EUROPEAN JOURNAL OF ANALYTIC PHILOSOPHY

UDC 101 | ISSN 1845-8475

Vol. 20, No. 1, 2024

Research Articles

ARBITERS OF TRUTH AND EXISTENCE, **Nathaniel Gan** | BOOK REVIEW UMBERTO GALIMBERTI L'ETICA DEL VIANDANTE, **Martina Blečić** | IS L.A. PAUL'S ESSENTIALISM REALLY DEEPER THAN LEWIS'S?, **Cristina Nencha** | THE LOGICAL POSSIBILITY OF MORAL DILEMMAS IN EXPRESSIVIST SEMANTICS: A CASE STUDY, **Ryo Tanaka** | TWO PROBLEMS ABOUT MORAL RESPONSIBILITY IN THE CONTEXT OF ADDICTION, **Federico Burdman** | NONTRIVIAL EXISTENCE IN TRANSPARENT INTENSIONAL LOGIC, **Miloš Kosterec** | CAN WE DEFEND NORMATIVE ERROR THEORY?, **Joshua Taccolini** | BETTER-MAKING PROPERTIES AND THE OBJECTIVITY OF VALUE DISAGREEMENT, **Erich H. Rast** | INTEGRATIVE BIOETHICS: A BLIND ALLEY OF EUROPEAN BIOETHICS, **Tomislav Bracanović** | ARE COMPOSITE SUBJECTS POSSIBLE? A CLARIFICATION OF THE SUBJECT COMBINATION PROBLEM FACING PANPSYCHIST, **Siddharth S** | IS KINDNESS A VIRTUE?, **Kristján Kristjánsson**

EUROPEAN JOURNAL OF
ANALYTIC PHILOSOPHY

UDC 101
ISSN (Print) 1845-8475
ISSN (Online) 1849-0514
https://doi.org/10.31820/ejap

Editor-in-Chief
Marko Jurjako
University of Rijeka, mjurjako@ffri.uniri.hr

Associate editors
Elisabetta Lalumera
University of Bologna

Lovro Savić
University of Oxford

Editorial administrators and proofreading
Mia Biturajac
Institute of Philosophy, Zagreb

Iva Martinić
University of Rijeka

Publisher, Editorial office

University of Rijeka, Faculty of Humanities and Social Sciences, Department of Philosophy
Address: Sveučilišna avenija 4, 51000 Rijeka, Croatia (EU)
Phone: +385 51 669 794
E-mail: eujap@ffri.uniri.hr
Web address: https://www.eujap.uniri.hr
Printed by Grafika Helvetica d.o.o., Rijeka (Croatia)

The publication of the journal is financially supported by the Ministry of Science and Education of the Republic of Croatia.

TABLE OF CONTENTS

ARTICLES

Arbiters of Truth and Existence
Nathaniel Gan ……….……………………………………………………1-24

BOOK REVIEW Umberto Galimberti L'ETICA DEL VIANDANTE
Martina Blečić ………………..……………………………….……….25-29

Is L.A. Paul's Essentialism Really Deeper than Lewis's?
Cristina Nencha……………………..……………………………..31-54

The Logical Possibility of Moral Dilemmas in Expressivist Semantics: A Case Study
Ryo Tanaka …..……………..…………………………………………55-86

Two Problems About Moral Responsibility in The Context of Addiction
Federico Burdman.…..……….……………………………..……….…87-112

Nontrivial Existence in Transparent Intensional Logic
Miloš Kosterec …..………..…………………………...………………113-131

Can We Defend Normative Error Theory?
Joshua Taccolini …..………..…………………………..…………..…133-156

Better-Making Properties and the Objectivity of Value Disagreement
Erich H. Rast …..………..…………………………..………….……157-181

Integrative Bioethics: A Blind Alley of European Bioethics
Tomislav Bracanović …….……………………………..…………..……183-207

Are Composite Subjects Possible? A Clarification of the Subject Combination Problem Facing Panpsychist
Siddharth S …..………..…………………………..…………..………209-234

Is Kindness a Virtue?
Kristján Kristjánsson…….……………………………..…………..……235-254

ABSTRACTS (SAŽECI) ….….……………………………………..(AB)

AUTHOR GUIDELINES
AND MALPRACTICE STATEMENT …….………………………………(AG)

European Journal of Analytic Philosophy EuJAP | 2024 | Vol. 20 | No. 1 | 1-23
https://doi.org/10.31820/ejap.20.1.1

ARBITERS OF EXISTENCE AND TRUTH

Nathaniel Gan[1]

[1] National University of Singapore, Singapore

Original scientific paper – Received: 24/06/2023 Accepted: 04/01/2024

ABSTRACT

Call the epistemological grounds on which we rationally should determine our ontological (or alethiological) commitments regarding an entity its arbiter of existence (or arbiter of truth). It is commonly thought that arbiters of existence and truth can be provided by our practices. This paper argues that such views have several implications: (1) the relation of arbiters to our metaphysical commitments consists in indispensability, (2) realist views about a kind of entity should take the kinds of practices providing that entity's arbiters to align with respect to their metaphysical dependencies, (3) if realists take a kind of practice to provide grounds on which to affirm the existence of a kind of entity, they should turn to those same grounds when seeking to provide an epistemology of the relevant domain.

Keywords: naturalism; Carnapian realism; indispensability arguments; epistemic problems.

Correspondence: nat_gan@hotmail.com

Introduction

Call the epistemological grounds on which we should rationally hold (or withhold) ontological commitments to a kind of entity that entity's *arbiter of existence*. Roughly, an entity's arbiter of existence provides the primary reasons for which we should affirm or deny its existence.

Independently of whether we affirm or deny the existence of a kind of disputed entity, we might also be interested in affirming or denying sentences that syntactically appear to ascribe properties to those entities. Call the epistemological grounds on which we should hold (or withhold) alethiological commitments to such sentences the *arbiter of truth* of the relevant entities.[1] If we affirm or deny claims regarding what an entity is like, these affirmations and denials should rationally be justified with reference to that entity's arbiter of truth.

Three questions might be raised regarding arbiters of existence and truth:

(a) What provides an entity's arbiters of existence and truth?
(b) How do these arbiters inform our ontological and alethiological (non-)commitments?
(c) How are an entity's arbiters of existence and truth related?

According to some popular approaches to ontology, we can answer (a) by noting that some things we do—that is, some of our *practices*—are epistemologically privileged when it comes to our metaphysical commitments, and can provide arbiters of existence and truth. For instance, scientific naturalists hold that science, broadly speaking, should provide the epistemological grounds for many of our metaphysical commitments (Armstrong 1968; Quine 1951, 1963). Hence, scientific naturalists take our scientific practices to be able to provide arbiters. Carnap and his followers, alternatively, hold that with some disputed entities, our ontological commitments should correspond to the existential statements that meet the acceptability standards of our

[1] To be precise, the class of sentences governed by an entity's arbiter of truth should be limited to just the property-ascription sentences that do not merely make claims about the existence of the entities in question—our attitude toward 'Phlogiston exists' should be governed by phlogiston's arbiter of existence, not its arbiter of truth. Also, this class should be delineated based on whether the relevant sentences have the *syntactic* structure of property-ascription sentences—our attitude toward 'Phlogiston has negative mass' should be governed by phlogiston's arbiter of truth, even if we wish to adopt a semantics under which 'Phlogiston has negative mass' does not actually ascribe properties to phlogiston. For brevity, this paper will refer to such sentences as sentences about the nature of a kind of entity, but this should not be taken to presuppose the existence of those entities. Thanks to a reviewer for pressing for clarity on this point.

discourse (Carnap 1950; Thomasson 2014). Thus, Carnapians take our discursive practices to be able to provide arbiters of existence.

This paper will describe a general framework for views that take our practices to provide arbiters and discuss the implications of holding such views (remaining neutral on whether such views are in fact right). §1 will flesh out the above answer to (a) by describing more precisely what it would mean for some of our practices to be privileged in the relevant sense. We will also consider how this answer to (a) bears on (b). It will be argued that when our practices provide an entity's arbiters of existence and truth, the relation between those arbiters and our metaphysical commitments consists in indispensability. Namely, we can delineate our metaphysical commitments regarding that entity by considering the privileged practices to which that entity and sentences about it are indispensable. §2 then turns to (c), arguing that if we accept a realist commitment regarding a kind of entity, and the epistemological grounds for that commitment are provided by our practices, the arbiters of existence and truth for those entities should align in some way.

§§3–4 explore implications of these results for some attempts to separate arbiters of existence and truth. Let *metaphysical realism* about a kind of entity be a view that affirms the (objective, mind-independent) existence of those entities, and *semantic realism* about a kind of entity be a view that affirms the truth of some sentences concerning the nature of those entities.[2] §3 considers views that hold semantic realism about a kind of entity without committing to either metaphysical realism or metaphysical anti-realism about those entities. Such views have been considered regarding mathematics (Dummett 1979; Putnam 1979), ethics (Ridge 2019; Sayre-McCord 1986) and science (Devitt 1991; Leplin 1984). Views like these seem to take the arbiters of existence and truth for the entities in question to be somewhat independent, such that we can justify an alethiological commitment to the relevant sentences while remaining neutral on the ontological aspect. §3 argues that if proponents of such views have adequate grounds on which to hold semantic realism about the target entities, and they take such grounds to be provided by our practices, they can get a reasonably clear idea of how we may adjudicate between metaphysical realism and metaphysical anti-realism about those entities.

[2] The term 'semantic realism' has been used variously in the literature. For instance, Michael Dummett calls 'realist' any view under which sentences in a relevant class have determinate truth values (e.g., Dummett 1982), while Herbert Feigl uses the term to refer to a view on the relationship between sentences containing observational and theoretical terms (e.g., Feigl 1950). The term as used here is intended to be distinct from these other uses. Thanks to a reviewer for highlighting this point.

It is perhaps less common to find views that hold metaphysical realism about a kind of entity without committing to either semantic realism or semantic anti-realism about those entities. Nevertheless, nearby views have been advanced that affirm the existence of a kind of entity while remaining neutral on the truth of *some* sentences about the nature of those entities. §3 also argues that if proponents of such views have adequate grounds on which to hold metaphysical realism about the target entities, then they should also have a reasonably clear idea of how we may determine our alethiological commitments regarding sentences about their nature.

§4 considers epistemological objections against metaphysical realism. It is sometimes argued that because certain disputed entities are epistemically inaccessible in some way, metaphysical realism about those entities would make it difficult to provide a plausible epistemology of the relevant domain. §4 argues that if metaphysical realists take our practices to provide the relevant arbiters of existence, then in view of the relations that may be expected to hold between arbiters, they should turn to those same practices when responding to epistemological objections.

1. Arbiters from practices

To see how the things we do can inform our metaphysical commitments, consider the following pair of hypothetical scenarios.

Scenario 1. In our best scientific theories, some electromagnetic phenomena are explained in terms of the electron. Suppose that one purpose for which we have scientific theories is to explain observed phenomena. Further suppose that we are for now somewhat uncertain of our understanding of electromagnetic phenomena, but our past successes in understanding and navigating our world using our scientific theories somewhat (even if not completely) justifies our belief in our current best scientific theories. Under these suppositions, should we say that electrons exist? It seems that insofar as we are inclined to say that other scientific posits exist, we should say the same of electrons. Since explanation is among our purposes for having scientific theories in the first place, the explanations in our scientific theories are key components of those theories. So, the justification our best theories have extends to our explanations of electromagnetic phenomena. If we take such justification to be reason to affirm the existence of some other scientific posits, then, it seems we should do the same for electrons. In this case, the epistemological grounds for our belief that electrons exist is part of our scientific practices, namely our use of electrons to explain electromagnetic phenomena.

Scenario 2. In our best scientific theories, some electromagnetic phenomena are explained in terms of the electron. Suppose that one purpose for which we have scientific theories is to explain observed phenomena by identifying the relevant dependency relations in the world. Further suppose that whether a scientific explanation succeeds in identifying dependency relations depends on the existence of its explanantia. That is, if it turns out that electrons do not actually exist, explanations of electromagnetic phenomena in terms of electrons would fail.[3] Under these suppositions, should we say that electrons exist? It seems that we should. Given that we intend scientific explanations to identify dependency relations in the world, our use of electron-based explanations assumes (perhaps tacitly) that those explanations can identify the relevant dependency relations. And since this assumption depends on the existence of electrons, we also tacitly assume in our use of electron-based explanations that electrons exist. It thus seems that we should affirm the existence of electrons to align our ontological beliefs with our tacit assumptions, at least for as long as we use electrons in scientific explanations. Here again, the epistemological grounds for our belief that electrons exist is provided by our scientific practices, namely by our use of electrons to explain electromagnetic phenomena.

These hypothetical scenarios illustrate two possible ways in which our practices can provide arbiters of existence. In both cases, our scientific practices can inform our ontological beliefs because they are somehow privileged with respect to our ontological commitments. In *Scenario 1*, our best scientific theories are privileged in the sense that their past successes justify our belief in them. In *Scenario 2*, our scientific explanations are privileged in the sense that their dependence on the existence of their explanantia implies that they carry tacit ontological assumptions. Either way, the fact that a scientific posit is involved in a particular way in our scientific practices can give us reason to affirm its existence.

Toward a generalisation, call a kind of practice *ontologically relevant* if an entity’s involvement in that practice can constitute good reason to

[3] To be sure, even if there were no electrons in the world, we would still be able to perform the act of explaining electromagnetic phenomena in terms of the (hypothetical) electron. The sense in which these explanations would fail is that (under the supposition above) they would be unable to identify the relevant dependency relations correctly, and hence unable to attain the purpose for which we have scientific explanations. Thanks to two anonymous reviewers for pressing for clarity on this point.

affirm the existence of that entity. The hypothetical scenarios above illustrate two (not necessarily exhaustive) ways in which some of our practices might be ontologically relevant. In cases where we affirm the existence of a kind of entity, a necessary condition for our practices to provide that entity's arbiter of existence is ontological relevance. For example, we do not typically think that an entity's appearance in fiction is a reason to affirm its existence, so although we affirm that human detectives exist, the epistemological grounds for this affirmation cannot be that a human detective appears in stories about Sherlock Holmes—the things we do with fictional stories cannot provide arbiters of existence because they are not ontologically relevant.

It might be wondered if any of our practices are ontologically relevant. Given that ontological claims are claims about what exists in the world, and our practices consist in human activities that may have little to do with worldly facts, it might seem odd to think that our practices can justify ontological claims. The scenarios above, however, suggest that it can sometimes be reasonable to think that our practices bear an epistemic connection to worldly facts. Namely, if a kind of practice has had a track record of success that indicates reliability regarding worldly facts, or if it depends on worldly facts in such a way that it would not be rational to engage in that kind of practice without believing those facts, it seems reasonable to consider that kind of practice a reliable guide for what some of our ontological beliefs should be. Indeed, it will be seen shortly that many do argue for the ontological relevance for some of our practices.

Another similarity between the two hypothetical scenarios is that in both cases, electrons were stipulated to play a key role in contributing to the purpose of our best scientific theories. It was supposed, in both scenarios, that some of our best scientific explanations depend on electrons, and that part of the purpose of our scientific practices is to provide explanations. These suppositions imply that if our scientific practices were revised such that our scientific theories did not contain apparent reference to electrons (in *Scenario 1*) or our scientific explanations were not given in terms of the electron (in *Scenario 2*), our ability to attain the very purposes for which we do science would be compromised.[4]

[4] To be sure, our *ability* to do science would not be affected even if electrons did not exist—indeed, it is possible that we were in fact wrong about electrons all along. What would be affected is our attainment of the *purposes* for which we do science. If it turns out that we were wrong about electrons, our scientific theories would be unable to serve their intended purposes, and hence should be revised. Thanks to two reviewers for highlighting this point.

In general, say that an entity is *indispensable* to a kind of practice if, were the relevant practices revised to avoid the use of that entity, or the use of theories containing apparent reference to that entity, the purpose for which we engage in those practices would be compromised. In the scenarios above, electrons are indispensable to our scientific practices.[5] In cases where we have reasons to affirm the existence of a kind of entity, a necessary condition for a kind of practice to provide that entity's arbiter of existence is for the entity in question to be indispensable to those practices. Someone having a hallucinatory experience might be justified in believing that there are tables in the world, but the epistemological grounds for their belief cannot be that they have table-like experiences, because tables are not indispensable for making sense of those experiences.

So when we are ontologically committed to a kind of entity, two necessary conditions for a kind of practice to provide that entity's arbiter of existence are ontological relevance and indispensability. It turns out that these conditions are also jointly sufficient. If a kind of practice is ontologically relevant, we have reason to affirm the existence of some entities involved in that kind of practice. And if a kind of entity is indispensable to that kind of practice, then we are justified in accepting an ontological commitment to those entities in particular.

In fact, some realist arguments in ontological debates proceed along these lines—they argue for an ontological commitment to a kind of entity on the grounds that those entities are indispensable to an ontologically relevant practice. Consider, as an example, the *Quine-Putnam indispensability argument* sometimes advanced in favour of mathematical Platonism, according to which we should affirm the existence of Platonic mathematical entities because those entities are indispensable to our best scientific theories (Quine 1981, 1986). The reasoning behind this argument is often understood in one of two ways. On one reading, the argument is that our scientific theories have some independent justification, which extends to mathematical entities on account of their indispensability to those theories (Baron 2013; Colyvan 2001). Alternatively, the argument may be understood as saying that our very use of scientific theories carries a metaphysical commitment to the mathematics on which those theories depend (Azzouni 2009; Panza and Sereni 2016; Resnik 1995). Either way, this argument attempts to make

[5] Some might think it more natural to describe Scenario 1 by saying that electrons are indispensable to our scientific *theories*. Given that we employ those theories as part of our scientific practices, it is also legitimate (albeit slightly less precise) to say of that scenario that electrons are indispensable to our scientific practices.

the case that our scientific practices are ontologically relevant, such that an entity's being involved in our scientific practices may be reason to affirm its existence (depending on which interpretation is adopted, the argument makes a case for ontological relevance similarly to either *Scenario 1* or *Scenario 2*, respectively). Then, according to the argument, mathematical entities are indispensable to our best scientific theories, from which it is concluded that we have reason to affirm the existence of mathematical entities, with our scientific practices providing their arbiter of existence.

Another example is David Lewis' (1986) argument for modal realism, according to which we should affirm the existence of concrete possible worlds because a realist view can provide a straightforward interpretation of our modal discourse. If this argument goes through, our discursive practices are ontologically relevant: the fact that we engage in modal discourse gives us reason to believe in the existence of some entities involved in that discourse. Lewis also argues that concrete possible worlds are indispensable to our modal discourse, in that interpretations of our modal discourse not involving concrete possible worlds are inferior in important respects to interpretations in terms of concrete possible worlds. From this it follows that we should affirm the existence of concrete possible worlds, with their arbiter of existence given by our modal discourse.

Apart from these examples, arguments have also been advanced for mathematical Platonism (Baker 2005; Colyvan 2010; Lyon 2011), moral realism (Enoch 2011; Majors 2003), scientific realism (Smart 1963), realism about grounding relations (Audi 2012), and theism (van Holten 2002) on the grounds that the respective entities are indispensable for some of our practices. The fact that connections between our practices and ontological commitments are often made via indispensability arguments lends further support to the idea that if our practices can provide arbiters of existence, the relation between those arbiters and our ontological commitments consists in indispensability.

We may also consider how this framework can be extended to cases in which we deny the existence of a kind of entity. To see how our practices can provide the epistemological grounds for such a denial, consider the following hypothetical scenario. Suppose we have reasons to reject an ontological commitment to phlogiston, and that phlogiston's arbiter of existence is given by our practices. Further suppose the following about three kinds of practices:

(i) Our discursive practices are not ontologically relevant.

(ii) Our practice of moral deliberation is ontologically relevant.
(iii) Our best scientific theories are ontologically relevant.

Which of (i)–(iii) can provide the epistemological grounds for denying the existence of phlogiston? It seems clear that (i) cannot. For, if our discursive practices are not ontologically relevant, they do not provide the epistemological grounds for any ontological commitments at all, so our rejecting an ontological commitment to phlogiston has nothing to do with our discursive practices. This reasoning generalises: in cases where we reject an ontological commitment to a kind of entity, a necessary condition for a kind of practice to provide that entity's arbiter of existence is ontological relevance.

From (ii) and (iii) it follows that phlogiston is not indispensable to either moral deliberation or our best scientific theories, given the discussion above. There is a sense in which both are part of the reason for which we deny phlogiston's existence. If phlogiston had been indispensable to either, we would have had reason to affirm its existence. But we can be more precise in identifying phlogiston's arbiter of existence. Although we would have been ontologically committed to phlogiston had it been indispensable for moral deliberation, it sounds odd to say that we should not be ontologically committed to phlogiston because it is dispensable for moral deliberation. For, given the conditions under which our concept of phlogiston was introduced, if we had been ontologically committed to phlogiston, this ontological commitment is more likely to have been underwritten by our best scientific theories than by our moral deliberation.[6] So, it is more natural to say that our best scientific theories provide the *primary* reason for which we are not ontologically committed to phlogiston—(iii) rather than (ii) is our epistemological grounds for denying phlogiston's existence. This reasoning also generalises: in cases where our practices give us reasons to reject an ontological commitment to a kind of entity, that entity's arbiter of existence is provided by the practices to which it would have been indispensable, had we had reasons to affirm its existence; and the reason for our actually denying that entity's existence is that it is in fact not indispensable to its arbiter of existence.

