that the individuals are given in toto, since some, such as e.g. the numbers, are only yielded by thought.

One possibility though which arises in connection with Wittgenstein's remarks is to interpret them in a completely different way, as I will do here: one can see the table only as a device for making sure that to each argument there is a corresponding line. When doing this, one abstracts from the *nature* of the connection between arguments and values. Recalling Frege's words, with the specification of a law by an equation, "variability has dropped out of sight, and instead generality comes into view". The present conception goes in the reverse direction, that is, from generality to functional dependency ("variability"). As we observed in Hintikka and Sandu (1992, p. 305), this conception made possible the gradual disentanglement of the notion of a set from intensional ingredients such as concepts, properties, etc., and laid the ground for the definition of sethood in an alternative way. In a parallel development, the modern notion of arbitrary function emerged through the gradual disentanglement of the notion of correlation from Fregean concepts, equations and other formal rules, or from requirements like predicativity. Ramsey's notion of a propositional function in extension is one step in this process of emancipation. Modern logic has developed Ramsey's idea and taken functional dependencies as arbitrary correlations between values and arguments. This move is important for it allows, among other things, to talk about *all* functions and to study all the properties they have in common. Here is one account which illustrates this, originally due to Hodges (1997) and further developed by Grädel and Väänänen (2013), among others.

6. Arbitrary correlations: Dependence logic

Functional dependencies are not definable in ordinary first-order logic. To make them definable, we introduce new logical constants in the object language and give them meaning through semantical clauses. More specifically, the syntax of first-order logic is extended with an infinite number of atomic formulas of the form

$=(x_1, x_2, \dots, x_n; y)$ intended to express arbitrary functional dependence: the (values of the) variables x_1, x_2, \dots, x_n totally (functionally) determine (the value of) y . Alternatively but equivalently: y is a function of x_1, x_2, \dots, x_n . Such an atom is interpreted in a model by a set X of (partial) assignments in the universe of the model (hence the higher-order source of the definition). The corresponding semantical clause is very intuitive and it is given in terms of a set of assignments (hence the second-order interpretation). Let X be a set of assignments to the variables x_1, x_2, \dots, x_n and y . Then:

(+) X makes the formula $=(x_1, x_2, \dots, x_n; y)$ true if and only if for any two distinct assignments s and s' in X the following holds: whenever s and s' agree on the values of the variables x_1, x_2, \dots, x_n , then they also agree on the values of y .

This semantical definition is just the definition of the notion a functional correlation in purely extensional terms, without appealing to any particular relation or property between an argument and its value. Here is an example. Let X consists of the following four assignments:

X	x_1	x_2	x_3	y
s_1	2	1	1	3
s_2	2	1	3	3
s_3	1	1	1	4
s_4	3	2	1	2

It can be checked that X makes true the formula $=(x_1, x_2; y)$ but not the formula $=(x_2, x_3; y)$. Notice that for the second claim, $s_1(x_2) = s_3(x_2)$ and $s_1(x_3) = s_3(x_3)$ but $s_1(y) = 3 \neq 4 = s_3(y)$.

Thus, returning to Frege’s remarks “In the equation ‘ $y = x^2 + 3x$ ’...variability has dropped out of sight, and instead generality comes into view, for that is what ‘law’ indicates”, in dependence logic we abstract from the form of the function and we represent only variability, which is technically encoded by (a functional) dependence atom: $=(x; y)$

7. Functional equations and counterfactuals

In the previous sections we have run through some basic differences between, on one side, functions in extension that we represented by dependence atoms $=(x_1, x_2, \dots, x_n; y)$ which represent explicitly which variables affect which, and, on the other side, functions specified by laws (e.g., Ramsey's predicative propositional functions). The latter are standardly represented by equations, which allows, as Frege pointed out, to express generality. We also discussed some arguments to the effect that laws account for substitution and containment, which, in turn, make the functions so introduced accessible ("practicable").

In the remainder of the paper, I will discuss one more difference between the two types of functions, which, to the best of my knowledge, has never been addressed in the literature. It appears in the application of functional equations to experimental science and causal reasoning, in particular. It operates in the reverse direction to the one considered by Ramsey. Ramsey showed that if one wants to reduce mathematics to logic and show mathematical truths to be necessities in Wittgenstein sense, then predicative functions are not enough but functions in extension are also needed. In the present case the argument is that if one needs to represent certain forms of causal and counterfactual reasoning, functions in extension (functional dependencies) are not enough, but one needs to specify also the form of the function, that is, the underlying law (equation). As we see, in Fregean terms the direction here is from variability (dependency) to generality.

As a starting point, let me mention that in the applications considered here, the variables x_1, x_2, \dots, x_n, y represent physical magnitudes, states or events, which may take different values, as e.g., in the statement:

(++) Weather depends on air pressure and air temperature.