[6] Slightly more precisely, in terms of possible worlds: holding fixed the way our concept of phlogiston was introduced, some counterfactual world in which we are ontologically committed to phlogiston and phlogiston is indispensable to our best scientific theories is closer to actuality than any world in which we are ontologically committed to phlogiston and phlogiston is indispensable to moral deliberation.

Taking stock: if an entity's arbiter of existence is given by our practices, it is given by the ontologically relevant practices to which that entity would be indispensable, assuming we are ontologically committed to it. And, in these cases, the relation between arbiters of existence and our ontological commitments consists in indispensability: we should accept an ontological commitment to a kind of entity iff it is indispensable to the practices that provide its arbiter of existence. The discussion above suggests that these relations hold regardless of whether we have reasons to accept or reject an ontological commitment to the entity in question. Therefore, as may be expected, the identification of an entity's arbiter of existence can be epistemologically prior to the determination of our ontological commitments.

Three loose ends to tie up. First, the above account assumes that arbiters of existence are given by our practices. This might not always be the case. We might affirm the existence of some entities simply because of a favourable pre-theoretic intuition, or we might deny the existence of some entities because their existence would entail a contradiction. In such cases, our intuitions or logical constraints provide arbiters of existence, and the epistemological grounds for our ontological beliefs have little to do with ontological relevance or indispensability. The above account is not intended to apply to cases like these.

Second, whenever arbiters of existence are provided by our practices, entities can always be expected to have an arbiter of existence. Earlier, it was argued that a sufficient condition for accepting an ontological commitment to a kind of entity is that it be indispensable to an ontologically relevant aspect of our practices. This condition is also necessary: if a kind of entity is not indispensable to any ontologically relevant kind of practice, our practices would not give us reason to affirm the existence of those entities. For, in such cases, it would not make a significant difference to our practices whether the entities in question exist.[7] So if our practices do give us reasons to affirm the existence of a kind of entity (whether actually or counterfactually), that entity would be indispensable to some ontologically relevant practice, which would then be its (actual) arbiter of existence.

[7] In cases where our practices give no indication as to the existence of a kind of entity, views are divided as to whether we should *deny* the existence of those entities (Field 1989, 45; Leng 2010, 258-260), or remain *agnostic* on their existence (Bueno 2009, 79; van Fraassen 1989, 193), or take there to be *no fact of the matter* (Carnap 1950; Yablo 2009). The argument here requires only the weaker conclusion, compatible with all three options, that we have *no reason to affirm* the existence of those entities.

Third, it might sometimes be unclear how practices should be individuated. For instance, there might be several viable ways of delineating our scientific practices: our visual perception in ordinary contexts does not seem to fall squarely within our scientific practices, but it might be considered scientific under some broad construal of science. For the purpose of examining arbiters, what we require is a sufficiently fine-grained delineation of practices that respects epistemological differences. That is, practices should be treated as distinct insofar they provide different kinds of epistemological grounds for the existence of a kind of entity. To be sure, this is not a fully precise account of how to individuate practices, since there might be disagreement over whether some practices are sufficiently similar to be identified, or sufficiently different to be distinguished. Nevertheless, this constraint provides a rough principle for assessing delineations, at least for present purposes, and rules out delineations that are arbitrary or gerrymandered.

Relatedly, under some delineations of practices, an entity might be indispensable to several ontologically relevant practices. The more precise way of stating the earlier result is to say that an entity's arbiter of existence is provided by the *disjunction* of all the ontologically relevant practices to which it is indispensable. If all our best scientific theories are ontologically relevant, and electrons are indispensable to both our best theory of electric fields and our best theory of molecular energy states, the electron's indispensability to *either* theory would have provided sufficient epistemological grounds on which to be ontologically committed to electrons. If we had been ontologically committed to phlogiston, this would have been because it is indispensable either to our best theory of combustion or our best theory of rusting; the reason we are not so ontologically committed is because phlogiston is indispensable to *neither*. For simplicity, we will speak of arbiters as though they are provided by particular practices, though in fact they may be provided by disjunctions thereof.

Having examined arbiters of existence, we can explicate the notion of an arbiter of truth analogously. There might be cases in which the things we do give us reason to affirm the truth of some sentences regarding the nature of entities involved therein—call such practices *alethiologically relevant*.[8] If we are somehow justified in believing our best scientific theories, those theories would be alethiologically relevant. If we use scientific theories for the purpose of prediction, and the predictive

[8] That there are such sentences, or that such sentences are true, should not be taken to imply that the entities in question exist—see n.1.

accuracy of a theory depends on the truth of sentences therein, those theories would again be alethiologically relevant. A necessary condition for a kind of practice to provide an arbiter of truth is that it be alethiologically relevant.

A kind of sentence is indispensable to a kind of practice if any revision of those practices to eliminate dependence on those sentences, if even possible, would compromise our ability to attain the purposes for which we engage in those practices. As we currently perform moral deliberation, we might rely on the idea of some outcomes being better than others. If it is not possible to perform moral deliberation without relying on this idea, or if any way of performing moral deliberation without relying on this idea is inferior in important respects to the way we presently perform moral deliberation, then sentences about the relative superiority of outcomes are indispensable to moral deliberation. By arguments similar to those above, a necessary and sufficient condition for us to be alethiologically committed to a kind of sentence is that those sentences be indispensable to an alethiologically relevant practice.

An entity’s arbiter of truth is provided by the practices to which sentences about the nature of that entity would be indispensable if, whether actually or counterfactually, we have reason to affirm the truth of such sentences. We should accept an alethiological commitment to sentences about that entity iff that sentence is indispensable to that entity’s arbiter of truth.

2. Relations between arbiters

We now turn to the issue of relations between arbiters. If an entity’s arbiters of existence and truth are given by our practices, might we expect any relation between the two? It will be argued in this section that under either metaphysical or semantic realism about a kind of entity, the arbiters for that entity may be expected to align in some way.

To illustrate the difficulties that potentially arise for realist views if the two kinds of arbiters are misaligned, consider the mathematical Platonist view formulated (though not endorsed) by Penelope Maddy (1992), under which we should affirm the existence of mathematical entities on scientific grounds but be informed as to their nature on mathematical grounds:

> We could argue, first, on the purely ontological front, that the successful application of mathematics [to science] gives us good reasons to believe that there are mathematical things.

> Then, given that mathematical things exist, we ask: by what methods can we best determine precisely what mathematical things there are and what properties these things enjoy? To this, our experience to date resoundingly answers: by mathematical methods; the very methods mathematicians use. (Maddy 1992, 279)

This view takes our scientific practices to provide the arbiter of existence for mathematical entities and our mathematical practices to provide their arbiter of truth. The results in §1 imply that under this view, we should affirm the existence of all and only mathematical entities that are indispensable to our best scientific theories, and affirm the truth of all and only sentences about the nature of those entities that are indispensable to our best mathematical theories.

Now consider what would follow under this view if our scientific and mathematical practices are misaligned with respect to their dependencies on mathematical alethiology.[9] Suppose, for a simplified example, that only the real number structure, and no other mathematical entity, is indispensable to our best scientific theories. Depending on whether the continuum hypothesis is false, there might be a subset of the real numbers whose cardinality is strictly between the cardinalities of the natural numbers and of the real numbers. It is known that the continuum hypothesis is independent of the axioms of Zermelo-Fraenkel set theory with choice—the most widely accepted foundation for mathematics—so our present understanding of the real numbers underdetermines the existence of such a subset. Suppose mathematicians were to decide, on mathematical grounds, that we should take the continuum hypothesis to be false. It would then seem that we should be ontologically committed to a set of real numbers with cardinality between the naturals and the reals. For, we are ontologically committed to the real numbers on account of our scientific practice, and we should, on account of our best mathematical practices, attribute to the real numbers properties according to the falsity of the continuum hypothesis.

But on the view under consideration, our practices do not warrant an ontological commitment to such a set. Our scientific practices provide the arbiter of existence for mathematical entities, but a set with cardinality

[9] It might be argued that our mathematical and scientific practices are sufficiently similar that they may be taken to constitute just one kind of practice, under some broad delineation of practices. In that case, our mathematical and scientific practices will, trivially, be aligned in their dependencies. For the purposes of this section, we set aside this possibility and assume, as Maddy does, that our mathematical and scientific practices constitute different kinds of practices.

between the naturals and the reals is not indispensable to our scientific practices—the real numbers would be able to play their role in our best scientific theories even if the continuum hypothesis were true. The fact that mathematical practice assumes the falsity of the continuum hypothesis does not underwrite an ontological commitment to the sets in question, because our mathematical practices do not provide the arbiter of existence for numbers. In this way, the alethiological misalignment between the arbiters leads to a tension over whether to affirm the existence of some mathematical entities.

The scenario above was one in which sentences about a kind of entity are indispensable to the aspect of our practices providing the entity's arbiter of truth but not that providing its arbiter of existence. The reverse situation might also be possible. Suppose that mathematicians are indifferent as to whether we should take the continuum hypothesis to be true, but a set with cardinality between the naturals and the reals is indispensable to our best scientific theories. Here again, a tension arises between the arbiters, this time over whether to affirm the falsity of the continuum hypothesis. To say that the continuum hypothesis is false would be to affirm some sentences about numbers that are not indispensable to our mathematical practices. But not to say so would mean not accepting an ontological commitment to sets that are indispensable to our best scientific theories, which provides the arbiter of existence for mathematical entities.

In short, if we affirm the existence of numbers on scientific grounds and turn to mathematical practice when seeking to determine the precise properties of numbers, it seems that we are justified in following the dictates of mathematical practice only insofar as its claims about numbers have bearings on scientific practice. Attributing properties to numbers beyond that would seem to entail holding unwarranted beliefs about the existence of numbers. Similar considerations apply to views that affirm the existence of a kind of entity while taking our practices to provide its arbiters. Namely, if the kinds of practices providing each arbiter are misaligned with respect to their dependencies on the alethiology of those entities, tensions may arise over the precise set of properties to be attributed to the entities in question. One way to avoid these difficulties is simply to drop metaphysical realism.[10] Another way is to revise the target

[10] The argument above shows that difficulties arise even in cases where some sentences about the nature of the target entities are not indispensable to the aspect of our practices providing their arbiter of truth, so dropping semantic realism about the target entities would not avoid the difficulties completely. Both cases in the argument, however, assumed that the target entities are indispensable to their adopted arbiter of existence. Hence, the difficulties are limited to metaphysical realist views.

view such that the arbiter of existence for the entities in question is not given by our practices. Yet another way is to hold a view under which the kinds of practices providing the arbiters of existence and truth are aligned with respect to their alethiological dependencies. While this does not require that the *same* practices provide both arbiters, it requires that the same sentences regarding the target entities be indispensable for both kinds of practices. That is,

(i) if a view affirms the existence of an entity, and the adopted arbiter of existence is given by a kind of practice, then the view should affirm a sentence about the nature of those entities iff that sentence is indispensable to that kind of practice.

Returning to the Platonist view above, consider now what would follow if our scientific and mathematical practices differ in their ontological dependencies. For simplicity, assume for this and the next paragraph that we are ontologically committed to all and only mathematical entities that are the referents of mathematical sentences that we affirm.[11] One way for the two aspects of our practices to be misaligned is for there to be parts of mathematics that are indispensable to our best mathematical theories but not to our best scientific theories. Suppose, for instance, that some very large infinities in set theory have no application to our best science. Then, there would be tension over whether to affirm the existence of such numbers (and hence sentences about them). An ontological commitment to such numbers would be unwarranted by their arbiter of existence because they are not indispensable to our best scientific theories, but to reject an ontological commitment would be also to withhold affirmation from all sentences about those entities, violating the dictates of their arbiter of truth.

The opposite misalignment might also be possible: there might be mathematical entities that are indispensable to our best scientific theories but not our best mathematical theories. Suppose that our best scientific theories require the use of large cardinals that are not given by our current set theory. Similar tensions between the two adopted arbiters arise here. To reject an ontological commitment to large cardinals would be to violate the dictates of the adopted arbiter of existence, but to accept the ontological commitment would be also to affirm sentences about them, at least some of which are not indispensable to the adopted arbiter of truth.

[11] A similar argument to what follows would go through without this assumption, albeit with the appropriate restrictions to mathematical sentences that are true in virtue of mathematical entities.

So if we affirm the truth of number-sentences on mathematical grounds and turn to science to justify an ontological commitment to numbers, it seems that we are justified in following the dictates of science only insofar as its claims about which numbers exist agree with the sentences we know to be true from mathematical practice. This generalises to other views that affirm sentences about the nature of a kind of entity while taking its arbiters of existence and truth to be given by our practices. Namely, if the kinds of practices providing the two arbiters are misaligned with respect to their ontological dependencies, tensions may arise over the precise nature of the entities whose existence is to be affirmed. Analogous to the above, these difficulties can be avoided by dropping semantic realism or by locating the adopted arbiter of truth outside of our practices. Or, the view in question should take the target entity's arbiter of truth to depend on its ontology in the same way as does its arbiter of existence:

(ii) if a view affirms sentences about the nature of an entity, and the adopted arbiter of truth for that entity is given by a kind of practice, then the view should affirm the existence of that entity iff that entity is indispensable to that kind of practice.

3. Implications for realist views

(i) and (ii) bear most directly on views that, like the Platonist view considered in §2, hold both metaphysical realism and semantic realism about a kind of entity while taking different kinds of practices to provide arbiters of existence and truth for those entities. David Enoch (2007, 2011) argued for moral realism on the grounds that objective moral properties are indispensable for deliberation (Enoch 2011, 72-74), and suggests that our moral judgments are reliable guides to moral facts (ibid., 168). Enoch's view takes our deliberative practices and our moral judgments to provide the arbiters of existence and truth (respectively) for objective moral properties. Thus, if different moral properties or moral claims are indispensable to our deliberation and moral judgments, difficulties similar to the above might arise over whether to attribute those properties or affirm those claims. Another example is the Platonist view defended by the early Maddy, who argued that we should be ontologically committed to mathematical entities because they are indispensable to our best scientific theories, and that we can know about mathematical entities by sense perception (Maddy 1990). For this view to avoid difficulties analogous to those above, it will have to be argued that our scientific theories and sense perception have the same ontological and alethiological dependencies.

(i) and (ii) also have implications for views that either hold metaphysical realism about a kind of entity while remaining neutral on their alethiology, or hold semantic realism about a kind of entity while remaining neutral on their ontology. The latter is perhaps more common. Hilary Putnam, for instance, argued that we should affirm mathematical sentences because ‘a reasonable interpretation of the application of mathematics to the physical world requires a realistic interpretation of mathematics’ (Putnam 1979, 74). He also held, however, that the applicability of mathematics to science does not commit us to Platonism, because it is possible for mathematical sentences to be true in virtue of possible structures in modal space rather than Platonic mathematical entities (ibid., 72).[12] Together with this non-commitment to Platonism, it might be thought that there is also no good reason why mathematical sentences *cannot* be true in virtue of Platonic mathematical entities. (Putnam did not think this—he held that mathematical sentences are in fact true in virtue of possible structures.) So there might be a view that takes our best scientific theories to provide the arbiter of truth for mathematical entities, holds semantic realism about mathematical entities, but remains neutral between metaphysical realism and metaphysical anti-realism about *Platonic* mathematical entities.

The argument of §2 implies that neutrality is an unstable position for such a view. In particular, (ii) implies that we should affirm the existence of Platonic mathematical entities under this view iff they are indispensable to our best scientific theories. Suppose first that Platonic mathematical entities are not so indispensable, say because our best scientific theories are indifferent between mathematical sentences being true in virtue of either Platonic mathematical entities or possible structures. It follows from this that sentences about mathematical entities that distinguish Platonic entities from possible structures (such as ‘2 is a Platonic entity’) are not indispensable to our best scientific theories. This is a reason *against* an ontological commitment to *either* Platonic entities *or* possible structures. For, it implies that to affirm the existence of Platonic entities rather than possible structures (or *vice versa*) would be to affirm sentences about mathematical entities in violation of their adopted arbiter of truth. Now suppose that Platonic mathematical entities are indispensable to our best scientific theories. That is, our best scientific theories require that mathematical entities bear properties of Platonic

[12] Also see Putnam (1967) and (2006).

entities rather than possible structures. In this case, since our best scientific theories provide the arbiter of truth for mathematical entities, we should identify mathematical entities with Platonic entities and accept an ontological commitment to the latter.

The upshot is that under the view in question, we should affirm the existence of Platonic entities iff they are indispensable to our scientific practices, which gives us a reasonably clear idea of what it would take for us to be ontologically committed to Platonic entities. This does not mean that we cannot be agnostic regarding this ontological commitment, to be sure, since we might not as yet have determined whether the indispensability claim is true. But it does mean that neutrality on the existence of Platonic entities cannot be the end of our inquiry.

Analogously, there might be views that adopt an arbiter of existence for a kind of entity and affirm the existence of those entities, but remain neutral as to whether we should affirm any sentences about the nature of those entities. Against such views, (i) says that we should affirm all and only sentences about the nature of those entities that are indispensable to the practices providing their adopted arbiter of existence. This yields a reasonably clear idea of what it would take to adjudicate between semantic realism and semantic anti-realism, implying that neutrality between the two is also unstable.

Views of the kind just described are less common in the literature, since it might seem rather odd to affirm the existence of a kind of entity without saying anything about what those entities are like. Nevertheless, nearby views have been advanced that affirm the existence of a kind of entity while committing to *little* by way of their alethiology. Consider, for instance, the form of Platonism defended by Mark Colyvan (2001), who argues for an ontological commitment to mathematical entities on the grounds that they are indispensable to our best scientific theories. Colyvan holds that the argument does not commit us to any particular view about the nature of mathematical entities:

> [The argument] simply asserts that there *are* mathematical objects. They might be constituted by more mundane items such as universals and/or relations…patterns or structures…or the part/whole relation. Perhaps they are constituted by *more* exotic items such as possible structures (…). In short, any (realist) account of mathematical objects is all right by the indispensability argument. (Colyvan 2001, 143; emphasis original)

This view takes our best scientific theories to provide the arbiter of existence for mathematical entities and affirms that we are ontologically committed to such entities. It also affirms mathematical sentences insofar as they are indispensable to our best scientific theories. The view is neutral, however, regarding whether we are alethiologically committed to sentences that have implications for the precise nature of mathematical entities. For instance, this Platonist view neither affirms nor denies ‘2 has two members’, which would be true if the natural numbers are the von Neumann ordinals and false if they are universals.

The argument of §2 also implies that neutrality is an unstable position for such a view. If indeed sets and universals can play the role of mathematical entities in our best scientific theories equally well, then neither is indispensable to our best scientific theories. This would be a reason against alethiological commitments to sentences about mathematical entities that imply their being (say) universals. For, to affirm such sentences while holding an ontological commitment to mathematical entities would imply an ontological commitment to universals, which would be unwarranted by the adopted arbiter of existence. Conversely, if universals are indispensable to the role of mathematical entities in our scientific practices, then we should be ontologically committed to mathematical entities as universals, and affirm sentences attributing (or implying) all the relevant properties. Either way, our scientific practices under this view can provide sufficient grounds to determine what we should hold about the precise nature of mathematical entities.

4. Implications for epistemological objections

(i) also bears on a line of objection sometimes raised against metaphysical realism. It is sometimes argued that we should not affirm the existence of a disputed entity because affirming its existence would make it difficult to account for our knowledge in the corresponding domain. For, if we affirm the existence of a kind of entity, it seems natural to hold also that some of our knowledge in the corresponding domain is knowledge regarding those entities. But in some cases, the entities in question might be abstract, causally isolated, unobservable, or epistemically inaccessible in some way. Consequently, it might be unclear how our beliefs about such entities can be reliable, and hence how knowledge about them is possible. Insofar as a plausible epistemology of the domain in question is not forthcoming under metaphysical realism, this casts doubt upon the view. Objections along these lines have been raised against metaphysical realism about mathematics (Benacerraf 1973; Field 1989), objective moral properties (Mackie 1977), concrete possible worlds (Peacocke 1997, 1999), and objective logical facts (Schechter 2010), among other things.

(i) implies that if metaphysical realists take our practices to provide the grounds on which to affirm the existence of a kind of entity, they should turn to those same grounds when responding to such objections. For, (i) implies that we should, under their view, affirm all and only sentences about the nature of the entities in question that are indispensable to the practices that provide its arbiter of existence. Thus, those practices may provide epistemic access to the target entities—insofar as our beliefs in the corresponding domain align with the alethiological dependencies of those practices, those beliefs reliably track what we should affirm about the nature of the entities in question.

As an illustration, consider a Platonist view under which our best scientific theories provide the arbiter of existence for mathematical entities. Under such a view, we can take our best scientific theories as a guide to what we should believe about mathematics. For, when our best scientific theories depend on the existence of mathematical entities, those theories also require that mathematical entities play a particular role, and whether mathematical entities can fulfil this role will depend on whether they bear certain properties. So our best scientific theories dictate not only that we affirm the existence of mathematical entities, but also that we take mathematical entities to exist with a particular set of attributes. Platonists who hold this view may thus say that we can attain mathematical knowledge reliably by considering what mathematical entities have to be like to play their role in science. Indeed, some Platonists do account for our mathematical knowledge in this way (e.g., Colyvan 2001, 151-155).

To be sure, the practices taken to provide the arbiter of existence may, in fact, not be a reliable guide to knowledge in the target domain. However, the implication in this case is not that metaphysical realists should turn to other kinds of practices to construct an epistemology for that domain, it is that the realist view itself has been undermined. According to (i), if the sentences we should affirm about an entity is misaligned with the alethiological dependencies of a kind of practice, then that kind of practice cannot provide the arbiter of existence for those entities. For instance, if our best scientific theories are unreliable guides to mathematical truth, then those theories do not depend on how mathematical entities are like. But then the argument for (i) in §2 implies that the success of our scientific theories is also independent of the existence of mathematical entities, and hence that those theories are inadequate grounds on which to hold Platonism. So it would be mistaken for Platonists to continue holding their view on grounds of our scientific practices while acknowledging that those practices cannot provide a plausible epistemology. And in general, if the supposed arbiter of existence for a metaphysical realist view cannot provide an adequate response to epistemological objections, it in fact cannot support the view at all.

5. Conclusion

It is commonly thought that arbiters of existence and truth are given by certain privileged kinds of practices. This paper has attempted to flesh out this view of arbiters and draw its implications. The sense in which our practices may be privileged and in a position to inform our metaphysical commitments consists in ontological or alethiological relevance. And, the relation between arbiters and our metaphysical commitments consists in indispensability: we delineate our metaphysical commitments according to the entities (or sentences) that are indispensable to ontologically (or alethiologically) relevant aspects of our practices. Taking arbiters to be given by our practices has implications for how arbiters of existence and truth should relate: if a view holds a realist commitment to an entity, it should also take the kinds of practices providing that entity's arbiters to align with respect to their metaphysical dependencies. This has two further implications. First, views holding an ontological or alethiological commitment to an entity have sufficient grounds in principle to arbitrate on the other realist commitment, and thus should not seek to maintain neutrality on the latter. And, metaphysical realists about an entity who take a kind of practice to provide its arbiter of existence should turn to those same practices when responding to epistemological objections.

Acknowledgments

Thanks to Ben Blumson and three anonymous reviewers for helpful comments on earlier drafts.

Funding information. This work was supported by the National Research Foundation, Singapore, under its Medium Sized Centre Programme—Centre for Advanced Robotics Technology Innovation (CARTIN), Subaward A-0009428-08-00.

REFERENCES

Armstrong, David M. 1968. *A Materialist Theory of the Mind.* London: Routledge & Kegan Paul.

Audi, Paul. 2012. “A clarification and defense of the notion of grounding.” In *Metaphysical Grounding: Understanding the Structure of Reality,* edited by F. Correia and B. Schnieder, 101–121. Cambridge: Cambridge University Press.

Azzouni, Jody. 2009. “Evading truth commitments: The problem reanalyzed.” *Logique et Analyse* 52 (206): 139–176.

Baker, Alan. 2005. “Are there genuine mathematical explanations of physical phenomena?” *Mind* 114 (454): 223–238.

Baron, Sam. 2013. “A truthmaker indispensability argument.” *Synthese* 190: 2413–2421.

Benacerraf, Paul. 1973. “Mathematical truth.” *The Journal of Philosophy* 70 (19): 661–679.

Bueno, Otávio. 2009. “Mathematical fictionalism.” In *New Waves in Philosophy of Mathematics*, edited by O. Bueno and Ø. Linnebo, 59–79. UK: Palgrave Macmillan.

Carnap, Rudolf. 1950. “Empiricism, semantics, and ontology.” *In Philosophy of Mathematics: Selected Readings*, edited by P. Benacerraf and H. Putnam, 241–257. Cambridge: Cambridge University Press.

Colyvan, Mark. 2001. *The Indispensability of Mathematics.* New York: Oxford University Press.

Colyvan, Mark. 2010. “There is no easy road to nominalism.” *Mind* 119 (474): 285–306.

Devitt, Michael. 1991. “Aberrations of the realism debate.” *Philosophical Studies* 61: 43–63.

Dummett, Michael. 1979. *Truth and Other Enigmas*. London: Duckworth.

Dummett, Michael. 1982. “Realism.” *Synthese* 52 (1): 145–165.

Enoch, David. 2007. “An outline of an argument for robust metanormative realism.” In *Oxford Studies in Metaethics*, vol. 3, edited by R. Shafer-Landau, 21–50. Oxford: Oxford University Press.

Enoch, David. 2011. *Taking Morality Seriously: A Defense of Robust Realism*. Oxford: Oxford University Press.

Feigl, Herbert. 1950. “Existential hypotheses. Realistic versus phenomenalistic interpretations.” *Philosophy of Science* 17 (1): 35–62.

Field, Hartry. 1989. *Realism, Mathematics and Modality.* Oxford: Blackwell.

Leng, Mary. 2010. *Mathematics and Reality*. Oxford: Oxford University Press.

Leplin, Jarrett. 1984. *Scientific Realism.* Berkeley: University of California Press.

Lewis, David. 1986. *On the Plurality of Worlds.* Oxford: Basil Blackwell.

Lyon, Aidan. 2011. "Mathematical explanations of empirical facts and mathematical realism." *Australasian Journal of Philosophy* 90 (3): 559–578.

Mackie, John L. 1977. *Ethics: Inventing Right and Wrong*. London: Penguin Books.

Maddy, Penelope. 1990. *Realism in Mathematics.* New York: Oxford University Press.

Maddy, Penelope. 1992. "Indispensability and practice." *The Journal of Philosophy* 89 (6): 275–289.

Majors, Brad. 2003. "Moral explanation and the special sciences." *Philosophical Studies* 113: 121–152.

Panza, Marco, and Andrea Sereni. 2016. "The varieties of indispensability arguments." *Synthese* 193: 469–516.

Peacocke, Christopher. 1997. "Metaphysical necessity: Understanding, truth and epistemology." *Mind.* 106: 521–574.

Peacocke, Christopher. 1999. *Being Known.* Oxford: Oxford University Press.

Putnam, Hilary. 1967. "Mathematics without foundations." *The Journal of Philosophy* 64 (1): 5–22.

Putnam, Hilary. 1979. "What is mathematical truth?" In *Mathematics Matter and Method*, 60–78. Cambridge: Cambridge University Press.

Putnam, Hilary. 2006. "Indispensability arguments in the philosophy of mathematics." In *Philosophy in an Age of Science: Physics, Mathematics, and Skepticism*, 181–201. Massachusetts: Harvard University Press.

Quine, Willard Van Orman. 1951. "Two dogmas of empiricism." *The Philosophical Review* 60: 20–43.

Quine, Willard Van Orman. 1963. "Carnap and logical truth." In *Philosophy of Mathematics: Selected Readings*, 355–376. Cambridge: Cambridge University Press.

Quine, Willard Van Orman. 1981. "Success and limits of mathematization." In *Theories and Things*, 148–155. Cambridge: Harvard University Press.

Quine, Willard Van Orman. 1986. "Reply to Charles Parsons." *In The Philosophy of W. V. Quine*, 396–403. La Salle: Open Court.

Resnik, Michael. 1995. "Scientific vs mathematical realism." *Philosophia Mathematica* 3 (3): 166–174.

Ridge, Michael. 2019. “Relaxing realism or deferring debate?” *Journal of Philosophy* 116 (3): 149–173.

Sayre-McCord, George. 1986. “The many moral realisms.” *The Southern Journal of Philosophy* 24: 1–22.

Schechter, Joshua. 2010. “The reliability challenge and the epistemology of logic.” *Noûs* 24: 437–464.

Smart, J. J. C. 1963. *Physical Objects and Physical Theories*. London: Routledge.

Thomasson, Amie L. 2014. *Ontology Made Easy*. Oxford: Oxford University Press.

van Fraassen, Bas C. 1989. *Laws and Symmetry*. Oxford: Oxford University Press.

van Holten, Wilko. 2002. “Theism and inference to the best explanation.” *Ars Disputandi* 2: 1–20.

Yablo, Stephen. 2009. “Must existence-questions have answers?” In *Metametaphysics: New essays on the foundations of ontology*, edited by D. Chalmers, D. Manley, & R. Wasserman, 507-526. Oxford: Oxford University Press.

BOOK REVIEW

Umberto Galimberti
L'ETICA DEL VIANDANTE
Feltrinelli, 2023

ISBN: 9788807493645 (paper)
ISBN: 9788858858530 (e-book)
Paperback, 20,90 EUR
e-book: 12,99 EUR

Martina Blečić[1]

[1] University of Rijeka, Faculty of Humanities and Social Sciences, Croatia

The Nietzschean quote from "Human, All Too Human", which Galimberti uses to inaugurate his latest book, establishes its prevailing tone: the apex of human reason's freedom takes the form of a wanderer, progressing toward an undefined path as a concreate goal cannot be individuated.

Just as Nietzsche is a controversial thinker and, at the same time, a philosopher who continues to captivate new generations, Galimberti himself has faced public accusations of appropriation of others' ideas, however, his book "L'etica del viandante" is hailed as having "all the prerequisites to become a classic of contemporary philosophical thought". While Galimberti may not be Nietzsche, let's explore what he brings to the table.

In the book, he provides an overview of the historical development of Western thought, starting from its two sources: Greek culture and the Judeo-Christian tradition. Despite their differences, Galimberti identifies a common thread between the two: the pursuit of order and stability. Greek thought is tied to nature, an immovable backdrop witnessing human endeavours. This is linked to a cyclical understanding of history, which lacks an ultimate goal but sees death as the conclusion of individual efforts.

Correspondence: martinablecic@gmail.com

The awareness of death leads to the ethics of limits, warning humans not to exceed their boundaries. The fact that man is mortal and just a part of nature combined with the quest for truth and rational knowledge leads to dualism, where man consists of both body and soul.

At a certain point, Greek philosophy encountered the ideas of Judeo-Christianity. The cyclical view of time was replaced by eschatological time, tracing its path from Earth to heaven in anticipation of salvation. This vision replaces nature with God and shifts the focus from the past to the future.

Such a perspective dominated until the modern era, with the realization that Earth revolves around the Sun, which has no inherent purpose which lead to the acknowledgment of the relativity of all motion. The world loses its enchantment, and the dominant narrative sees humanity's goal as mastering nature. The mastery of nature was supposed to contribute to human emancipation, freeing individuals from religious beliefs and superstitions. The use of reason allows the replacement of divine laws, which previously governed lives, with the laws of mathematics. If mathematics is the language of nature, then nature can be understood. Man, now at the centre of history, can overcome all negatives like ignorance, poverty, and disease, achieving complete liberation. However, even this faith in science begins to waver—Freud explores the role of the subconscious, and Mach, Hilbert, and Planck question the previously laid foundations of science.

Nevertheless, the final blow to faith in the science nurtured by modernity was dealt by Nazi ideology. The collapse of faith in universal reason led to the end of modernity and opened the doors to postmodernity, bringing cultural relativism and the complete dominance of individualism.

Galimberti's insights into our time, that he calls the technological age are perhaps more intriguing than his historical reconstruction. He believes that technology has now taken on the role of a subject, not just a means of human action. It is not just a subject but also the ultimate goal.

He criticizes the idea that technology can liberate us, help us overcome obstacles as it once seemed. "Are we truly free today not to use a computer or a mobile phone?", Galimberti asks. The author argues that we are not; we cannot choose another means to communicate with, for example, the government or a bank. Technology is not just the application of scientific results; it is the essence of science. Of course, this has significant moral implications. Technological experimentation is not conducted in the safe conditions of a laboratory but throughout the world. If we add to that the idea that today the human ability to do something is much greater than its ability to foresee the consequences of what is done, the future can be worrisome.

In the pre-technological era, man dominated nature through the use of technological tools; today, technique dominates man with its rationality, which does not recognize anything beyond itself. In the technological age, humanism is lost, not because technology is not yet developed enough, but because it does not concern itself with it at all. Technology does not strive for a goal, does not promote meaning, and does not open possibilities for salvation; it simply acts, prompting questions about concepts like individual, identity, freedom, truth, meaning, morality, politics, democracy, and others. For example, ethics in the technological age becomes powerless due to the technological imperative to know everything that can be known and to do everything that can be done.

Again, the goal of technology is work, production, which no longer stems from human rationality but from the rationality of the machine. Traditional ethics is no longer applicable because it cannot transcend its anthropocentrism and regulate knowledge and power beyond the space of the planet and the time of human life. Ethics once could guide us on how to act, how to use technology, but today, it has no influence because action is inhuman. Consequences are no longer the product of human decision and conscious action but the result of a process. The idea of human responsibility for one’s actions is behind us. In the technological age, responsibility concerns only the proper performance of the machine’s action. Technology feeds itself and leads to consequences independent of any direct intention.

Therefore, Galimberti calls on us to return to the ancient virtue of measure. Giving oneself measure becomes urgent. To achieve this, we need the ethics of the wanderer, who, without using a map, faces difficulties one by one as they come. This is our limit. Ethics cannot be prescriptive; it must try to catch up with technology. We can no longer speak of a goal, but anyone focused only on the goal does not enjoy the journey. They travel to arrive, not to travel. To achieve this, we must abandon deeply rooted beliefs; we must not appeal to rights but to experience and the observation of its diversity.

In the elaboration of his planetary ethics, the author brings concrete examples of the dangers that technology brings, mentioning global warming, the consequences of genetically modified organisms, and nuclear energy. He also cites ozone holes, water pollution, and glacier melting. They testify to the development of scenarios that do not unfold due to the power humans have over nature but due to the power that technology has over humans and nature.

To prevent a scenario in which progress approaches catastrophe, it is not enough to reduce the use of technology. We need to radically change the paradigm that guided the relationship between man and nature, moving from anthropocentrism to biocentrism. The wanderer knows this; he knows that life belongs to nature that preceded man and that will exist after him. Such ethics can be called planetary because the life of the Earth becomes the measure of all things.

However, it does not forget man. To ensure that all living beings live under suitable conditions, brotherhood, a sense of unity among people, needs to be fostered. We need to abandon the idea that one culture is superior to others, and for peace to reign, we need to give up on states because the peace they want to achieve within themselves leads to war with others. To achieve this, education is needed, teaching us from an early age that we are all equal. Just as individuals renounce some of their freedom to be part of society, nations must now renounce some of their interests to join an ecological culture.

To achieve this, cultural evolution is needed to tear down divisions between races, religions, nations, and states. Furthermore, we need to replace the logic of the enemy with the logic of brotherhood. This brotherhood includes not only humans but all living and non-living things on Earth. The wanderer does not see Earth as a source of resources but as a value to be preserved, and we all must strive to be like him.

Certainly, only the rough outlines of Galimberti's ideas are presented here, and I certainly hope that the most important ideas have been covered. What can we say about them? Although the message of universal ethics that applies to all beings on Earth, including the Earth itself, is attractive and I believe in the correctness of such views, its feasibility remains uncertain. According to the author we should become wanderers, nomads, aware of our transience and equality with others that do not look for goals or prescribed norms. However, if we reject all known ethics and their prescriptiveness, how will we act in specific situations? What will guide our actions if not an awareness of the universality of our rationality and moral sense? And if we talk about universality, are we not talking of a goal? Isn't the fraternal society that Galimberti calls for a goal in itself?

The author himself points out the difficulties and contradictions of today's capitalist system, which goes hand in hand with the technological age but presents it as subordinate to technology. Perhaps it would be good to reverse the story here. Critiquing and pointing out the enormous shortcomings of capitalism as the dominant system of values is necessary, but if we support a cyclical view of history, as the author does, things could fall into place on their own. Perhaps we should not do anything? Contrary to that, Galimberti argues that we should let go of the ideas of states and nations. Calling for the abolition of states and nations, as noble as it may sound, is unrealistic because it is contrary to the human need for association. If it not feasible, as idealistic as it sounds it is not a solution.

Furthermore, is his personified portrayal of technology, as something greater than man, as something that governs human lives, one that leads to a feeling of helplessness? I would argue that it is. How, as individuals, as wanderers, do we confront such a Goliath? Aren't we the ones who created technology and the ones who still control it? Keeping that in mind we should be optimistic in our hope to use it for good, since technology itself is amoral.

When we look at the examples of technology that the author offers, we may be surprised at how little space is provided for presenting their threat. In the last decades discussions about, for example, genetically modified organisms or nuclear energy have been numerous, and their advocates are not only technology enthusiasts but also those who believe that these technologies can bring benefits to humans and nature.

In conclusion, we can say that Galimberti's heart is certainly in the right place, and any reflection on the contemporary world and its future is more than welcome. Still, perhaps proposals for its improvement should be based on a broader and more nuanced approach to its understanding.

European Journal of Analytic Philosophy EuJAP | 2024 | Vol. 20 | No. 1 | 31-54
https://doi.org/10.31820/ejap.20.1.2

IS L.A. PAUL'S ESSENTIALISM REALLY DEEPER THAN LEWIS'S?

Cristina Nencha[1]

[1] University of Bologna and University of Bergamo, Italy

Original scientific paper – Received: 06/10/2023 Accepted: 06/03/2024

ABSTRACT

L.A. Paul calls "deep" the kind of essentialism according to which the essential properties of objects are determined independently of the context. Deep essentialism opposes "shallow essentialism", of which David Lewis is said to be a prominent advocate. Paul argues that standard forms of deep essentialism face a range of issues (mainly based on an interpretation of Quinean skepticism) that shallow essentialism does not. However, Paul claims, shallow essentialism eliminates the very heart of what motivates essentialism, so it is better to be deep than shallow. Accordingly, she proposes a very sharp novel account of essentialism, which, while attempting to preserve some of the advantages of shallow essentialism over the classical forms of deep essentialism, can be deemed to be deep.

In this paper, I compare Paul's proposal for a kind of deep essentialism with Lewis's account, as it is presented by Paul. My aim is to show that the differences between the two approaches are not as significant as Paul takes them to be, and that Paul's account can be taken to be deeper than Lewis's only at the cost of sacrificing the very idea at the bottom of deep essentialism.

This might be taken to suggest that, if Paul is correct in asserting that shallow essentialism is better equipped to address some skeptical challenges, but it is generally preferable to be deep than shallow, then Lewis's account should be re-evaluated, since, as shallow as it can be, it might be deeper than it looks.

Keywords: David Lewis; essentialism; L.A. Paul; context-sensitivity.

Correspondence: cristina.nencha@gmail.com

1. Introduction

Let us take essentialism to be the doctrine that at least some non-trivial property is determined to be essential to some objects, where trivial properties are properties such as being either *F* or non-*F*, for any property *F*.[1] L.A. Paul distinguishes between "deep" and "shallow" essentialism. Deep essentialism is the thesis according to which properties are determined to be essential to an object *O* independently of the context.[2] Shallow essentialism opposes deep essentialism: it rejects the view that properties can be determined as essential to *O* independently of the context. David Lewis is said to be a shallow essentialist.

Paul argues that standard forms of deep essentialism face a range of issues (mainly based on an interpretation of Quinean skepticism) that shallow essentialists à la Lewis do not. However, she claims, deep essentialism is what gives us those properties that define the real, ultimate, nature of the objects, while shallow essentialism eliminates the very heart of what motivates essentialism. Therefore, as Paul states, "it is better to be deep than to be shallow" (Paul 2006, 347). Accordingly, she proposes a very sharp novel account of essentialism, which, while attempting to preserve some of the advantages of shallow essentialism over the classical forms of deep essentialism, can be deemed to be deep.

While I concur with Paul regarding the challenges traditional forms of deep essentialism encounter in addressing skepticism, I will not delve into this matter. My focus in this paper will be on comparing Paul's proposal for a kind of deep essentialism with Lewis's account as presented by Paul. I will demonstrate that the differences between the two approaches in terms of depth are not as significant as Paul takes them to be, and that Paul's account can be taken to be deeper than Lewis's only at the cost of sacrificing the very idea at the bottom of deep essentialism.

This might be taken to suggest that, if Paul is correct in asserting that shallow essentialism is better equipped to address some skeptical challenges (as I think she is), but it is preferable to be deep than shallow,

[1] In the example, the triviality of the property of being *F* or non-*F* relies on the fact that this property belongs to all things (see Della Rocca 1996). For issues regarding the condition of triviality, see, for instance, Wildman (2016) and De (2020). In the following, I will take for granted the reference to non-trivial properties.

[2] The condition about the context-independency is generally attributed to W. V. O. Quine (1953), and the contextual factors that are usually regarded as relevant are those relative to the ways *O* is represented (namely, thought or described) in a context.

then Lewis's account should be re-evaluated, since, as shallow as it can be, it might be deeper than it looks.

Before moving on, some quick terminological clarifications. Given an object *O*, I will distinguish between "the properties that are said/determined to be essential to *O*" and "the essential properties of *O*". Unlike the latter, the first are those properties that are characterized as "essential" to *O,* only according to some context. For the sake of clarity, I will sometimes use expressions like "context-independent essential properties" or "essential properties determined independently of the context", even though they are only redundant ways to say "essential properties". Claims that attribute essential properties to objects, or properties that are said/determined to be essential, are "essentialist claims".