We have here an example of causal dependence between three physical magnitudes, weather, air pressure and temperature. The example is taken from Abramsky, Kontinen, Väänänen, and Vollmer (2016).

In Pearl and McKenzie (2018) the authors introduce what they call 3 *ladders of causation*:

1. *Associations*: correlations, dependencies
2. *Interventions*: actions and their consequences

Sandu

3. *Imagining* (reasoning about hypothetical situations): Things that could happen, were certain actions to be performed.

According to the authors, at ladder 1 we are confronted with queries of the form

4. Given that a patient is sick ($Sick = 1$), would he die ($Dies = 1$), if he is not receiving treatment ($Treated = 0$)?

and their probabilistic variants

5. Given that a patient is sick ($Sick = 1$), what is the probability that he dies ($Dies = 1$), if he is not receiving treatment ($Treated = 0$).

We are told that such queries, which involve states and events represented by Boolean variables, are usually answered by *collecting data* and analysing them using *conditional probabilities*. Pearl and McKenzie represent answers to such queries by conditional probability statements, $\Pr(A/B) > t$, (the probability of A conditional on B). Here I shall use, instead, a more logical notation, $B \supset \Pr(A) > t$, where \supset is a kind of indicative conditional and the interpretation of $B \supset \dots$ is, roughly, “if B happens (is observed), then the probability of the event A is greater than t ”.

At ladder 2 we encounter queries like

6. Would the agent recover, if we start treating him?

and its probabilistic variants

7. What is the probability that a patient recovers, if we start treating him?

In this case one evaluates the impact on the patient of the action (*intervention*) of treating him or her. The intervention forces a change in the state of the patient from not being treated, $Treated = 0$ to being treated, $Treated = 1$. In our logical notation we can represent answers to such queries by:

$$X = x \mapsto \alpha$$

which represents what we shall call an *interventionist counterfactual*, where the antecedent indicates that the value of the state denoted by X is surgically set to the value x and the consequent α represents the consequence of this intervention. (Pearl, Glymour, Jewell, 2016.)

Finally at level 3, we encounter more complex statements, such as

8. A subject who died after treatment would have recovered had he not be treated

and its probabilistic variants

9. The probability that a subject who died after treatment would have recovered had he or she not been treated is ε .

Thus in (8), we first make an observation (“the subject died after treatment”), followed by an intervention (changing the state of the subject from treated to non-treated), after which we observe the consequence of this action. In our logical notation we can render the first statement by:

$$(Dies = 1 \wedge Treated = 1) \supset (Treated = 0 \mapsto Dies = 0).$$

More generally, statements at this level have the general form:

$$\alpha \supset (X = x \mapsto \beta)$$

which more perspicuously renders the logical embedding of observations and counterfactuals whose antecedents are interventions.

Now Pearl and McKenzie make an important claim:

- To evaluate a counterfactual statement at level 3 (i.e., statements of the form $\alpha \supset (X = x \mapsto \beta)$), one needs to have *knowledge of the specific functional form* of the laws (equations).
- To evaluate the probabilities of counterfactuals at level 2, it is enough to know which variables affect which. This can always be computed from observational data by mere knowledge of the *causal graph*.

Let me illustrate shortly what is going on here with a simple example. Consider the following table (called a *team* in the literature) T which consists of 3 assignments to the variables X, Y, Z and W :

	X	Y	Z	W
$T:$	0	2	1	1
	1	1	0	1
	1	0	0	0

Several functional correlations hold in the team, that is, the team T makes true the following dependence atoms (among others):

$$= (Y; Z), = (X, Y, Z; W).$$

We can further study these correlations with the help of conditional probabilities, e.g., the probability that $W = 1$ conditional on $X = 1$, ($X = 1 \supset (\Pr(W = 1) = 0,5)$), but this will not be our concern here.

At the next step we select (designate) some functional dependencies. They are marked with arrows:

	X	Y	Z	W
$T:$	0	2	1	1
	1	1	0	1
	1	0	0	0

For instance, the arrow from Y to Z indicates that Z depends functionally on Y , and the arrows from X, Y, Z to W indicate that W depend functionally on X, Y and Z . The arrows form an (acyclic) *graph* and are consistent with the assignments in the team. In other words, the previous dependence atoms $= (Y; Z), = (X, Y, Z; W)$ continue to be true in the team, but there is now explicit information about the functional dependencies associated with the team.

We can now take one further step and describe explicitly the *equations* behind the designated functional dependencies. These are called *structural equations* in the literature. For instance, one way to do so is:

$$\begin{cases} Z = 1, \text{ if } Y = 2, \text{ and } Z = 0, \text{ otherwise} \\ W = X + Z, \text{ if } Y \neq 0, \text{ and } W = 0, \text{ otherwise} \end{cases}$$

It may be checked that these equations are consistent with the assignments of the team.