2. Deep essentialism

Paul (2004, 2006) aims to defend a kind of deep essentialism (from now on "DE"). She writes that, according to DE: "essential properties are *absolute*, i.e., are not determined by contexts of describing (or thinking, etc.) about the object, and truths about such properties are absolute truths" (Paul 2006, 333). Similarly, she says that "deep essentialism takes objects' natures and claims about them to be independent of context (…)" (Paul 2006, 358).

From these definitions, it seems reasonable to infer that DE, for Paul, is a thesis that holds at two levels of analysis: semantics and metaphysics. Indeed, Paul claims that the properties that define objects' natures as well as "the truths about such properties" or the "claims about [objects' natures]" are supposed to be "absolute", namely, context-independent. However, in a footnote to the first quote, Paul claims that DE "should defend a certain sort of semantic indeterminacy consistent with this view (…)" (Paul 2006, 366). And, in some other places (for instance, see Paul 2004, 180), she clearly admits that, in her account, the truth-values of essentialist claims are inconstant.

I believe that, in order to understand what is really at stake here, we should clearly distinguish between the two different understandings of DE. Broadly speaking, semantics is about the semantic values of expressions. Semantically speaking, DE might thus be intended as the thesis that the truth-values of essentialist sentences must be context-independent (I will refer to the semantic understanding of DE as "DE semantical"). Metaphysics can be thought to concern the nature of the facts in the world, which are the truth-makers for sentences (the potentially truth-making properties, if we are going in for the truth-maker talk). Therefore, metaphysically speaking, DE might be interpreted as the thesis that what makes essentialist sentences true must be facts in the worlds, which are independent of the context (I will refer to the metaphysical reading of DE as "DE metaphysical"). In the following, I will compare Lewis's stance towards these theses with Paul's.

3. Lewis's account

Let us consider essentialist sentence type 1):[3]

1) Socrates is essentially human.

According to Lewis, the truth-conditions for 1) are given in terms of counterpart relations of Socrates. Counterpart relations are similarity relations that Socrates entertains with objects of (usually) other worlds, namely, counterparts (see, for instance, Lewis 1968, 1986).

Now, similarity is defined in terms of properties sharing. The fact that two individuals have some properties in common, that they are similar in some way, does not depend, in general, on the context;[4] it depends on how the world(s) is(are) made. In other words, it is the business of metaphysics to establish similarity relations between objects. Similarity, as it has been defined, is very easy to get: almost anything is similar to anything else, under some respect (see, for instance, Goodman 1972). Therefore, given an individual like Socrates, there are a lot of similarity relations that he entertains with possible objects, and such similarities are established independently of the context.

Then, there is the further question of which similarity relations are relevant, and relevance is a matter of context. Accordingly, as Divers puts it, what may change from one context to another "are facts about which [similarity] relations are relevant in a context, not the facts about the obtaining or otherwise of [similarity] relations" (Divers 2007, 18).

[3] I use the distinction type/token in order to point out the fact that, according to Lewis, the logical form of an essentialist sentence is incomplete. This completion happens only at the level of specific tokens of that sentence.

[4] To be sure, in some special (maybe uninteresting) cases, the fact that two individuals share a property *does* depend on some contextual facts. My arguments in the following will ignore these special cases.

Now, retracing Paul's (2006, 344ff.) distinction between an "evaluative" and an "antiessentialist" kind of shallow essentialism, there are two plausible readings of how Lewis intends that we think of similarity relations and their selection, when it comes to essentialist claims:[5]

A. Only when a similarity relation is deemed to be contextually relevant, then that similarity relation enjoys the status of a counterpart relation and enters in the determination of the truth-conditions for 1). Accordingly, an individual *O* is a counterpart of Socrates if and only if (hereafter, "iff") *O* is similar to Socrates in a contextually relevant way. Thus, while there are similarity relations *simpliciter* (namely, established independently of the context), there are no counterparts of Socrates *simpliciter*, but only counterparts of him relative to some context.

B. The similarity relations *simpliciter* are regarded as counterpart relations. Therefore, there are counterpart relations *simpliciter*, namely, that obtain independently of the context. However, only the counterparts that are determined as contextually relevant are employed in determining the truth-conditions of essentialist claims. (Interpretation A corresponds to evaluative shallow essentialism, and B to the antiessentialist kind).

Paul is more inclined to interpret Lewis according to A, though she is not firm about this point (Paul 2006, 370).[6] In this paper, I will discuss both interpretations, and I will call the Lewisian who accepts A "Lewis$_A$", and the Lewisian who endorses B "Lewis$_B$". If I do not specify the interpretation, it means that I am referring to both and, in order to do that, I will use the devise of putting between brackets what holds only for Lewis$_B$.

Lewis$_A$ and Lewis$_B$ give different truth-conditions for essentialist claims, such as 1). According to Lewis$_A$:

> 1$_A$. Socrates is essentially human iff every counterpart of Socrates is human.

[5] Interesting discussions about this topic can be found in Hazen (1979) and Heller (2005).

[6] As Paul rightly says, there are relevant differences in how Lewis accounts for essentialism in his works (starting from Lewis 1968 to Lewis 1986) that make differences in how to interpret him.

For Lewis$_B$:

> 1$_B$. Socrates is essentially human iff every relevant counterpart of Socrates is human.

However, in both cases, the same counterparts matter in order to determine the truth-values of essentialist claims: given a context C, the possible objects whose similarity to Socrates is relevant according to C. Therefore, 1$_A$ and 1$_B$ end up being equivalent for determining the truth-conditions of 1). This implies, as we will see presently, that there are no significant differences between Lewis$_A$ and Lewis$_B$, when it comes to the evaluation of the truth-values of essentialist claims.

Let us call "*de re* representations" of an object *O* the similarity relations that determine the *de re* modal properties of *O*. Therefore, in Lewis's view, the *de re* representations of *O* are the similarity relations that are relevant in a context.

4. Lewis and DE semantical

Let us go back to 1):

> 1) Socrates is essentially human.

As we saw, for Lewis, 1) is true iff all the (relevant) counterparts of Socrates are human. This means that the general form of the truth-conditions for an essentialist sentence type is incomplete: it needs to be completed with the input of a (relevant) counterpart relation, and we know that that is a contextual matter: O is a (relevant) counterpart of Socrates iff O is similar enough to him under relevant respects, but it is a matter of context which respects of similarity are salient and which grades of similarity are enough under such respects. The (relevant) counterparts of Socrates are therefore determined to a large extent by the contexts in which 1) is produced and evaluated. Therefore, Lewis gives complete truth-conditions only for specific tokens of 1). Different tokens of the same sentence type about Socrates might be produced and evaluated in different contexts and, thus, evoke different (relevant) counterparts of him. So, according to different contexts, different *de re* representations may figure in the content of the utterance of 1), and, hence, 1) might have different truth-values in different contexts. Accordingly, Lewis rejects the semantic constancy of essentialist claims, namely, he rejects DE semantical.

This means that, for Lewis, according to different contexts, different properties are said to be essential to an object. Note that I am not saying that an object is said to have different essential properties according to different contexts. Indeed, if essentialist claims are semantically inconstant, it would be misleading to talk about "essential properties", since there are only properties that, according to a context, are said to be essential to an object.

5. Lewis and DE metaphysical

DE metaphysical is the thesis that what makes essentialist sentences true (the truth-making properties, if you like) must be facts in the worlds which are independent of the context.

So, let us see, in Lewis's view, what ultimately makes essentialist sentences, such as 1), true. We know that what would make sentence 1) true is the fact that all the (relevant) counterparts of Socrates are human. Let us suppose that, in a context C, Socrates's (relevant) counterparts are Socrates, *O* and *P*. Well, if 1) is true in C, this is so by virtue of the fact that Socrates, *O* and *P* are all human, that is by virtue of the fact that they are all similar by being all human. However, we saw that it is the job of metaphysics to establish similarity relations between objects, since they are based on the fact that objects have the properties they have, which, in turn, depends on how the worlds are made. So, it is a matter of metaphysics if *O* and *P* are similar to Socrates under some respect (they all share the property of being human) that happens to be contextually relevant. Therefore, 1), if true, is made true by facts in the worlds which are independent of the context.

Accordingly, Lewis accepts DE metaphysical: for Lewis, what makes an essentialist sentence true in a context are facts in the worlds which are independent of the context. This means that metaphysics can establish the properties that, according to some context, are determined to be essential to Socrates. Note that I am not saying that metaphysics can establish which *essential properties* Socrates has, because (and again) it would be misleading to talk, in this view, about "essential properties": metaphysics can only establish which properties the objects (Socrates, *O* and *P*) exemplify; then those properties, according to some context (namely, when *O* and *P* are deemed to be (relevant) counterparts of Socrates), are required by the truth-conditions of 1) and, hence, are determined to be essential to Socrates. In other words, there are only properties (that are exemplified independently of context) that, according to a context, are determined to be essential to objects.

Saying that Lewis accepts DE metaphysical means to say that, at the level of metaphysics, no contextual facts are involved in the attributions of essentiality to some properties of individuals. And, in this sense (maybe a shallow sense), DE survives in Lewis's metaphysics. Note also that, not even at the metaphysical level, there are relevant differences between Lewis_A and Lewis_B.[7]

6. Paul's account

Let us see now Paul's account of how objects have properties as a matter of *de re* modality, which is crucial to her defence of DE (I will refer mainly to Paul 2004, 2006).

Paul takes ordinary objects to be nothing more than bundles of properties, such that bundling is a type of mereological fusion. The sum of the basic non-modal properties of an object *O* is its "core" (the composition that gives rise to the cores is called "qualitative composition", that is a fusion of properties). The fact that *O*'s core is in a counterpart relation (namely, a similarity relation) to some possible object which has a property *F*, generates the relational property (*Rprop*) of being *de re* represented as having *F* (*Rprop*F). In this way, *O* is *de re* represented as having *F*. The *Rprops* are thus monadic relational properties ontologically generated by the core of the object standing in a counterpart relation to *possibilia*, and they are included in the sum that is the object (the composition between a core and the *Rprops* it generates is called "modal composition", which is a kind of qualitative composition). Therefore, *O* is a sum of its core properties plus the *Rprops* of being *de re* represented in certain ways.

Then, if *O* includes *F* and the *Rprop* of being represented as not-*F*, then *O* is accidentally *F*. If *O* includes *F* and lacks the *Rprop* of being represented as not-*F*, then *O* is essentially *F*.

Therefore, in this perspective, for an object to have a *de re* modal property, both the core and the *Rprops* must be included in the sum that is the object. That is to say that, while we would intuitively take an object to be only its core (and Lewis with us), for Paul objects are sums of their cores plus the *Rprops*. The *de re* modal properties of an object supervene on the sum of object's core plus the *Rprops* and, as such, they are included as well in the sum that is the object (see Paul 2006, 353).

[7] The differences between them will become clear later.

Paul then claims that modal composition is unrestricted. Therefore, given a core (what we would take to be the object), and given that anything is similar to anything else under some respect, there are many objects composed by that core. In other words, where we thought there was only one object, there are instead many more objects. By having the same basic non-modal properties, namely the same core, these objects have a lot in common: they occupy precisely the same spatiotemporal region. Nonetheless, they occupy different regions of "modal space": they differ from each other in terms of inclusion or exclusion of some *Rprop* and, so, in which *de re* modal properties they include. Therefore, we have a proliferation of objects and modal profiles (these different sums are different objects with different natures): many different objects with clear modal boundaries, metaphysically carved at their joints. However, Paul claims, in most non-philosophical contexts we would deny the existence of many of these objects. Nonetheless, according to her, they exist.[8]

Let "*O*" be the name of an object. Since the sum of that object's core and the *Rprops* generated by the core is unrestricted, then there are many sums which share the same core, and they are all plausible candidates to be the denotation of "*O*". Hence, there are many different sums that we can pick out when we use "*O*": for instance, the maximal sum of core properties and all the *Rprops* generated by that core, or some proper parts of this maximal sum. Paul maintains that it is the context that helps us to determine the denotation of "*O*".

Paul claims that, thanks to her theory of objects, she can disagree with Lewis about an important point that marks the passage from a shallow to a deep kind of essentialism. For Paul, the reason why Lewis is a shallow essentialist is that he accepts the inconstancy of *de re* representations. And Lewis overtly admits that he endorses inconstancy of *de re* representations. We have seen, indeed, that the reason why Lewis rejects DE semantical is precisely because objects, according to different contexts, may have different (relevant) counterparts and, so, the *de re* representations that figure in the content of the utterance of an essentialist sentence may change according to different contexts. Paul claims that, unlike Lewis, she is able to take *de re* representations to be constant.

[8] According to Leslie (2011), any characterization of essentialism is bound to accept such a proliferation of entities, each with distinct modal profiles, that she calls "plenitude". Interestingly, she claims that Lewis's theory, instead, can avoid the argument to plenitude: where essentialism postulates multiple entities, Lewis postulates multiple counterparts. As we will see in the following, I will argue that this is one of the reasons why we should prefer Lewis's proposal of essentialism over Paul's.

Before delving into Paul's strategy for achieving the hoped constancy of *de re* representations, it is useful to stress what I take to be the key moves of such a strategy. The first key move is semantical: Paul postulates that ordinary names such as "Socrates" are vague inasmuch they can refer (in this world) to different objects. The second key move is metaphysical: she includes in the sums she takes to be the objects the *Rprops* and the *de re* modal properties and, from this assumption, she derives that there exist many more objects that we thought, metaphysically carved at their joints.

Paul's proposal is clever, captivating, ambitious, and deserves careful consideration. Despite my forthcoming arguments that Paul can (partially) eliminate the inconsistency of *de re* representations only by sacrificing the very idea at the core of deep essentialism, I believe it is her great merit to have acknowledged the significance of including contextual factors in the analysis of essentialism.

In the following pages, I will analyze what we can derive from Paul's key moves with regard to the coveted constancy of *de re* representations, by starting from Paul's stance toward DE semantical.

7. Paul's semantic move and DE semantical

Let us go back to our essentialist sentence type 1):

1) Socrates is essentially human.

According to Paul, context helps us to determine which object the name "Socrates" refers to, among the multitude of sums (in this world) that are eligible candidates for the denotation. Therefore, the general form of the truth-conditions for 1) is incomplete: it needs to be completed with the input of the denotation of "Socrates". However, different tokens of the same term type "Socrates" might refer to different objects (in this world), which have different *de re* modal properties. This means that, in Paul's view, 1) is such that in some contexts is true, and in some other context may be false.[9] Therefore, Paul also embraces the inconstancy of the truth-values of essentialist claims. In other words, Paul also rejects DE semantical.[10]

[9] It should be evident that Paul, as well as Lewis, also accepts variations in truth-values across referentially equivalent essentialist sentences: different coreferential expressions can clearly evoke different sums as well.

[10] Paul might deny that there is such a context-dependence in the individuation of the reference of ordinary names, by saying, for instance, that there are many homonymous names which give

However, since Paul is not totally clear about her rejection of DE semantical, it is important to see if there is any relevant difference between Paul and Lewis's accounts, when it comes to semantics.

Let us consider Paul's semantic move of postulating the vagueness of ordinary names. Well, the first semantic consequence that such a move has is to shift the source of the semantic inconstancy of essentialist sentences (I will simply refer to it as "the shift"): while in Lewis's case the semantic inconstancy is due to the variability of Socrates's *de re* representations, in Paul's view, it is due to the semantic indeterminacy of "Socrates" (see Paul 2004, 180). Might it be the case then that, even if Paul rejects DE semantical, she can still guarantee, by virtue of the shift, the constancy of *de re* representations, as she claims? No, she cannot.

Indeed, in her view, given the semantic vagueness of the name "Socrates", different contexts can pick out different objects as the reference for that name. However, by selecting an object, the context also selects its *Rprops* (which are included in the object). But, since the *Rprops* are generated from the object's core standing in some *de re* representations, then, by selecting an object, the context also selects the object's *de re* representations. Therefore, *de re* representations figure in the content of the utterance of an essentialist sentence for Paul as well as they do for Lewis. And, in both accounts, their inconstancy is the source of the rejection of DE.

So far, the only way in which Paul can get constancy is to say that, once we have specified which object "Socrates" denotes, the same *de re* representations will figure in all the utterances of essentialist claims about that object. Indeed, as we saw, once the context picks out one object, it also selects the object's *de re* representations. Then, of course, in all the essentialist claims about that object, the same *de re* representations will figure. But the same happens with Lewis. Once a context is fixed, we know which *de re* representations figure in the content of the utterance of an essentialist claim about, say, Socrates. Hence, the same *de re* representations will figure in all the essentialist claims about Socrates. For instance, let us suppose again that, in the context of evaluation of 1), Socrates's (relevant) counterparts are Socrates, *O* and *P*. Well, given that context, the same *de re* representations will figure in all the utterances of essentialist claims about Socrates.

different references. However, in this case, she should convince us that, in our ordinary language, there are many more homonymous names than we expected there to be.

So, for instance, Paul can claim that:

> on the shallow essentialist view, an object like Humphrey can be *de re* represented in many different, conflicting, ways. Suppose that the person I refer to as 'Humphrey' is essentially descended from his parents, Ragnild and Hubert. A (…) deep essentialist will thus hold that this person, Humphrey, is not *de re* represented as having counterparts with different parents, no matter what the context. (Paul 2006, 350)

But, in this quote, the only difference between the two cases is that Paul supplies the context only for the latter example (her case). Indeed, by saying "Suppose that the person I refer to as 'Humphrey' is essentially descended from his parents", Paul is supplying the context: she is individuating the reference of the name "Humphrey" as that sum that includes Humphrey's core and does not include the *Rprop* for different origins. Of course, given a context and, with it, the object that the context picks out, then, in every essentialist claim about that very same object, the same *de re* representations that the object's core entertains will occur. But the same would apply to the first case (Lewis's case) if we supply a context: given a context and, with it, Humphrey's (relevant) counterparts (say that they all have the same origins as Humphrey), in every essentialist claim about that very same object, the same *de re* representations will figure.

Therefore, so far, the difference between the two approaches is quite superficial. Paul maintains that *de re* representations are constant; however, what she really believes is that, once a context is supplied and, with it, the object that is the reference of the name (and, with it, the counterpart relations that the object's core entertains) they are fixed. Lewis explicitly claims that *de re* representations are inconstant; nonetheless, what happens is that, once a context is supplied and, with it, the (relevant) counterparts, they are fixed. And it is precisely because they both accept, at the end of the day, the inconstancy of *de re* representations that they both reject DE semantical.

This means that, semantically speaking, Paul's stance mirrors Lewis's: (i) the selection of the *de re* representations is a contextual matter, therefore (ii) essentialist claims are inconstant, and (iii) according to different contexts, different properties are said to be essential to objects. Let us move on, now, to the metaphysical level, in order to see if, at that level, there are significant differences between the two approaches.

8. Paul's metaphysical move and DE metaphysical

We know that similarity relations are metaphysically determined, and that they are easy to get. According to Paul, the similarity relations *simpliciter* can enjoy the status of counterpart relations. This means that, given Socrates's core, (CoreS), such a core can be fused with many different *Rprops*. For instance, since there is a similarity relation between CoreS and a cat, under some respect, CoreS will also generate the *Rprop* for being represented as the cat (*Rprop*C). Modal composition is unrestricted, so CoreS can combine with the *Rprops* that it generates in many different ways. So, for instance, we have the maximal sum of CoreS plus all the *Rprops* that it can generate (Sum$_n$), and all the proper subsets of Sum$_n$ (Sum$_1$, Sum$_2$, and so on). Sum$_n$ will have all its properties accidentally, while its subsets, by ruling out some *Rprop,* will have some properties essentially.

Let us take one of these subsets: Sum$_1$, which has CoreS plus the *Rprop* for being represented as a musician (*Rprop*M). Sum$_1$ will be accidentally a philosopher and essentially human (there is no *Rprop* for being represented as non-human: for instance, the *Rprop*C is not included in this sum). The fact that Sum$_1$ is essentially human, and here is Paul's metaphysical move, is determined by virtue of *constant de re* representations. Indeed, Sum$_1$ is carved at its joints, metaphysically speaking. Surely, metaphysics cannot establish if Sum$_1$ is the best reference for "Socrates". As Paul says (see 2006, 362), there is nothing that allows us to select one object rather than another, when we talk about Socrates: no object stands out as the privileged reference of "Socrates". Nonetheless, metaphysics can establish the boundaries between Sum$_1$ and, say, Sum$_n$: these objects are carved at their joints. This means that it is not the context that determines which *de re* representations of Sum$_1$'s core are relevant: they are all at the same level. And, since metaphysics distinguishes between Sum$_1$ and Sum$_n$, and since Sum$_1$ includes only some of the *Rprops* generated by its core (only the *Rprop*M), then we get *metaphysically selected de re* representations, namely constant *de re* representations. And such constant *de re* representations are able to determine the *essential properties* of Sum$_1$.

Let me stress the point that, in Paul's view, should make the difference: Sum$_1$ has essential properties independently of context. That is, we are no longer talking about properties that, according to a context, are determined to be essential to Sum$_1$, but about its *essential properties*. And this can happen because such properties are established by virtue of constant *de re* representations: similarity relations, whose relevance is metaphysically determined. Indeed, the *Rprop*M is the only *Rprop* included in Sum_1, which metaphysics carves at its joints; so, the similarity relation that Sum_1's core entertains to the musician is metaphysically picked out, among all the similarity relations that that core entertains, since only the *Rprop* that it generates is included in Sum_1.

However, does this really amount to saying that Paul can guarantee constant *de re* representations, as she intends to? Does this make any relevant difference with Lewis for what regard DE metaphysical?

Well, recall that, in Lewis's view, according to different contexts, different *de re* representations of Socrates are deemed to be relevant. Because of this, essentialist claims are inconstant and, hence, according to different contexts, different properties are said to be essential to Socrates. However, and here is Lewis's acceptance of DE metaphysical, it is the business of metaphysics to establish the properties that, according to some context, are determined to be essential to Socrates.

Is the situation really different on Paul's view? Well, I just said that metaphysics can establish essential properties of Sum_1, inasmuch they are determined by constant *de re* representations. However, only according to some context, Sum_1 is the reference of "Socrates". This means that we are back where we were: according to different contexts, different *de re* representations of Socrates are deemed to be relevant (since different objects are regarded as relevant). Because of this, essentialist claims are inconstant and, hence, according to different contexts, different properties are said to be essential to Socrates. And Paul accepts DE metaphysical, for the very same reasons why Lewis accepts it: metaphysics establishes the properties that, according to some context, are determined to be essential to Socrates.

This means that, in both accounts, *de re* representations are inconstant and, in neither account, metaphysics can determine the *essential properties* of Socrates, but only the properties that, according to some context, are required by the truth-conditions of an essentialist sentence about him and, thus, are determined to be essential to him.

The only way to make a difference between the two accounts, therefore, is to insist on focusing on what happens to, say, Sum_1, which is an object for Paul, and, as I said, does have *essential properties*, determined by constant *de re* representations. Indeed, when Paul says that she can guarantee that objects have essential properties and are *de re* represented in a constant way, she does not intend to talk about what we would take to be objects (like Socrates), rather she means to talk about objects like

Sum$_1$. And, surely, Lewis cannot guarantee that any of his objects have any of those characteristics.
However, is there really any improvement if we focus on objects, like Sum$_1$, which have essential properties and constant *de re* representations? I say no. I will claim that, far from being an improvement, it goes to Paul's disadvantage that she guarantees that. Before going there, a little detour might be useful.

9. Lewis$_A$ and Lewis$_B$

We need to go back to the distinction between Lewis$_A$ and Lewis$_B$. Recall that the main difference between them lies in the fact that only Lewis$_B$ accepts counterpart relations *simpliciter*. I said that, from this difference, we cannot infer any significant consequence with regard to their stances toward DE, neither semantical or metaphysical. However, Paul draws an important consequence.

From Lewis$_B$'s perspective, and from the fact that anything is similar to anything else under some respect, we inferred that there will be as many counterparts of Socrates as many relations of similarity there are. Now, according to Paul's understanding of Lewis$_B$, such counterparts *simpliciter* determine the *de re* modal properties of *O* (see Paul 2006, 344ff.).[11] I claim that, if Paul so interprets Lewis$_B$, then she should accept that Lewis$_B$ has constant *de re* representations as well. Especially because this does not change her supposed advantage over Lewis$_B$.

Indeed, if counterparts *simpliciter* determine the *de re* modal properties of *O*, then such *de re* modal properties are individuated once and for all, independently of the context: they are determined by similarity relations, whose relevance is established independently of context. Indeed, we know that, in this view, metaphysics cannot distinguish among *O*'s counterparts, since no one stands out as the privileged counterpart. However, if they are all kept at the same level, without any arbitrary privileging, then they are all determined to be equally relevant, from a metaphysical point of view. This means that Lewis$_B$ too has constant *de re* representations.[12] However, I do not call such properties "essential",

[11] I do not agree with this step in Paul's interpretation of Lewis$_B$. However, my aim in this paper is to show that Lewis's account, as Paul defines it, is not less deep than Paul's. According to my interpretation of Lewis$_B$, I might make things easier for me.

[12] The claim that Lewis$_B$ has constant *de re* representations clearly clashes with what has been said so far: Lewis$_B$ endorses inconstancy of *de re* representations. However, as I understand Paul, there is a way to make the two things compatible. As we will see, such constant *de re* representations

despite their being context-independent, for an important reason. *O* will have so many counterparts (and metaphysics cannot privilege one over another) that almost no property will come out as essential to it: they all will turn out to be accidental. So, *O* will have context-independent *accidental* properties.[13]

Therefore, here is the difference between Paul and Lewis$_B$, that is relevant for Paul. In both cases, we have constant *de re* representations: *de re* representations that determine the *de re* modal properties of objects, whose relevance is determined independently of context. However, Paul has constant *de re* representations by virtue of the fact that metaphysics can pick out some *de re* representation as relevant, by carving the objects at their joints. By contrast, Lewis can have constant *de re* representations only because, since metaphysics cannot discriminate among them, then it establishes that they are all equally relevant. The consequence is that only in Paul's case, constant *de re* representations can determine *essential properties* of objects (recall Sum$_1$, which is metaphysically determined to be essentially human), while in Lewis's view, metaphysics establishes that all the properties of Socrates come out as *accidental*, and we need to appeal to the contextual selection of the relevant *de re* representations in order to talk, in some context, of properties that are determined to be essential to him.[14]

So, Paul is entitled to claim that, only according to her account, (some) objects have essential properties that are determined by constant *de re* representations. In the next Section, I am going to argue that, rather than being an asset to Paul's account of essentialism, this aspect actually worsens the situation (in Section 11, I will discuss other problems for Paul's account).

determine that all Socrates's properties are accidental. So, in some sense, such constant *de re* representations are not to be taken into account if we want to talk about the *essential* properties of Socrates (or the properties that are determined to be essential to him). By contrast, if the context selects the relevant *de re* representations, then we are allowed, for instance, to say that Socrates is essentially human, since, in common contexts, we can ignore some of Socrates's counterparts. So, it is only by virtue of the *de re* representations, whose relevance is determined by the context, that we can have properties that are said to be essential to Socrates in a context.

[13] The reason why these passages are not doable for Lewis$_A$ should be clear: according to Lewis$_A$, there are no counterparts *simpliciter* that might determine context-independent *de re* modal properties of individuals.

[14] Note, anyway, the parallel with Paul: she does have constant *de re* representations that determine *essential properties* of what she takes to be objects. Nonetheless, she needs to appeal to inconstant *de re* representations in order to talk about the properties that are determined to be essential to Socrates.

10. A problem with Paul's account of DE

I explained why Paul's metaphysical move of having objects (cores plus *Rprops*) carved at their joints is relevant for her account of DE: since metaphysics establishes the boundaries between one sum and another, Paul can have objects (like Sum_1) with constant *de re* representations that determine essential properties.

Therefore, Paul needs to be cautious about the question as to whether modal composition, namely the fusion of a core with the *Rprops* it generates, is unrestricted or restricted. Indeed, if metaphysics must establish the boundaries between objects, then it must be a matter of metaphysics if a composition between a core and some *Rprops* is admissible or not. Therefore, either (i) modal composition is unrestricted and, hence, any sum is allowed, so that there is no arbitrary privileging of some sum over another, or (ii) if modal composition is restricted and, hence, not all the sums are allowed, there must be some metaphysical justification for limiting the composition.

Despite the fact that Paul clearly recognizes how important this matter is for her defence of DE (Paul 2006, 359), I do not find her discussion about this point very clear.[15] In the following, I raise a problem for modal composition that has to do with sums of "modal incompatibilities". I will argue that (i) if modal composition is totally unrestricted, then there are sums with modal incompatibilities, to the effect that the very idea at the bottom of DE is jeopardized; (ii) if modal composition is somehow restricted, the restrictions we would need in order to avoid those sums are either not metaphysically justified, or somehow available to Lewis too, so to cancel the differences between the two accounts.

Let us start by assuming that modal composition is unrestricted. Recall the previous example, in which metaphysics establishes both Sum_n and Sum_1's *de re* modal properties, and while Sum_1 is essentially human,

[15] She clearly claims that qualitative composition that generates the objects' cores is restricted in order to avoid sums of incompatible properties, such as the golden mountain or the round square (see Paul 2006, 360; 2016, 41). However, since she recognizes how problematic is to determine the conditions under which composition occurs, she endorses a brute restriction (Paul 2016, 39) (or, for instance in Paul 2002, she works with unrestricted composition, for the sake of simplicity). Then, she claims that, for the reasons I explained in the main text, she accepts unrestricted modal composition (see Paul 2006, 360-366). Nonetheless, since modal composition is a kind of qualitative composition, it seems to inherit some restrictions. So, modal composition is said to be only "minimally restricted" (Paul 2006, 361). However, as I said, restrictions for qualitative composition are taken to be a brute fact, and there is no indication as to whether such restrictions apply to modal composition only because they apply to the composition of the cores, or also in order to avoid sums of incompatible *modal* properties (that Paul, at any rate, does not discuss).

Sum_n is accidentally human. Now, it is bad enough that it is metaphysically established that a human being could have been a cat. However, as bad as it is, it is nothing new for us. We saw that Lewis_B, according to Paul's interpretation, also accepts that (and this holds regardless of whether or not I am right in saying that Lewis_B has constant *de re* representations). In addition, Lewis_B has no way to say that, metaphysically speaking, some human is essentially human. Therefore, so far so good for Paul: her account is deeper than Lewis's.

However, let us take the sum of CoreS (Socrates's core) plus the *Rprop*C (the *Rprop* for being represented as a cat) and nothing else (we are assuming that composition is unrestricted, therefore, there must be such a sum). This sum (Sum_2) is carved at its joints. So, metaphysics can establish which *de re* modal properties Sum_2 has. Well, Sum_2 will be accidentally human (being represented as a cat). However, it will be essentially, say, intelligent and hairy. Here, I am assuming that CoreS is composed by non-modal properties such as "being intelligent", "being hairy", "being a philosopher" and so on. Since it has only a cat as a counterpart, it is represented by this counterpart as not-human, perhaps as not-philosopher and so on. However, it is not represented as not-hairy and not-intelligent (and here I am assuming that the cat is intelligent and hairy). Therefore, metaphysics allows for the existence of a human being (recall that we cannot call it "Socrates", otherwise we are back to semantics), which is accidentally human but, at the same time, essentially hairy. Therefore, Paul has objects with essential properties determined by constant *de re* representations. However, such essential properties are not the properties we expected them to be. And they are not even the properties Paul wanted.

Indeed, Paul aims to defend DE because she wants essential properties to define the nature of things. For this reason, Paul claims, such properties cannot be context-dependent (Paul 2006, 345). But what do we have now that DE is back? We do have context-independent essential properties of objects, but they do not always define the nature of things. Indeed, if there is something that defines the nature of a human being, it should be his being human, and certainly not his being hairy. Using her own words, "there is very little content in the idea that object has to be a certain way in order for it to exist" (Paul 2006, 346).

To sum up, totally unrestricted modal composition allows for sums of modal incompatibilities. The fact that metaphysics allows for the existence of a human being that has all of his properties accidentally, the property of being human included, is bad enough, both for Lewis_B and

Paul. But saying that metaphysics allows for the existence of a human being that could have been a cat but not bald is definitely worse.

Let us suppose, then, that modal composition is minimally restricted. Are there metaphysical justifications for limiting modal composition to the effect that sums of modal incompatibilities are not admissible?

Note that it might be said that, if a sum combines incompatible properties, then it simply does not exist: metaphysics does not carve its joints. Therefore, the thought goes, there is no need to put restrictions on what sums there exist in order to avoid such sums: they simply are impossible. Accordingly, modal composition can be unrestricted, since such sums cannot exist. In response to such a thought, I would like to say that DE, as an account of the nature of objects, should explain why the nature of an object stops it from being essentially hairy and accidentally human: DE should explain why the combination of CoreS only with *Rprop*C would make for an impossible object (Sum$_2$). If DE's explanation were only that, if the combination of CoreS only with *Rprop*C makes for an impossible object, then there is no fusion of them, there would be no explanation of what DE was supposed to explain. So, it cannot be simply said that Sum$_2$, since it is impossible, it does not exist: we need to know *why* CoreS cannot combine only with *Rprop*C. Therefore, if there are metaphysically justified restrictions on modal compositions, and if, according to these restrictions, Sum$_2$ does not exist, only then could it be claimed that Sum$_2$ is impossible.[16]

So, let us see if there are metaphysically justified restrictions that can be imposed on modal composition, to the effect that Sum$_2$ is ruled out:

a) It might be said that, for any sortal property *S* included in a core *C*, there can be no fusion between *C* and the *Rprop* for being *non-S*. So, if a core includes the property of being human, it can never combine with the *Rprop* for being not-human (in this way, we also exclude Sum$_n$). However, one thing is to say that it is clearly wrong to have objects that are accidentally human and essentially

[16] Note that, in a context where she is discussing only about qualitative composition (2016, 43), Paul talks about primitive modal constraints that limit the composition: "deep modal facts" that should prevent, for instance, two properties from combining. But, of course, recurring to such brute deep modal facts for explaining modal composition would be circular: if there are such facts, we expect DE to explain them. To be sure, Paul says that such deep modal facts are supposed to be *de dicto* modal truths. But without any account of such alleged facts, we cannot rely on them for explaining modal composition.

hairy. Another thing is to determine the properties we want to come out as essential to an object and, then, establish some rule for making it impossible that they come out as accidental. Unless there is some metaphysical justification for the fact that a core that includes *S* cannot combine with a *Rprop* for being *non-S*, we cannot assume something that DE was supposed to explain.

b) If we appeal to natural properties, then there might be something in the metaphysics that allows us to say that an object that is *S* cannot combine with the *Rprop* for being *non-S*. For instance, it might be said that, if "being human" is a natural property and a core includes such a property, then that core cannot combine with a *Rprop* for being not-human (again, we would exclude also Sum$_n$). And this might solve the problem: it is metaphysics that selects which properties are natural, then it would be the job of metaphysics to establish that similarity relations that generate *Rprops* for being not-human are not acceptable. However, if this solution works for Paul, then it should hold for Lewis$_B$ too (and perhaps for Lewis$_A$ as well), who acknowledges natural properties in his metaphysics (see, for instance, Lewis 1983): he would have *natural counterpart relations* that stand out as metaphysically privileged for determining the essential properties of individuals (despite my belief that this interpretation does not faithfully capture Lewis's thought—Nencha 2017—see Buras 2006, who defends such an argument). But then, Lewis's essentialism would be as deep as Paul's.

Therefore, (i) if modal composition is unrestricted, then sums of modal incompatibilities are admissible, to the effect that the very idea at the bottom of DE gets lost; (ii) if modal composition is restricted, the restrictions put in place in order to avoid sums of modal incompatibilities are either not metaphysically justified or available to Lewis too, to the effect that the difference between the two accounts would be lost.

What happens, instead, from Lewis$_A$ and Lewis$_B$'s perspectives, with regard to objects with incompatible modalities? Well, also Lewis might say that a human being is essentially hairy but accidentally human. This happens if there is a context that makes it true. However, Lewis does not intend to say that such *de re* modal properties metaphysically define the object's nature: it is only a matter of context.

Therefore, in order to draw a relevant difference between Paul's account and Lewis's, modal composition must be taken to be unrestricted and, so, sums of modal incompatibilities are to be accepted. In this way, it can be said that Paul, unlike Lewis, accepts that some of the sums that she takes to be objects have context-independent essential properties determined by constant *de re* representations. However, such an achievement plays against Paul herself: it would have been better if it were only a matter of context. Indeed, what Paul can really guarantee is that some of the things she takes to be objects have essential properties, and such essential properties, *only in some case,* define their natures. Therefore, the sense itself of DE is lost.

In the following, I will briefly discuss some other aspects of Paul's theory that make, in my opinion, her account worse than Lewis's. Perhaps, some of the following disadvantages could have been acceptable for the benefit of having DE, if DE had not had the consequence just discussed.

11. Lewis's view is in better shape than Paul's

Let us start from semantic aspects. I said that, from a semantic point of view, both authors reject DE semantical. However, I think that the context-dependence that Paul hypothesizes is more problematic than the one postulated by Lewis. Indeed, Paul seems to be relying on the hypothesis that names are vague in ways that we do not expect them to be. That is, she needs to convince us that ordinary names are candidates for semantic indeterminacy. By contrast, Lewis's postulation that the relevance of an object's similarity relations changes according to different contexts leaves the reference theory unchanged. That is, according to this perspective, ordinary names, such as "Socrates", are not treated as candidates for semantic indeterminacy, precisely as we do not expect them to be.

Moreover, it is not only that we are said that ordinary names, contrary to what we expect, are vague. But we are also said that such a vagueness is very pervasive: *all* the ordinary names come out as semantically vague. Indeed, in Paul's view, vagueness would be a matter of any name whatsoever insofar as it denotes an object that is eligible for having accidental properties. Perhaps, "God" would not be a vague name in this perspective, as long as the object that it is supposed to denote might be expected to have all of its properties essentially. So, Paul should convince us to believe in a semantic vagueness so pervasive that basically all ordinary names in our language are vague.

Finally, ordinary names tend to occur in the superficial form of the essentialist sentences. Yet we do not have any evidence of such alleged indeterminacy. For Lewis, instead, the indeterminacy concerns a predicate (the predicate for the counterpart relation) that emerges only when the sentence is analyzed in counterpart theoretic terms. Therefore, since such a predicate does not occur in the superficial form of the essentialist sentences, it is less surprising that there are not evidences of its indeterminacy.[17]

From a metaphysical point of view, the alleged multiplication of entities in Paul's account is replaced by a multiplicity of counterpart relations in Lewis's view, which seems to be a less expensive ontological commitment. Note also that Paul's strategy to make us accept such a proliferation of objects is the standard strategy to distinguish between the things that exist and the things over which we quantify in usual contexts: such a multitude of objects does exist, but we usually ignore them. And by "usual contexts" she mainly means "non-philosophical contexts" (see for instance, Paul 2004, 183; 2006, 361). However, Paul seems to disregard the fact that, since these objects occupy different regions of modal spaces, we would be confronted with such a multitude every time we want to talk about the *de re* modal properties of objects. But, far from being specific of philosophical contexts, modal talk is everyday talk. I agree with Paul that, in non-philosophical contexts, we would not select as relevant references for "Socrates" those objects according to which Socrates could have been non-human. However, there are many other sums that we would have to take into account in our everyday talk.

Therefore, not only is Paul able to (partly) rule out from her account of essentialism the context-dependence only with high costs, but also the place where she puts such a dependence, namely, the reference theory, with all the consequences that derive from such a move, worsens the situation.

12. Conclusion

In this paper, I discussed Paul's proposal for a form of deep essentialism that, according to her expectations, retains some of the advantages of shallow essentialism while being classified as deep. The key difference,

[17] What is more, the context-sensitivity of modal expressions is almost universally accepted, and, in Lewis's view, the predicate for the counterpart relation explains *de re* modal expressions.

according to Paul, between her proposal and Lewis's should revolve around the postulation of constant *de re* representations.

I argued that, ultimately, Paul's proposal either is not effectively different from Lewis's account, as Paul assumes, or, if there are indeed substantial differences, they are to Paul's own detriment. Moreover, there are difficulties that her proposal faces, both from a semantical and a metaphysical perspective, that Lewis's account does not encounter.

Therefore, if Paul is correct (and I think she is) in asserting that certain sceptical challenges pose problems for standard forms of deep essentialism that they do not for shallow essentialism, but it is better to be deep than to be shallow, then the arguments presented in this paper could be interpreted as suggesting that Lewis's account should be reconsidered, since, as shallow as it can be, it might be deeper than it looks.

Acknowledgments

I am deeply grateful to Giorgio Lando, Daniele Sgaravatti, Pasquale Frascolla, and Andrea Bottani for dedicating their time to help me enhance this paper. I would also like to thank the anonymous reviewers of this journal for their valuable remarks.

REFERENCES

Buras, Todd. 2006. "Counterpart theory, Natural Properties and Essentialism." *The Journal of Philosophy* 103 (1): 27–42. https://doi.org/10.5840/jphil2006103138.

De, Michael. 2020. "A Modal Account of Essence." *Metaphysics* 3 (1): 17–32. https://doi.org/10.5334/met.33.

Della Rocca, Michael. 1996. "Recent Work in Essentialism, Part 1." *Philosophical Books* 37 (1): 1–13.

Divers, John. 2007. "Quinean Skepticism about De Re Modality after David Lewis." *European Journal of Philosophy* 15 (1): 40–62. https://doi.org/10.1111/j.1468-0378.2007.00229.x.

Goodman, Nelson. 1972. "Seven strictures on similarity." In *Problems and Projects*, edited by Nelson Goodman, 437–447. Indianapolis: Bobbs-Merrill.

Hazen, Allen. 1979. "Counterpart-Theoretic Semantics for Modal Logic." *The Journal of Philosophy* 76 (6): 319–338. https://doi.org/10.2307/2025472.

Heller, Mark. 2005. "Anti-essentialism and Counterpart Theory." *The Monist* 88 (4): 600–618. https://doi.org/10.5840/monist200588428.

Leslie, Sarah-Jane. 2011. "Essence, plenitude, and paradox." *Philosophical Perspectives* 25 (1): 277–296. https://doi.org/10.1111/j.1520-8583.2011.00216.x.