What does the specification of the equations add? Pearl and McKenzie’s answer is that it allows the expression of *interventionist counterfactual reasoning*. Here is a simple-minded example just to help fixing the intuition. Consider the statement:

10. If X were 1 and Y were 2, then W would be 2 ($X = 1 \wedge Y = 2 \mapsto W = 2$).

To establish the truth of this statement, the graph of dependencies, that is, the functional dependencies $= (Y; Z), = (X, Y, Z; W)$ are not enough, but one needs the specification of the functions (‘structural equations’; strictly speaking, this is not true, the structural equations are needed for the probabilistic variants of (a)). Indeed, once we have the two structural equations mentioned above, we can run the following computation: if $Y = 2$, then by the first equation, $Z = 1$. From the second equation we now get $W = X + Z = 1 + 1 = 2$. As we see and as Frege observed, the equations ensure “generality”, for they cover cases not covered by the observations encoded in the Table.

8. Conclusions

We have argued that Ramsey’s notion of a propositional function in extension was one step in the process of emancipation of sets and functions from intensional ingredients such as concepts, properties, laws, etc. Modern (model-theoretic) logic has developed Ramsey’s idea. According to this conception, the crucial aspect of an arbitrary correlation is the idea of co-variability (dependence). Ramsey needed arbitrary sets and arbitrary correlations to show that mathematics is reducible to logic, which in turn shows that mathematical truths are not only general but also necessary (“in Wittgenstein’s sense”). Let me point out in this connection that

functions in extension were also needed for other foundational (“completeness”) results. For instance, it is known that only if we assume the standard interpretation, we can formulate descriptively complete and categorical axiomatizations of e.g., number theory and the theory of real numbers (e.g., Dedekind’s theorem). And a result in Ramsey’s spirit, only if we assume the standard interpretation, we can show that higher-order logics are reducible to second-order logic (Hintikka, 1955). These results should not be surprising. Think, for instance, about the (semantic) completeness theorem which states that every (formally) consistent set of sentences has a model. In the Henkin type completeness, one needs many models to match a formally consistent sets of sentences. Obviously not all of them would be definable in the underlying language.

We saw that one of the problems with arbitrary correlations (functions in extension) is their “impracticability”. Several philosophers have rejected them on the ground that they are not epistemically accessible. We tried to circumvent this problem by arguing that one can still investigate the logical properties of functional dependencies as done recently in *Dependence Logic* where partial axiomatization of the notion of dependence has been given. The problem, however, is that the “impracticability” steps in through the backdoor: the set of logical truths of dependence logic is non-recursively axiomatizable. Väänänen (2012) describes a way out of this predicament using a notion of internal categoricity but we shall not discuss this problem here.

In the second part of the paper, we were concerned with applications of mathematics to empirical sciences (causal explanation, counterfactual reasoning, and predictions). The counterfactual reasoning and the notion of causation it sustains was modelled by changes in the values of certain variables. The strategy here was completely opposite to the one in the first part of the paper. We mention an argument due to Pearl according to which for those specific purposes, one needs not only knowledge of the functional dependencies (i.e., the graph), but also the form of the structural equations themselves. This result should not be surprising either: the examples we gave make it clear that the underlying counterfactual reasoning and the notion of causation it sustains, requires not only a specification of the variables which co-vary but also with *how much*.