Lewis, David Kellogg. 1968. "Counterpart Theory and Quantified Modal Logic." *The Journal of Philosophy* 65 (5): 113–126. https://doi.org/10.2307/2024555.

Lewis, David Kellogg. 1983. "New Work for a Theory of Universals." *Australasian Journal of Philosophy* 61 (4): 343–377. https://doi.org/10.1080/00048408312341131.

Lewis, David Kellogg. 1986. *On the Plurality of Worlds*. Malden, Massachusetts: Wiley-Blackwell.

Nencha, Cristina. 2017. "Natural Properties Do Not Support Essentialism in Counterpart Theory: A Reflection on Buras's Proposal." *Argumenta* 2 (2): 281–292. https://doi.org/10.23811/46.arg2017.nen.

Paul, Laurie Ann. 2002. "Logical Parts." *Noûs* 36 (4): 578–596. https://doi.org/10.1111/1468-0068.00402.

Paul, Laurie Ann. 2004. "The Context of Essence." *Australasian Journal of Philosophy* 82 (1): 170–184. https://doi.org/10.1080/713659794.

Paul, Laurie Ann. 2006. "In Defense of Essentialism." *Philosophical Perspectives* 20 (1): 333–372. https://doi.org/10.1111/j.1520-8583.2006.00110.x

Paul, Laurie Ann. 2016. "A one category ontology." In *Being, Freedom, and Method: Themes from the Philosophy of Peter van Inwagen*, edited by John A. Keller, 32–61. New York: Oxford University Press.

Quine, Willard Van Orman. 1953. "Three Grades of Modal Involvement." *Proceedings of the XIth International Congress of Philosophy* 14: 65–81. https://doi.org/10.5840/wcp11195314450.10.5840/wcp11195314450

Wildman, Nathan. 2016. "How (not) to be a modalist about essence." In *Reality Making*, edited by Mark Jago, 177–196. Oxford, United Kingdom: Oxford University Press.

European Journal of Analytic Philosophy EuJAP | 2024 | Vol. 20 | No. 1 | 55-86
https://doi.org/10.31820/ejap.20.1.3

THE LOGICAL POSSIBILITY OF MORAL DILEMMAS IN EXPRESSIVIST SEMANTICS: A CASE STUDY

Ryo Tanaka[1]

[1] University of Tokyo, Japan

Original scientific paper – Received: 09/09/2023 Accepted: 26/02/2024

ABSTRACT

In this paper, using Mark Schroeder's (2008a) expressivist semantic framework for normative language as a case study, I will identify difficulties that even an expressivist semantic theory capable of addressing the Frege-Geach problem will encounter in handling the logical possibility of moral dilemmas. To this end, I will draw on a classical puzzle formulated by McConnell (1978) that the logical possibility of moral dilemmas conflicts with some of the *prima facie* plausible axioms of the standard deontic logic, which include *obligation implies permission*. On the tentative assumption that proponents of ethical expressivism should be generally committed to securing the logical possibility of moral dilemmas in their semantic theories, I will explore whether and how expressivists can successfully invalidate *obligation implies permission* within the framework developed by Schroeder. The case study eventually reveals that this can indeed be a hard task for expressivists. Generalizing from the case study, I will suggest that the source of the difficulty ultimately lies in the mentalist assumption of the expressivist semantic project that the logico-semantic relations exhibited by normative sentences should be modeled in terms of the psychological attitudes that speakers express by uttering them. My final goal will be to show that the difficulty expressivists face in dealing with the logical possibility of moral dilemmas is a reflection of the more general problem that their commitment to the mentalist assumption prevents them from flexibly adopting or dropping axioms in their semantic theories to get the right technical results.

Keywords: expressivism; moral dilemmas; metaethics; semantics; deontic logic.

Correspondence: tanaka.r29@gmail.com

1. Introduction

In this paper, using Mark Schroeder's (2008a) expressivist semantic framework for normative language as a case study, I will identify difficulties that even an expressivist semantic theory capable of addressing the Frege-Geach problem will encounter in handling the logical possibility of moral dilemmas. To this end, I will draw on a classical puzzle formulated by McConnell (1978) that the logical possibility of moral dilemmas conflicts with some of the *prima facie* plausible axioms of the standard deontic logic, which include *obligation implies permission*. On the tentative assumption that proponents of ethical expressivism should be generally committed to securing the logical possibility of moral dilemmas in their semantic theories, I will explore whether and how expressivists can successfully invalidate *obligation implies permission* within the framework developed by Schroeder. The case study eventually reveals that this can indeed be a hard task for expressivists. Generalizing from the case study, I will suggest that the source of the difficulty ultimately lies in the mentalist assumption of the expressivist semantic project that the logico-semantic relations exhibited by normative sentences should be modeled in terms of the psychological attitudes that speakers express by uttering them. My final goal will be to show that the difficulty expressivists face in dealing with the logical possibility of moral dilemmas reflects the more general problem that their commitment to the mentalist assumption prevents them from flexibly adopting or dropping axioms in their semantic theories to get the right technical results.

In the remainder of the introduction, I will address three preliminary issues. First, I will introduce the puzzle concerning the logical possibility of moral dilemmas that I will discuss in this paper. Second, I will explain and briefly justify the tentative assumption of the paper that expressivists generally need to secure the logical possibility of moral dilemmas in their semantic theories. Lastly, I will explain why I specifically draw on Schroeder's framework to develop my discussion. The subsequent sections will be devoted to the case study: I will show how the problem concerning the logical possibility of moral dilemmas for ethical expressivism arises taking a specific shape in Schroeder's framework and explore how one can respond to it.

1.1 Ethical expressivism and the logical possibility of moral dilemmas

A moral dilemma is defined as a situation where incompatible courses of action A and B are both morally obligatory for an agent. The reality of such dilemmas in human life seems indubitable. One can easily imagine, and often find oneself in, situations where general moral precepts such as "Do not lie", "Do not steal", "Help your family and friends", may come to conflict with one another. One can see the reality of moral dilemmas most vividly in situations where one and the same moral precept seems to generate conflicting but equally strong demands. Consider the often-cited case of Sophie's choice (Styron 1979): Sophie and her two children are at a Nazi concentration camp, and a guard tells Sophie that only one of her children will be allowed to live but the other will be killed. For each child, Sophie has an obligation to save him/her, but she cannot save both. In this case, it is implausible to think that Sophie can resolve the conflict by thinking that one of her obligations overrides the other, because there is no obvious reason why either of them should be stronger than the other. This and similar examples suggest that genuine, that is, irresolvable dilemmas are possible and often real.[1]

As McConnell (1978, 2022) points out, however, the possibility of irresolvable moral dilemmas apparently conflicts with some of the *prima facie* plausible axioms of the standard deontic logic. (The presentation of the problem below follows McConnell (1978)). Crucially, the problem concerns the *logical* possibility of moral dilemmas—adopting the relevant axioms leads to the result that moral dilemmas are impossible as a matter of *logic and the meanings of the relevant normative expressions alone*. For the purpose of this paper, I will specifically focus on the two axioms, which are meant to capture the following *prima facie* plausible theses: (1) *permission can be defined in terms of obligation*, and (2) *obligation implies permission*.[2] Let OA stand for "A is obligatory" and PA stand for "A is permissible". First, a moral dilemma is a situation where incompatible courses of action are both obligatory. To capture its troublesome nature, one can characterize a moral dilemma as a situation where the following holds:

(MD) $OA \wedge O{\sim}A$.

A moral dilemma is a situation where A is obligatory, but not doing A is obligatory as well because it is necessary for doing B, another obligatory action—e.g., saving one child's life is obligatory for Sophie, but not doing so is obligatory as well because it is a necessary means for saving the life of the other child, which is another obligatory act for her. Second,

[1] For influential arguments for the logical possibility of moral dilemmas that invoke the notion of moral residue, see e.g., Marcus (1980), Tessman (2015), and Williams (1966).

[2] Brink (1994), for example, takes these as conceptual truths concerning the notions of obligation and permission.

(1) states that doing A is permissible if and only if it is not the case that not doing A is obligatory:

(1) $PA \Leftrightarrow \sim O \sim A$.

Lastly, the symbolic representation of (2), *obligation implies permission*, is the following:

(2) $OA \Rightarrow PA$.

(1) and (2) entail $OA \Rightarrow \sim O \sim A$. Assuming the material conditional, $OA \Rightarrow \sim O \sim A$ is equivalent to $\sim(OA \wedge O\sim A)$. This, however, directly contradicts (MD). Hence, (1) and (2) are jointly inconsistent with (MD). Adopting the apparently intuitive theses (1) and (2) as axioms in one's theory thus rules out the logical possibility of moral dilemmas.

Since McConnell (1978) provided a formal presentation of the puzzle, many answers have been proposed in the literature.[3] Those who think moral dilemmas should be at least logically possible seek ways to justify abandoning (at least) one of the axioms that give rise to the inconsistency, whereas opponents argue that any putative solution to the puzzle is bound to be *ad hoc*.[4] (At this point, it may need to be noted that there are also other combinations of axioms that are known to be inconsistent with the logical possibility of moral dilemmas. I will come back to this point below.)

In this paper, for the sake of discussion, I tentatively assume that expressivists should generally side with the pro-dilemma view.[5] A *prima facie* justification for this assumption stems from the fact that ethical expressivism is usually construed as a *semantic view,* that is, a view about *the meanings of normative sentences in a natural language,* such as English. (In the next section, I will elaborate on this point in more detail.) As such, proponents of ethical expressivism should be committed to formulating a theory that correctly reflects ordinary speakers' use of, and

[3] For a helpful survey, see McConnell (2022). For various responses to McConnell (1978), see e.g., Almeida (1990), Conee (1982), Goble (2009), Hansson (2019), Holbo (2002), Marcus (1980), McConnell (1978), Nair (2016), Sinnott-Armstrong (1988), Vallentyne (1989), and Zimmerman (1996). See e.g., Lemmon (1962) and van Fraassen (1973), for influential discussions of this topic that precede McConnell (1978).

[4] Those who deny the logical possibility of moral dilemmas often point to the fact that historically influential philosophers such as Aquinas, Mill, Kant, Ross etc. seem to hold similar views. See Marcus (1980), McConnell (2022, 6), Sinnot-Armstrong (1988) for discussion.

[5] As I will explain, the plausibility of the conclusions of the paper will not depend on the truth of this assumption. Certainly, all things considered, it might turn out that one's semantic theory should give up the possibility of moral dilemmas, rather than *obligation implies permission*. However, to avoid unnecessary complications, I will not address this point until Section 4.

linguistic intuitions about, the target normative expressions—they constitute the data that should guide one's theory construction. Crucially, ordinary speakers' language use and linguistic intuitions seem to suggest that they take moral dilemmas to be at least conceptually possible. Ordinary speakers do often find themselves in dilemmatic situations and describe them as such, and, accordingly, they do not seem to take "A is obligatory for S, but not-A is also obligatory for S" as an utterly confused, non-sensical statement. To accommodate the data, one's expressivist semantic account should not entail that moral dilemmas are impossible, *as a matter of logic and meaning alone*. When conjoined to the puzzle concerning the logical possibility of moral dilemmas presented above, this means that expressivists should construct their semantic theories so that *obligation implies permission* will not turn out to be formally valid (assuming that they do not want to reject the interdefinability of obligation and permission).[6]

To see the intuitive plausibility of this assumption, it may also be helpful to point out the fact that those who deny the possibility of moral dilemmas are usually motivated by substantive moral-theoretic concerns that do not necessarily coincide with ordinary speakers' language use and linguistic intuitions (McConnell 2022, Sec. 3 and 4). For example, a Kantian theorist might argue that moral dilemmas should be impossible, because one's (deontological) moral theory should be *uniquely action-guiding* in the sense that it will never prescribe incompatible actions for an agent in a given situation. To meet this requirement, one might propose that there should be some way of hierarchically structuring moral precepts so that irresolvable conflicts will never arise (for an influential critique of this idea, see e.g., Ross 1930, Ch. 2). Whether this line of reasoning is plausible or not, it clearly concerns a requirement that is to be imposed on one's *substantive theory of morality*, not one's *semantic account* of normative expressions in a natural language. Semantics must respect the fact that ordinary speakers often talk about moral dilemmas meaningfully and they do seem to take dilemmas to be at least conceptually possible. The assumption of this paper is that this

[6] One might question why expressivists need to reject *obligation implies permission* rather than the interdefinability of obligation and permission. In this paper, I will explore the former option mainly for pragmatic reasons (I leave open, but will not discuss in detail, the possibility of pursuing the other option). As I explain in the next paragraph, Schroeder himself discusses this problem by focusing on the question how one may (in)validate *obligation implies permission* in his framework. Schroeder does not pursue the alternative route since he thinks it "is an old observation that 'permissible', 'impermissible', 'obligatory', and 'unobligatory' can all be interdefined using negation" (Schroeder 2008a, 46). Furthermore, as I will explain in Section 3, rejecting the interdefinability of obligation and permission in Schroeder's framework actually turns out to be more difficult compared to *obligation implies permission*.

constitutes a *prima facie* reason for expressivists to try to secure the logical possibility of moral dilemmas in their semantic accounts. Whether the possibility of moral dilemmas should be ultimately excluded in one's substantive moral theory is a separate question, which I do not intend to address in the current paper.

In the rest of the discussion in this paper, as a case study, I will examine how this general problem concerning the logical possibility of moral dilemmas arises taking a specific shape in the expressivist semantic framework developed by Mark Schroeder (2008a) and explore how one might respond to it. There are two reasons why I specifically focus on Schroeder's expressivist semantics to explore this issue. The first is that, although ethical expressivism is one of the most discussed views in the literature on the semantics of normative language, apparently no proponents of expressivism have done a better job than Schroeder in actually constructing a semantic theory that seems to provide a promising response to the so-called Frege-Geach problem (Geach 1960, 1965). The Frege-Geach problem, in short, is a demand for expressivists to develop a *compositional* semantic theory that correctly captures logical-semantical relations between non-atomic, logically complex normative sentences (e.g., negations, disjunctions, conditionals, sentences with quantifications, etc.). In most expressivist proposals before Schroeder, the details of expressivist semantic theories were not fully spelled out, and it was even unclear if they could adequately explain such basic semantic phenomena as the logical inconsistency of "A is wrong" and "A is not wrong"—as I will explain in the next section, this is what Unwin (2001) calls the "negation problem" for expressivists, which is an instance of the Frege-Geach problem as applied to negation. An important contribution of Schroeder's work is that it identifies a *structural* requirement that any expressivist semantic account should meet to deal with the Frege-Geach problem.[7] Exploring the issue of the logical possibility of moral dilemmas in his framework will provide a useful case study, because my initial goal is to show that an expressivist semantic theory that provides an adequate solution to the Frege-Geach problem does not necessarily succeed in securing the logical possibility of moral dilemmas as well. The second reason is purely pragmatic: at one place in his book, Schroeder (2008a, Ch. 5, Sec. 4) himself discusses the relevant thesis, *obligation*

[7] For this reason, Schroeder's expressivist semantics is worth exploring also for those who are attracted to the recent movements of applying the expressivist idea to other types of discourse than the ethical. See e.g., Bar-On and Sias (2013) for discussion.

implies permission, in connection to the logical possibility of moral dilemmas. I will develop my own discussion by building on Schroeder's.[8]

Relatedly, as I noted earlier, there are also other combinations of intuitively plausible axioms that are known to be inconsistent with the logical possibility of moral dilemmas. For example, *obligation implies possibility* (i.e., *ought implies can*) and the principle of agglomeration, which states that *if A is obligatory and B is obligatory, then A and B is obligatory* (Williams 1966), are jointly inconsistent with the possibility of moral dilemmas (I will not provide a proof here—interested readers should consult surveys such as McConnell 2022, Sec. 4 and McNamara and Van De Putte 2023, Sec. 6.4.) If expressivists need to side with the pro-dilemma view, they will ultimately have to find ways to avoid such combinations as well. Due to the limit of space, I will not explore other combinations and focus only on *obligation implies permission.* In Section 4, I will briefly discuss how one might interpret the results of this paper in light of this broader point.

The rest of the discussion in this paper will proceed as follows. In Section 2, I will briefly review several basic features of ethical expressivism construed as a semantic project and introduce Schroeder's expressivist semantic theory against that general background. As noted, Schroeder's theory is mainly motivated as a response to the negation problem, which is a special instance of the Frege-Geach problem. The discussion may get technical in places, and to avoid unnecessary complications, in Section 2, I will *not* discuss how the technical tools developed by Schroeder connect with the issue of the logical possibility of moral dilemmas. I will turn to this issue in Section 3: I will explore how one might respond to the problem of securing the logical possibility of moral dilemmas within Schroeder's framework and make some general observations from the case study. In Section 4, I will conclude by articulating the moral of the case study in the most general terms.

The last caveat before the main discussion: in presenting Schroeder's expressivist semantics, I will, as Schroeder himself does, focus on the predicate "is wrong" as the main target of semantic analysis, instead of

[8] Also, it should be noted that the overall aim of Schroeder's book *Being For* is to illustrate the *costs* of ethical expressivism (see *Preface*). Schroeder's intention is to reveal the theoretical commitments that expressivists should make by actually developing a workable expressivist semantic theory on behalf of them. Throughout the book, he occasionally reminds readers to think about where to "get off the boat"—it is beyond the scope of the discussion in this paper to assess the overall plausibility of Schroeder's project construed as a *reductio* of ethical expressivism. See, however, Section 4 below for a brief discussion on the prospect of the expressivist semantic project in light of the result of the current paper.

predicates such as “is obligatory” or “is permissible” that I used in introducing the issue of the logical possibility of moral dilemmas. As I will explain in Section 3, however, there is an easy and relatively uncontroversial way to translate claims that contain the predicate “is wrong” to those that only contain “is obligatory” and “is permissible” (for example, “A is obligatory” could be rephrased as “Not A is wrong”). Although this invites some complication, it does not pose any serious problem for the main discussion of this paper. Schroeder himself also notes that his discussion is applicable to expressivist views that take different normative predicates (e.g., “is rational”, “is the thing to do”, “ought”, and so on) as basic (Schroeder 2008a, 7, see, also, 39). I will come back to this point in Section 3.2.

2. Schroeder’s expressivist semantics: A structural solution to the negation problem

Ethical expressivism is characterized by the idea that what one does when uttering a sentence with a normative predicate is to express one’s non-cognitive attitude toward an object of evaluation. The basic expressivist idea can be traced to early non-cognitivist views proposed by Ayer (1936), Hare (1952), and Stevenson (1937); Blackburn (1984, 1988, 1998) and Gibbard (1990, 2003) are known for more systematic formulations of contemporary expressivist views (for a brief history of ethical expressivism, see Schroeder 2010, Chapter 4). On a simple expressivist account, for example, “Murder is wrong” might express a speaker’s non-cognitive attitude of disapproval of murder. *Non-cognitive* attitudes contrast with cognitive attitudes (e.g., one’s belief that murder has such and such properties) in that the former do not have truth-evaluable propositions as their contents. Importantly, as I mentioned in the previous section, expressivism is usually construed as a *semantic* view. To develop a semantic theory for descriptive language, one can fruitfully invoke the notion of truth-evaluable proposition. To develop a semantic theory for normative language, expressivists insist, a different approach is called for —the meanings of normative sentences are *not* truth-conditions. Instead, their meanings are the non-cognitive attitudes that speakers express by uttering them.[9] Hereafter, I call this the *mentalist*

[9] It should be noted that some authors argue, against the orthodoxy, that expressivism need not be interpreted as a *semantic* thesis (Bar-On and Chrisman 2009, Bar-On et al. 2014). In Section 4 (in the last footnote), I will briefly discuss how one might interpret the results of this paper in connection with this kind of “neo-expressivist” positions.

assumption of the expressivist semantic project.[10] It is the assumption that normative sentences must get their meanings from speakers' *mental states*, instead of propositions, which are usually taken to be abstract entities that exist independently of speakers' psychology.

On this construal, one major task for proponents of expressivism is to provide a systematic account of the logical and semantic features that normative sentences in our natural language seem to exhibit, without assuming the standard truth-functional compositional semantics developed primarily for descriptive language. Specifically, the mentalist assumption of the project requires that they should somehow model the logico-semantic relations exhibited by normative sentences in terms of the psychological attitudes speakers express by uttering them. Logicians and formal semanticists can explain (among many other things) why a descriptive sentence, say, "Grass is green", is inconsistent with "Grass is *not* green", by appealing to the truth-functional definition of the meaning of "not" and the meaning (i.e., truth-condition) assigned to the original sentence. If expressivism is a view about meaning of normative language, it is expected that proponents of the view should be able to explain, in some parallel way, why "Murdering is wrong" is inconsistent with "Murdering is not wrong". Informally, in the standard truth-conditional semantics, what negation does when applied to a descriptive sentence is to "flip" the truth-value assigned to the sentence. However, because expressivists understand the meanings of normative sentences in terms of non-cognitive attitudes they express instead of truth-conditions, what negation does when applied to normative sentences should be explained in a different way.

Of course, there is no principled reason why one ought to think that it is impossible to develop an expressivist semantic theory for normative language that meets this challenge.[11] Attempts have been made, most notably by Simon Blackburn and Allan Gibbard, at sketching outlines of compositional semantic theories for ethical language based on the expressivist assumptions (Blackburn 1984, 1988, 1998; Gibbard 1990, 2003). However, as Schroeder contends (2008a, *Preface*), there was no consensus whether they even succeeded at providing a plausible explanation of how "Murder is wrong" should be logically inconsistent with "Murder is not wrong"—this is known as "the negation problem"

[10] This corresponds to what Sias (2024) calls "semantic ideationalism" in his survey entry on ethical expressivism.

[11] See, e.g., Hare (1970) for an expression of optimism about the prospect of non-cognitivism construed as a semantic project.

for expressivism (Unwin 2001). As mentioned above, the negation problem is an instance of the so-called Frege-Geach problem, which questions how proponents of expressivism (or non-cognitivism in general) could provide a compositional semantic account for non-atomic, logically complex normative sentences in such a way that the theory can correctly capture their logical-semantical relations. In Part II of the book, Schroeder sets out to develop his expressivist semantic account primarily by responding to the negation problem formulated by Unwin (see, also, Schroeder 2008b, 2008c). In the rest of this section, I will introduce Schroeder's account by explaining how it is tailored to deal with the negation problem.

The negation problem, as formulated by Unwin, is this. Expressivists maintain that "Murdering is wrong" expresses one's non-cognitive attitude toward murder—let us stipulate that it expresses the attitude, *disapproval of murder*. Then, what attitude should be assigned as the meaning of "Murdering is not wrong", the sentence that should turn out to be logically inconsistent with the original sentence? At the first glance, *disapproval of not murdering* might seem to be a good candidate, because it seems logically inconsistent to disapprove of both φ-ing and not φ-ing. However, this cannot be right, because *disapproval of not murdering* should be, intuitively, the attitude that is to be expressed by "Not murdering is wrong" instead of "Murdering is not wrong". These sentences clearly have different meanings, and no adequate semantic theory should conflate the meaning of one with that of the other. One might think that there should be some way of getting around the problem by inserting "not" in the right places in the attitudes expressed by the relevant sentences, but Unwin's discussion shows that the problem cannot be solved so easily.

According to Schroeder (following Unwin), the negation problem arises from the "*insufficient structure"* (Schroeder 2008a, 57) in the attitudes expressed by, and thereby assigned as the meanings of, normative sentences. (Hereafter, all the references are to Schroeder (2008a) unless otherwise noted.) One can best see this point by looking at the following table (45; slightly modified from the original):[12]

w Jon assents to "Murdering is wrong".

[12] Schroeder, following Unwin, uses "Jon thinks that murdering is wrong" and so on in demonstrating the negation problem. Here and in the relevant places below I will use "Jon assents to 'Murdering is wrong'" instead to highlight the fact that the problem primarily concerns which attitudes should be assigned as the meanings of normative *sentences*.

n1 Jon does not assent to “Murdering is wrong”.
n2 Jon assents to “Murdering is not wrong”.
n3 Jon assents to “Not murdering is wrong”.

w* Jon disapproves of murdering.
n1* Jon does not disapprove of murdering.
n2* ???
n3* Jon disapproves of not murdering.

The problem, in short, is that the account allows for too *few* ways to negate w*. There are only two places where one can insert “not” in *Jon disapproves of murdering* (which yield n1* or n3*), whereas there are three ways to negate w (n1, n2, and n3). Specifically, as I explained, expressivists need the attitude expressed by “Murdering is not wrong” (n2) to be inconsistent with *disapproval of murdering*, which is expressed by “Murdering is wrong” (w) —but, at the same time, the attitude in question cannot be *disapproval of not murdering*, because it should be assigned as the meaning of “Not murdering is wrong” (n3). Apparently, then, there seems to be no way of arriving at the correct semantic assignments for w, n1, n2, and n3, starting from the assumption that the meaning of “*φ*-ing is wrong” is one’s *disapproval of φ-ing*.[13]

One way to avoid the problem might be to think that “Murdering is not wrong” should express a different kind of attitude than disapproval, such as one’s *tolerance of murdering*. This might allow expressivists to explain the inconsistency between “Murdering is wrong” and “Murdering is not wrong” by appealing to the stipulation that disapproval of φ-ing is inconsistent with tolerance of φ-ing (Blackburn 1988). However, Schroeder argues that this is a problematic move because it leaves completely unexplained why “two *distinct* and apparently *logically unrelated* attitudes [i.e., disapproval and tolerance] toward the *same* content” (48) can be logically inconsistent with one another. Schroeder contrasts this to the unproblematic kind of inconsistency that holds between two attitudes of the *same* kind toward inconsistent contents (ibid.). In Schroeder’s terminology, these are “inconsistency-transmitting attitudes”:

[13] For more detailed presentations of the negation problem, see Unwin (2001) and Schroeder (2008a, Ch. 3, 2008b, 2008c).

> **Inconsistency-transmitting attitudes**: An attitude *A* is inconsistency-transmitting just in case two instances of *A* are inconsistent just in case their contents are inconsistent. (43)

Belief is a good example: believing that p is inconsistent with believing that not-p, *because* their contents, p and not-p, are logically inconsistent. In other words, in the case of belief, the inconsistency of the contents p and not-p *transmits* to one's attitudes towards these contents. And insofar as the idea that belief is an inconsistency-transmitting attitude is generally accepted, there is no reason why expressivists cannot treat, say, disapproval as an inconsistent-transmitting attitude and assume that one's *disapproval of murdering* is logically inconsistent with one's *disapproval of not murdering*. On the other hand, Schroeder contends, it is not justified for expressivists to take it for granted that *disapproval of murdering* should be logically inconsistent with *tolerance of murdering*. On his view, this is a purely *ad hoc* solution to the negation problem, because it is a mere convenient stipulation that there should be non-cognitive mental attitudes of disapproval and tolerance such that they are completely distinct but nonetheless can be logically inconsistent with one another in some way. What makes this stipulation particularly problematic is the fact that, unlike inconsistency-transmitting attitudes, there are no undisputed good examples of attitudes that exhibit the desired feature (47-9).[14]

Taking stock: on Schroeder's view, the negation problem arises from the lack of structure in the attitudes (e.g., *disapproval of φ-ing*) that expressivists assign as the meanings of normative sentences. Furthermore, there is also the constraint that expressivists should not explain the inconsistency between normative sentences by stipulating the existence of multiple attitudes (e.g., disapproval and tolerance), each of which are primitive but nonetheless can be logically related.

Schroeder's main positive proposal defended in the book is that one can resolve the negation problem on behalf of expressivists by replacing the attitude of disapproval with a primitive inconsistency-transmitting non-cognitive attitude that he calls "*being for*" (58). Schroeder's overall strategy is to use this, and *only* this, attitude as the basic tool for constructing meanings for all normative sentences (hence, the title of the book, *Being For*). For the purposes of this paper, it would not be necessary to discuss Schroeder's exposition on the psychological nature

[14] However, for an important critique of Schroeder's argument summarized in this paragraph, see Baker and Woods (2015).

of the attitude in question. The key point is that the attitude of being for creates the necessary *structure* that was missing in the expressivist semantic analysis that adopts the attitude of disapproval as the basic explanatory tool. On Schroeder's proposal, "φ-ing is wrong" expresses the attitude of *being for blaming for φ-ing*. Crucially, unlike *Jon disapproves of murdering*, there are three, instead of two, places to insert "not" in *Jon is for blaming for murdering.* The semantic analysis of w, n1, n2 and n3 that results from this proposal is shown in the following table (59):

w Jon assents to "Murdering is wrong".
n1 Jon does not assent to "Murdering is wrong".
n2 Jon assents to "Murdering is not wrong".
n3 Jon assents to "Not murdering is wrong".

w** Jon is for blaming for murdering.
n1** Jon is not for blaming for murdering.
n2** Jon is for not blaming for murdering.
n3** Jon is for blaming for not murdering.

Hereafter, following Schroeder's notation, I will abbreviate "being for blaming for φ-ing" as "FOR(blaming for φ-ing)". On Schroeder's proposal, one can explain the inconsistency between "Murdering is wrong" and "Murdering is not wrong" by the fact that the attitudes expressed by these sentences—i.e., FOR(blaming for murdering) and FOR(not blaming for murdering)—have inconsistent *contents*. On the assumption that the attitude of being for is (like belief) inconsistency-transmitting, these are inconsistent attitudes *because* they have inconsistent contents.

Thus, analyzing the meanings of w, n1, n2, and n3 in terms of being for attitudes provides expressivists with a systematic way of correctly capturing their logical relationships without making any controversial assumptions. More formally, the meanings of (i.e., the attitudes expressed by) any normative sentences that contain negation can be determined compositionally by applying the definition of negation, provided on p. 66:

(NEG) Where 'A' expresses FOR(α), '~A' expresses FOR(~α).[15]

[15] In a similar fashion, Schroeder provides recursive definitions for conjunction and disjunction on page 66; he also defines entailment relationship between sentences on page 70.

From this definition, it follows that “Murdering is not wrong” expresses FOR(not blaming for murdering), that is, (n2**). And, it is also natural to think that “Not murdering is wrong” expresses FOR(blaming for not murdering). This assignment of attitude is intuitive, and, more importantly, the assigned attitude is distinct from the one expressed by “Murdering is not wrong”. The analysis thus avoids conflating the meanings of “Not murdering is wrong” and “Murdering is not wrong”. This, Schroeder argues, resolves the negation problem on behalf of expressivists.

For the sake of discussion, I assume that Schroeder’s proposal summarized above provides a promising solution to the negation problem. Here, I want to highlight two features of Schroeder’s semantic framework that will be important for the purposes of the discussion below. Crucially, both derive from the fact that the negation problem is a *structural* problem and the solution requires adding the necessary structure to the attitudes assigned as the meanings of normative sentences. First, Schroeder’s discussion, if successful, implies that any proponent of expressivism should ultimately adopt a semantic theory that at least shares the basic structure with Schroeder’s account—that is, the structure that allows one to deal with the negation problem. As he puts it, adopting his semantic framework “isn’t just *a* way of making progress on the negation problem, for expressivists”—rather, it is “*the* expressivist solution to the negation problem” (61). Second, this need not mean, however, that expressivists should adopt the semantic theory that analyzes “Murdering is wrong” in terms of the attitude of being for blaming for murdering, *specifically*. Since the negation problem is a structural problem, any account that yields sufficient structure should be able to deal with it, at least in principle. Schroeder himself notes in passing that he sticks with this specific analysis “just to fix examples” (58) and one could adopt an alternative stipulation that analyzes the meaning of “Murdering is wrong” as “being for *disapproving* of [murdering]” (ibid., emphasis mine), instead of being for *blaming* for murdering. Elsewhere, he also considers a proposal that “Murdering is wrong” expresses “being for avoiding murdering” (74). So, depending on one’s interests and pre-theoretical intuitions, which *specific* kind of attitude/act should be taken as the target of the being for attitude may vary, as long as it retains the structure necessary for dealing with the negation problem.

To summarize, the general conclusion of Schroeder’s discussion is that the attitude that is to be assigned as the meaning of “φ-ing is wrong” should take the following form: *being for [one's preferred unary expression] φ-ing*.[16] This is the structural requirement that one's semantic theory should meet to avoid the negation problem (and, more generally, to deal with the Frege-Geach problem). However, at the same time, this structural requirement does not entail any strong material restriction on which specific kind of attitude/act should go into the placeholder. In principle, any unary expression that takes a gerund (φ-ing) as the object will do, as long as it does not yield an obviously implausible meaning assignment. [17] Some obvious candidates include blaming for φ-ing, disapproving of φ-ing, avoiding φ-ing, but there may also be others. With these in mind, in the next section, I will explore the issue of the logical possibility of moral dilemmas within Schroeder's semantic framework.

3. How expressivists can and should secure the logical possibility of moral dilemmas

As I noted in Section 1, in this paper I tentatively assume that an expressivist semantic account should aim to secure the logical possibility of moral dilemmas—as a semantic theory, it should respect the data that ordinary speakers do not take the conjunction of "φ-ing is obligatory" and "not φ-ing is obligatory" to be an utterly confused non-sensical statement. This at least requires that *obligation implies permission* should not turn out to be formally valid in the theory (again, there are also other combinations of axioms that one would need to invalidate, which I will not discuss in this paper—see Section 1). In this section, I will explore how one can achieve this task within Schroeder's framework.

The discussion will proceed in two steps. First, I will show that whether one can successfully achieve this task in Schroeder's framework ultimately depends on which specific kind of attitude/act one decides to put in the placeholder in the attitude assigned the meaning of "φ-ing is wrong", i.e., *being for [one's preferred unary expression] φ-ing*. Second, I will argue that it will be crucial for expressivists that their decision here should not turn out to be problematically *ad hoc*. Seen from a broader perspective, to decide which specific act/attitude should go into the

[16] Köhler (2017) also highlights the essentially structural nature of Schroeder's proposal to defend it from an objection raised by Skorupski (2012).

[17] There is a very weak material requirement on what kind of unary expression one can put into the placeholder: it should be, at least, some negative attitude/act toward the object of evaluation. (I thank an anonymous reviewer for pointing this out.) For example, putting *praising* into the placeholder would yield *being for praising for φ-ing*, which is structurally adequate but obviously implausible as a meaning assignment for "φ-ing is wrong". Notice that all of the candidates Schroeder considers (i.e., blaming for φ-ing, disapproving of φ-ing, avoiding φ-ing) meet this requirement.

placeholder is to answer a very basic question for expressivists: what is the non-cognitive attitude expressed by "φ-ing is wrong", after all? This is a question whose answer should be motivated by general semantic-psychological considerations, *not just* by whether or not the resulting theory can secure the logical possibility of moral dilemmas. The case study will eventually show that this in fact makes it difficult for expressivists to invalidate *obligation implies permission* to get the right technical result in their theory. My final goal will be to locate the ultimate source of the difficulty in the mentalist assumption of the expressivist semantic project itself.

3.1 The logical possibility of moral dilemmas in Schroeder's expressivist semantics

So far, following Schroeder, I have been focusing on the predicate "is wrong". To address the question whether one can make *obligation implies permission* formally invalid in Schroeder's framework, it is necessary to translate all the sentences that contain "is wrong" to the sentences that contain "is obligatory/permissible". The required translation is shown in the following table.

φ-ing is wrong	φ-ing is not permissible	Not φ-ing is obligatory
Not φ-ing is wrong	Not φ-ing is not permissible	φ-ing is obligatory
φ-ing is not wrong	φ-ing is permissible	Not φ-ing is not obligatory
Not φ-ing is not wrong	Not φ-ing is permissible	φ-ing is not obligatory

Table 1

The proof of the table only requires two uncontroversial assumptions: (a) "Not φ-ing is obligatory" is a translation of "φ-ing is wrong", and (b) obligation and permission are interdefinable (i.e., "φ-ing is permissible" can be defined as "Not φ-ing is not obligatory" and *vice versa*). These directly yield the result that "φ-ing is not permissible" is a translation of "φ-ing is wrong" (represented in the first row), and one can similarly prove the rest of the table.[18]

I believe that these are natural assumptions in extending Schroeder's analysis of "is wrong" to "is obligatory/permissible". In Schroeder's framework, (a) amounts to the idea that "φ-ing is wrong" and "Not φ-ing

[18] Here, I assume that the definition for negation (NEG) provided in the previous section is applicable to sentences with different predicates than "is wrong". Another important point is that one need not use the axiom *obligation implies permission* to prove this table.

is obligatory" express the same being for attitude, namely, FOR(blaming for φ-ing). (b) amounts to the idea that "φ-ing is permissible" and "Not φ-ing is not obligatory" express the same attitude, namely, FOR(not blaming for φ-ing). In fact, once one accepts (a), it is unclear how one can avoid (b) in Schroeder's framework: to reject (b), one would have to maintain that the pairs of the sentences in each of the rows (such as "φ-ing is permissible" and "Not φ-ing is not obligatory") express non-equivalent being for attitudes. It is questionable whether one can come up with any reasonably simple assignment of being for attitudes that meets this condition. (And, as I noted in footnote 6, Schroeder himself takes the interdefinability of obligation and permission to be uncontroversial anyway.)

As one can see from the table, the claim that "φ-ing is obligatory" implies "φ-ing is permissible" translates to the claim that "Not φ-ing is wrong" implies "φ-ing is not wrong". As I discussed in Section 1, whether *obligation implies permission* turns out to be formally valid in one's semantic theory is an important question, because of its connection to the classic puzzle concerning the logical possibility of moral dilemmas. The puzzle was that the logical possibility of genuine dilemmas conflicts with the apparently plausible theses, each of which one might be inclined to treat as an axiom in one's semantic theory: (1) permission can be defined in terms of obligation, and (2) obligation implies permission. As noted above, (1) is an independently plausible assumption, and, specifically in Schroeder's framework, it is difficult to find a way to reject it. Hence, one should either maintain that (2) *obligation implies permission* is not formally valid or admit that moral dilemmas are impossible as a matter of logic and semantics alone. As I will explain below, this also tracks how Schroeder himself pursues this matter.

At one point in the book, Schroeder (Ch. 5, Sec. 4) discusses a possible treatment of *obligation implies permission* in his framework, in connection with the issue of the logical possibility of moral dilemmas. There, Schroeder simply registers the fact that there are theorists who believe that moral dilemmas should be logically *impossible,* and he goes on to explore whether his semantics could accommodate such a claim. Unlike me, Schroeder does not make any assumption concerning whether or not expressivists in general should aim to secure the logical possibility of moral dilemmas in their semantics. (In the next subsection, I will explain the ramifications of the divergence in stance here.) His aim is, rather, to show that his semantic framework is compatible with either of the opposing views on this issue: it "can remain neutral on this question [about the logical (im)possibility of moral dilemmas], offering ways for those who like either result to capture their views" (74-5).[19] Specifically, Schroeder maintains that there is a way to "supplement our system with an auxiliary assumption that will yield the result that 'murdering is wrong' [i.e., 'not murdering is obligatory'] and 'not murdering is wrong' [i.e, 'murdering is obligatory'] turn out to be inconsistent" (72). So, on Schroeder's view, one can either adopt or reject the "auxiliary assumption" in question to reflect one's preferred view on the logical possibility of moral dilemmas. Below, let me introduce Schroeder's "auxiliary assumption" and explain how it will make moral dilemmas logically impossible in the current framework.[20]

The auxiliary assumption in question states that "blaming for not murdering entails not blaming for murdering" (73). According to Schroeder, this validates *obligation implies permission* in the current framework. Later, I will question exactly what the auxiliary assumption is claiming in substance and how one might justify it—for now, let us simply confirm the technical point first. Notice that (given Table 1) "φ-ing is obligatory" and "φ-ing is permissible" express the following attitudes, respectively.

φ-ing is obligatory (not φ-ing is wrong)	FOR(blaming for not φ-ing)
φ-ing is permissible (φ-ing is not wrong)	FOR(not blaming for φ-ing)

Table 2

The auxiliary assumption in question states that the following holds:

(AA) Blaming for not φ-ing entails not blaming for φ-ing.

This, Schroeder claims, results in the following entailment relation between being for attitudes that captures *obligation implies permission*:

(OP) FOR(blaming for not φ-ing) entails FOR(not blaming for φ-ing).

Strictly speaking, to move from (AA) to (OP), one would require an assumption that for any pair of being for attitudes, entailment relations that hold between the contents of the attitudes will reflect in the

[19] Schroeder's neutral stance toward this issue is also reflected in his comment that it should "pay to be cautious about building this [i.e., the result that 'murdering is wrong' and 'not murdering is wrong' turn out to be inconsistent] into our logic" (72, ft. 6).

[20] Unfortunately, the proof of this point is only sketched and is not fully worked out by Schroeder himself—I will try to remedy it here. I thank an anonymous reviewer for pointing out the need for making the proof more explicit.

corresponding entailment relations between the attitudes (that is, FOR(α) entails FOR(β) if and only if α entails β). Schroeder does not seem to explicitly discuss this, but let us accept it as a generalization of Schroeder's basic proposal that being for attitudes are inconsistency transmitting-attitudes—they are, recall, attitudes that are inconsistent with one another if and only if their contents are inconsistent. This is motivated by the general idea that being for attitudes are, like beliefs, attitudes such that their logical relationships (such as inconsistency) are reducible to the logical relationships that hold between the embedded contents. If one can assume this much, then the auxiliary assumption in question does capture *obligation implies permission* in Schroeder's framework.

Now, let us confirm how this will rule out the possibility of moral dilemmas. A moral dilemma is, again, a situation where φ-ing and not φ-ing are both obligatory for an agent. Below, I will demonstrate that "φ-ing is obligatory and not φ-ing is obligatoy" and *obligation implies permission* are jointly inconsistent in Schroeder's framework. The dilemma's conjuncts, "φ-ing is obligatory" and "not φ-ing is obligatory", express the following attitudes, respectively:

> "φ-ing is obligatory" expresses FOR(blaming for not φ-ing).
>
> "Not φ-ing is obligatory" expresses FOR(blaming for φ-ing).

Assigning the meaning for the dilemmatic statement, "φ-ing is obligatory and not φ-ing is obligatory", requires introducing the definition for conjunction that Schroeder provides on page 66:

> (AND) If 'A' expresses FOR(α) and 'B' expresses FOR(β), 'A&B' expresses FOR(α∧β).