University of Helsinki

References

- Abramsky & Kontinen, J. Väänänen, H. Vollmer (eds.), 2016, *Dependence Logic. Theory and Applications*.
- Carnap, R. 1931, Die logizistische Grundlagen der Mathematik, *Erkenntnis* 2, 91–105 (1931).
- Dummett, M. 1973, *Frege: Philosophy of Language*, Duckworth, London.
- Dummett, M. 1991, *Frege. Philosophy of Mathematics*, Duckworth, London.
- Frege, G. 1904a, Was ist eine Funktion?, in *Festschrift Ludwig Boltzmann gewidmet zum sechzigsten Geburtstag, 20. Februar 1904*, ed. by S. Meyer (J. A. Barth, Leipzig, 1904), pp. 656–666.
- Frege, G. 1904b, *Translations from the Philosophical Writings of Gottlob Frege*, ed. by P. Geach, M. Black, Basil Blackwell, Oxford, 1960.
- Frege, G. 1983, *Nachgelassene Schriften und Wissenschaftlicher Briefwechsel* (Meiner, Hamburg, 1969–1976) (2 vols; 2nd edn. of, vol. 1, 1983). Vol. 1, *Nachgelassene Schriften*, ed. by H. Hermes, F. Kambartel, F. Kaulbach; vol. 2, *Wissenschaftlicher Briefwechsel*, ed. by G. Gabriel, H. Hermes, F. Kambartel, C. Thiel, A. Veraar.
- Frege, G. *Posthumous Writings* (Basil Blackwell, Oxford, 1979). English Translation of vol. 1 of Frege 1983.
- Grädel, E. & J. Väänänen, 2013, ‘Dependence and independence’, *Studia Logica* 101, 399–410.
- Hintikka, J. & Sandu, G. 1992, ‘The skeleton in Frege’s cupboard: the standard versus nonstandard Distinction’ *J. Philos.* 89, 290–315.
- Hodges W., Compositional semantics for a language of imperfect information, *Logic Journal of the IGPL* 5 (1997) 539–563.
- Pearl, J. Glymour, M. & Jewell, N. P. *Causal inference in statistics: a primer*, John Wiley & Sons, 2016.
- Pearl, J. & McKenzie, 2018, *The Book of Why. The New Science of Cause and Effect*.
- Ramsey, F. P. 1925, ‘The foundations of mathematics’, *Proc. Lond. Math. Soc. Ser. 225* (Part 5), 338–384 (1925). Re-edited in F.P. Ramsey, *The Foundations of Mathematics* (Routledge and Kegan Paul, London, 1931), pp. 1–61. [We refer to this edition.]
- Sandu, G. 2005, ‘Ramsey and the notion of arbitrary function’, in *Critical Reassessment*, ed. by M.J. Frápolli, M. F.P. Ramsey, *Continuum*, London, 2005, pp. 237–256.
- Stenius, E. 1972, *Critical Essays*, North-Holland Publishing Corporation, Amsterdam.
- Trueman, R. 2011, Propositional Functions in Extension, *Theoria*, vol. 77, pp. 292–311.
- Wittgenstein, L. 1969, *Philosophische Grammatik* (Blackwell, Oxford, 1969), Herausgegeben von, R. Rhees.
- Väänänen, V. 2012, Second order logic, set theory and foundations of mathematics, in *Epistemology versus Ontology: Essays on the Philosophy and Foundations of Mathematics in Honour of Per Martin-Löf*, ed. by P. Dybjer, S. Lindström, E. Palmgren, G. Sundholm (Springer, Dordrecht, 2012), pp. 371–380.

Tabula Gratulatoria

Professor Valtteri Arstila is congratulated on 13. XII 2024 by:

niin & näin -lehti

Suomen tieteellisen metafysiikan seura

Hanne Appelqvist

Adrian Bardon

Ned Block

Alexander D. Carruth

Matti Eskelinen

Dmitri Filimonov

Tuukka Franck

Akiko Frischhut

Karl Friston

Heidi Haanila

Leila Haaparanta

Jani Hakkarainen

Mirja Hartimo

Santtu Heikkinen

Visa Helenius

Ilpo Hirvonen

Mika Hämäläinen

Samuli Isotalo

Roomet Jakapi

Haeran Jeong

Johanna K. Kaakinen

Marja-Liisa Kakkuri-Knuuttila

Sakari Kallio

Santtu Kauppila

Julian Kiverstein

Michal Klincewicz

Polaris Koi

Olli Koistinen

Mika Koivisto

Ville V. Kokko

Anssi Korhonen

Tapio Korte

Inkeri Koskinen

Eerik Lagerspetz

Hemmo Laiho

Maria Lasonen

Veikko Launis

Enyu Lin

Tabula Gratulatoria

Dan Lloyd

Kristjan Loorits

Vili Lähteenmäki

Paavo Maanpää

Sanna Mattila & Peter Myrdal

Mikko Metsälampi

Tuuli Mäkitie-Drechsler

Bruno Mölder

Riikka Möttönen

Ilkka Niiniluoto

Valdas Noreika

Markku Oksanen

Tuomas K. Pernu

Henri Pettersson

Kalle Pihlainen

Sami Pihlström

Piret Pöldver

Vivian Puusepp

Paavo Pyökkänen

Panu Raatikainen

Henry Railo

Antti Revonsuo

David Rosenthal

Juha Räikkä

Lassi Saario-Ramsay

Hanna-Mari Salonen

Nils Sandman

Gabriel Sandu

Anil Seth

Helena Siipi

Jani Sinokki

Avril Styrman

Tuukka Tanninen

Pii Telakivi

Aksel Tigerstedt

Tuukka Tomperi

Jarno Tuominen

Susanne Uusitalo & Sauli Mäkimattila

Matti Vaittinen

Katja Valli

Jyrki Vanhakartano

Markus Weckström

Saara Wuokko

Mikko Yrjönsuuri