Accordingly, "φ-ing is obligatory and not φ-ing is obligatory" expresses the following attitude:

> (MD*) FOR(blaming for not φ-ing and blaming for φ-ing).

Now, it needs to be shown that (MD*) and (OP) are jointly inconsistent. To proceed from here, one only needs to assume that having the attitude of FOR(α∧β) is equivalent to having the attitudes of FOR(α) *and* FOR(β). If this can be assumed, the proof is obvious. Having the attitude (MD*) amounts to having the following pair of the attitudes, (1) FOR(blaming for not φ-ing) and (2) FOR(blaming for φ-ing). (1) and (OP) immediately yield FOR(not blaming for φ-ing). This is directly inconsistent with (2). Hence, (OP) and (MD*) are jointly inconsistent.

This should suffice to show Schroeder is right to claim that the auxiliary assumption in question rules out the logical possibility of moral dilemmas in his framework. According to Schroeder, one can then either adopt or drop the auxiliary assumption depending on one's view on the logical possibility of moral dilemmas. Now, the question I want to pursue below is this. Can an expressivist really adopt or drop the auxiliary assumption that flexibly, as Schroeder seems to assume? A potential worry stems from the point that I set aside earlier. In claiming that *blaming for not murdering entails not blaming for murdering*, one actually seems to be making a substantive claim about blaming. That is, whether it is true or not seems to depend on what blame actually is, or how the notion of blame should be understood. If so, the auxiliary assumption is making a claim whose plausibility may need to be examined *independently of* one's view on the logical possibility of moral dilemmas. This might mean that expressivists cannot in fact adopt or drop the auxiliary assumption as they like to deal with the logical possibility of moral dilemmas. (In the current paper, unlike Schroeder, I am assuming that expressivists generally need to side with the pro-dilemma view. So, the question I will focus on in the next subsection is this: can expressivists reject the auxiliary assumption freely, just because they need to invalidate *obligation implies permission* and make moral dilemmas logically possible?)

Schroeder, in fact, seems to recognize this sort of concern himself. In the same section, Schroeder points out that if one wishes to adopt the auxiliary assumption to make moral dilemmas logically impossible in the proposed semantic framework, one may need to justify it by maintaining that "it is [as a matter of conceptual necessity] impossible to both blame for murdering and blame for not murdering" (73). This, in effect, is to justify the auxiliary assumption by maintaining that it expresses a conceptual truth about blaming (as Schroeder puts it, a truth in the "logic of blaming" (73)). In passing, however, Schroeder also notes that this may actually seem "a little too strong for plausibility" (74). Although Schroeder does not elaborate on this, certainly we may imagine someone who insists that one can consistently blame someone for not φ-ing *and* blame the same person for φ-ing. For example, in Sophie's choice, Sophie is forced to choose only one child from the two, and, whichever child she ends up choosing, she might blame herself for not choosing the other. To take a more mundane example, one can imagine, say, a poor heavy smoker who will be blamed by her family and friends anyway regardless of whether she continues smoking or refrains from doing so.[21] The existence of this kind of practice concerning blame can certainly cast doubt on the idea that the auxiliary assumption expresses a conceptual truth about blame.

Generalizing from this point, I think one can arrive at an important observation: the plausibility of the auxiliary assumption in question depends on one's decision as to which act/attitude should go into the placeholder in the attitude of *being for [some unary expression] φ-ing*. Schroeder puts this point in this way:

> A different idea about 'wrong' is that 'murder is wrong' expresses being for avoiding murdering. On this account, the assumption required to yield the inconsistency is that it is impossible to both avoid murdering and avoid not murdering, which *is*, in fact, a highly plausible assumption about the logic of avoiding. So how easy it is to get 'murdering is wrong' and 'not murdering is wrong' to turn out to be inconsistent will obviously turn on which account we give of the attitude expressed by 'murdering is wrong'. (74)

The auxiliary assumption originally states: "blaming for not murdering entails not blaming for murdering" (73). If one decides that *avoiding,* instead of *blaming*, should go into the placeholder, the auxiliary assumption would have a different content, correspondingly: *avoiding not murdering entails not avoiding murdering*. Here, the latter might actually appear more plausible than the former, because it is highly unintuitive to think that an agent can avoid not φ-ing and avoid φ-ing at the same time—this looks similar to the case of intending φ-ing and intending not φ-ing at the same time, which seems impossible or at least deeply irrational. This in turn means that it can be highly controversial to *reject* this version of the auxiliary assumption (i.e., avoiding not φ-ing entails not avoiding φ-ing), because it appears to capture an independently plausible claim that follows from the "logic" of avoiding. On the other hand, if one instead decides that, say, *disliking* should be put into the placeholder, one would get *disliking not φ-ing entails not disliking φ-ing* as the corresponding auxiliary assumption. This may seem rather implausible—we can coherently imagine a universal hater who dislikes

[21] Imagine the following: if she smokes after dinner, her family might blame her for doing so by claiming that it harms their health; if she decides to refrain from smoking, her family might blame her for not smoking, claiming that there is an important value to sticking with one's habit and she should not be influenced so easily by others' advice. So, they will blame her anyway. This, at least, seems to capture what people do sometimes.

pretty much everything, including both φ-ing and not φ-ing. It would of course depend on how one understands the notion of disliking in question, but the point is that rejecting this version of the auxiliary assumption seems less controversial compared to the different version that one gets by putting *avoiding* in the placeholder.

Let me summarize the current point by connecting it to my exposition of Schroeder's semantics in the previous section. To adequately deal with the negation problem, one's expressivist semantic theory needs to meet the structural requirement that the attitude assigned as the meaning of "φ-ing is wrong" should take the following form: *being for [some unary expression] φ-ing*. Recall that this in itself does not call for any strong material restriction on which specific kind of expression should go into the placeholder. But now, there is at least one important consideration that one should take into account in making one's decision here: the plausibility of the auxiliary assumption, which validates *obligation implies permission*, depends on which specific attitude/act gets plugged into the placeholder. I think this is an interesting result that one can extract from Schroeder's discussion, which is worth pressing further than he actually does. Building on this point, in the next subsection I will explore how expressivists should ultimately deal with this issue and explain how this seemingly technical point actually exposes a more general problem for the expressivist semantic project.

3.2 Basic meaning assignment and its empirical implications

As I mentioned, Schroeder neither endorses nor rejects the auxiliary assumption himself—he merely presents it as an option that one can either adopt or reject, depending on one's view on the logical possibility of moral dilemmas. Although Schroeder's neutral stance is justified given the overall aim of his discussion, I think one can actually push this point further than Schroeder himself does to pose a general challenge for proponents of expressivism. One can do so by, for the sake of argument, sharing the assumption of the current paper that it is in fact a requirement for expressivists to secure the logical possibility of moral dilemmas. As I explained, this in turn requires (at least) invalidating *obligation implies permission* in one's semantic theory. So, although Schroeder simply allows expressivists to either accept or reject the auxiliary assumption, one can advance the discussion further by assuming that they are actually committed to rejecting it.

This leads to an important question. From the discussion above, we know that the plausibility of the auxiliary assumption depends on which specific attitude/act gets plugged into the placeholder in *being for [some* *unary expression] φ-ing*, assigned as the meaning of "φ-ing is wrong (is not permissible)". Now, can expressivists justifiably decide to put e.g., *blaming* instead of *avoiding* into the placeholder, *solely on the basis of the fact* that this would make it easier for them to reject the auxiliary assumption and thereby invalidate *obligation implies permission*? The answer I defend below is *no*. More specifically, I argue that their decision about which specific act/attitude must go into the placeholder should be criticized as problematically *ad hoc,* if it is motivated *only* by the need for securing the logical possibility of moral dilemmas.

To see why, notice first that the question of which act/attitude should go into the placeholder is a question that concerns the theory's basic meaning/attitude assignment for an atomic sentence that contains its target normative expression: what is the non-cognitive attitude that ordinary speakers express by sincerely uttering, "φ-ing is wrong", after all? Here, recall also that expressivists are committed to the mentalist assumption about the meanings of normative sentences (see Section 2): the non-cognitive attitudes assigned as the meanings of normative sentences are *mental states* of speakers who express them via their utterances. Therefore, in deciding what to put into the placeholder in *being for [some unary expression] φ-ing*, expressivists are making a *substantive empirical-psychological claim* about the mental states that underlie the use of "is wrong" in the actual linguistic practice.[22] As such, naturally, their theoretical decisions need to be empirically well-motivated. Therefore, if securing the logical possibility of moral dilemmas in their semantic theory is the *only* reason for their decision in their basic meaning assignment for "is wrong", it is problematically *ad hoc*; this is because it simply ignores other equally important, notably psychological, considerations that expressivists need to take into account in motivating their basic meaning assignment.

Let me demonstrate this point in more concrete terms. Their decision in the basic meaning assignment will, for example, yield predictions concerning what kind of behavioral patterns are generally compatible with one's sincere utterance of "φ-ing is wrong". The plausibility of their decision should be then tested by examining whether the predictions it yields fit with ordinary speakers' actual behaviors as well as their intuitions on this matter. Suppose that one's expressivist semantic theory tells us that, as Schroeder supposes, a speaker's sincere utterance of "φ-ing is wrong" expresses the attitude of *being for blaming for φ-ing.* This proposal has an implication that a speaker who sincerely utters this

[22] I thank an anonymous reviewer for urging me to make this point more explicit.

sentence must be generally disposed to sanction actual instances of φ-ing performed by others. Surely, one who possess the attitude of being for blaming for φ-ing must feel compelled to actually blame others' performances of φ-ing, at least when circumstances permit. Likewise, people should be likely to find puzzling a situation where a speaker sincerely utters "φ-ing is wrong" but never cares to blame observed instances of φ-ing at all. The question is: does this in fact capture ordinary speakers' behavior and their intuitions on this matter? If yes, putting *blaming* into the placeholder is empirically well-motivated—if no, the choice may need to be reconsidered.

Now, contrast this to an alternative proposal, which assigns the attitude of FOR(*avoiding φ-ing*) as the meaning of "φ-ing is wrong".[23] This account now yields a different prediction concerning how ordinary speakers would react to the kind of situation described above. This is because avoiding φ-ing may be, unlike blaming, a matter of making personal plans for oneself, which may not necessarily concern whether one would also publicly sanction *others'* performances of φ-ing. If this is so, my having the attitude of FOR(avoiding meat-eating), for example, might simply mean my being committed to avoid eating meat myself (and, perhaps, vaguely hope others do the same). This attitude, unlike FOR(blaming for meat-eating), need not imply that I am committed to socially sanction those who do not act as I do. Accordingly, even if I am known for overtly asserting "Meat consumption is wrong", my not taking any corrective actions toward those who continue to consume meat need not appear so puzzling on the alternative proposal. Again, the question is: does this actually fit with ordinary speakers' behavior and intuitions?

This quick comparison between the two choices above should suffice to illustrate how one's decision about which attitude/act should be put into the placeholder needs to be motivated by general empirical-psychological considerations concerning the use of "is wrong" in ordinary speakers' linguistic practice. Whatever decisions they end up making in their basic meaning assignment for "is wrong", the mentalist assumption of the expressivist semantic project implies that they are *also* making claims about individual speakers' psychology. As such, they yield various predictions that need to be tested empirically in light of the data. What I discussed above is just one example, and I suspect that there are also

[23] As I explained in the previous section, choosing to assign FOR(avoiding φ-ing), instead of FOR(blaming for φ-ing), as the meaning of "φ-ing is wrong" leads to the result that the auxiliary assumption *obligation implies permission* will appear more plausible. The current point is that this kind of choice, when taken together with the mentalist assumption, also yields other predictions that should not be ignored.

other types of similar considerations that one should take into account (e.g., to what extent do ordinary speakers take a speaker’s sincere utterance of “φ-ing is wrong” to be compatible with her emotional neutral reactions to instances of φ-ing?). It is for this reason that, in deciding which specific act/attitude should go into the placeholder in *being for [some unary expression] φ-ing*, expressivists cannot simply insist that they are justified to choose whatever act/attitude that would invalidate *obligation implies permission* and make moral dilemmas logically possible.

It might be helpful to elaborate on the current point by connecting it to the fact that when Schroeder chooses to specifically assign FOR(blaming for φ-ing) as the meaning of “φ-ing is wrong”, he purports to be following the proposal by Gibbard (Schroeder 2008a, 58). For Gibbard, “to call something rational is to express one’s acceptance of norms that permit it” (Gibbard 1990, 7); and, accordingly, “φ-ing is irrational” (which should be read as “φ-ing is wrong”, given his overall picture) would express a state of *accepting a norm that forbids φ-ing,* which looks similar to *being for blaming for φ-ing.*[24] Again, Schroeder is not necessarily committed to this specific choice, and he draws on Gibbard just to “fix examples” (58). However, surely Gibbard himself should have some basic reasons and motivations (including considerations such as above) for analyzing “is wrong” ultimately in terms of blame/forbiddance, instead of avoidance, disapproval, disliking and so on.[25] And this means that Gibbard (in his 1990 book) and others who adopt the notion of blame/forbiddance in analyzing the meaning of “is wrong” are *prima facie* committed to accepting whatever theoretical consequences that follow from “the logic of blame/forbiddance”. If it tells them that blaming for (forbidding) not φ-ing and blaming for (forbidding) φ-ing are inconsistent, they are *prima facie* committed to accepting its consequence in their semantic theory: *obligation implies permission* turns out to be formally valid, which in turn makes moral dilemmas logically impossible. My contention is that even if something like this turns out to be the case, they cannot easily switch to a different

[24] More precisely, for Gibbard, moral judgements are “judgments of what moral feelings it is rational to have”, that is, “judgements of when guilt and resentment are apt” (1990, 6). Gibbard then analyzes an act of calling something rational or irrational in terms of a speaker’s expression of acceptance (which is a non-cognitive mental state) of norms that permit or forbid the object of evaluation. So, Gibbard’s analysis of “is wrong” is expressivist in somewhat indirect way, mediated by his expressivist understanding of the evaluation of rationality. I believe, however, that the overall plausibility of the discussion does not depend on the details of Gibbard’s theory.

[25] One might take issue with this point—perhaps, Gibbard may have no deep reason to invoke the notion of forbiddance in his analysis of “is irrational“. I will address this point at the end of this section.

analysis that invokes e.g., the notion of avoidance instead of blame/forbiddance just for the purpose of blocking this result. Such a response should be criticized as problematically *ad hoc*. As I argued, securing the possibility of moral dilemmas is only one of the considerations that one should take into account in determining the basic meaning assignment for "is wrong" in one's theory. If one hastily makes changes in the basic attitude assignment to deal with this particular technical problem, it is likely to produce unintended predictions in other places and even runs the risk of unintentionally abandoning whatever basic insights that motivated one's theory in the first place.

The current point can be generalized. Different expressivist accounts invoke different basic notions in analyzing the meaning of "is wrong" (or whatever normative predicate or operator that they take to be basic, such as "is irrational", "ought", "is obligatory", etc.). To take a few examples, Blackburn (1984, 1988) analyzes "φ-ing is wrong/impermissible" in terms of *booing/disapproving φ-ing*; Gibbard (2003) analyzes "φ-ing is the thing to do" in terms of a state of *planning to φ*, Horgan and Timmons (2006) analyze "One ought to φ" in terms of *an ought-commitment that one φ's*. Each of these different proposals should be motivated by some basic theoretical considerations that they take to be important, including observations of ordinary speakers' behavior and intuitions concerning the use of the target expressions. Depending on which of these proposals one finds plausible and which act/attitude one thinks should be put in the placeholder in *being for [some unary expression] φ-ing*, different results will follow as to whether moral dilemmas are logically possible. Even if they do not like the result, modifying their basic attitude assignment just for the purpose of blocking it would be problematically *ad hoc*. I have demonstrated this point in some detail, focusing on an expressivist semantic account that invokes the notion of blaming in its basic meaning assignment. I believe that one can pose, *mutatis mutandis*, the same point for any kind of expressivist semantic theory.[26]

[26] Let me briefly demonstrate this point focusing on Blackburn's proposal as an example. Blackburn stipulates, following Ayer, that the meaning of "φ-ing is wrong/impermissible" is a speaker's disapproval of φ-ing (Blackburn 1984, 195). Suppose that one is now convinced that the notion of disapproval should be invoked in one's expressivist semantic analysis of "is wrong". Since Blackburn's original proposal faces the negation problem, one would need to reformulate Blackburn's proposal using Schroeder's framework—one obvious way to do so is to think that "φ-ing is wrong" expresses FOR(disapproving φ-ing). Here, if one thinks that there are good reasons to believe that it is not inconsistent to disapprove φ-ing and disapprove not φ-ing at the same time, one would have to accept that FOR(disapproving not φ-ing) *does not* entail FOR(not disapproving φ-ing). As a result, *obligation implies permission* turns out to be invalid in this Blackburn-inspired semantics, and, accordingly, moral dilemmas turn out to be logically possible (again, assuming that the theory does not validate other combinations of axioms that make moral dilemmas logically impossible). If expressivists should side with the pro-dilemma view, the result must be a welcoming

Before closing this section, let me address one potential objection. In the discussion above, I assumed that expressivists such as Blackburn and Gibbard have some independent theoretical reasons and motivations to stick with specific notions (such as disapproval, blame/forbiddance and so on) in their analyses of "is wrong". One might find this assumption dubious and object that their choices are not really based on any substantive, let alone empirical, considerations, because their primary aim is merely to construct *structurally adequate* expressivist semantic accounts that can deal with the Frege-Geach problem (although, if Schroeder is correct, they do not succeed in achieving this aim either). For example, Blackburn does not provide any lengthy discussion to justify his choice—he merely notes in passing that he is following Ayer (Blackburn 1984, 167). So, one might say, expressivists are free to switch to whatever attitude/act that seems suitable for dealing with technical problems at hand (such as making moral dilemmas logically possible) and there is nothing *ad hoc* about this move.

My response to the objection would be that expressivists including Blackburn, Gibbard and others *should have* supported their choices by some non-trivial empirical-psychological considerations, even if they in fact did not do so. As I explained, when taken together with the mentalist assumption of the expressivist semantic project, one's choice in the basic meaning assignment will yield various predictions that should be empirically tested, whether they like it or not. I demonstrated this point by comparing analyses that invoke different notions (such as blame, avoidance) in their basic meaning assignments. The discussion in this section, if successful, shows that expressivists cannot remain indifferent to this issue and simply maintain that whatever attitude/act will do as long as it allows them to deal with technical problems in their semantic theories. It has to be recognized that, in the expressivist semantic project, one's decision in the basic meaning assignment always comes with psychological implications.

one for those who find Blackburn's choice generally convincing. Of course, if one thinks that disapproving φ-ing and disapproving not φ-ing *are* inconsistent, then the opposite result will follow. That is exactly my point—whichever turns out to be the case, one cannot simply change the basic meaning assignment for "φ-ing is wrong/impermissible" *just because* one wants to avoid some particular result. Any such move should be criticized as problematically *ad hoc*.

4. Concluding remarks: The mentalist assumption of expressivism and its costs

In this paper, I explored how expressivists can secure the logical possibility of moral dilemmas in their semantic theories, using Schroeder's framework as a case study. Even if one's expressivist semantic theory is structurally adequate in that it can deal with the Frege-Geach problem, securing the logical possibility of moral dilemmas remains as a separate task. Specifically, in Schroeder's framework, whether or not moral dilemmas turn out to be logically possible depends on which specific attitude/act one thinks should go into the placeholder in the attitude of *being for [some unary expression] φ-ing*, assigned as the meaning of "φ-ing is wrong". Due to the mentalist assumption of the expressivist semantic project, deciding what should be put into the placeholder involves making a substantive empirical claim about the psychology that underlies the use of "is wrong" in the actual linguistic practice. Expressivists then need to take many things into consideration in making their decision, including, for example, actual behavioral patterns that typically follow a speaker's sincere utterance of "φ-ing is wrong" and folk intuitions on this matter. Accordingly, their decision should be criticized as problematically *ad hoc* if it is solely motivated by the need for invalidating *obligation implies permission* to secure the logical possibility of moral dilemmas.

Let me conclude by articulating the moral of the case study in more general terms. Overall, the case study suggests that the difficulty for expressivists mainly derives from the mentalist assumption of their semantic project that logico-semantic relations exhibited by normative sentences should be captured in terms of the psychological attitudes that speakers express by uttering them. Whenever expressivists wish to make a certain theoretical move to deal with a technical problem (e.g., invalidating *obligation implies permissibility* to make moral dilemmas logically possible), they first need to confirm that their move is consistent with the basic meaning assignment in their semantic theory. If adopting the desired theoretical move requires changing the basic meaning assignment, expressivists will need to commit to whatever empirical-psychological claims entailed by such a change. This seems to capture how the mentalist assumption generally prevents expressivists from flexibly adopting or dropping axioms in their theories to get the right technical results.

Recall also that, as I noted in Section 1, there are other combinations of intuitively plausible axioms that are known to be inconsistent with the logical possibility of moral dilemmas (e.g., *obligation implies possibility* and the principle of agglomeration).[27] Extending the strategy of the current paper, one can similarly examine whether one's preferred expressivist semantic theory can find reasonable ways to avoid such combinations. The discussion in this paper might give the impression that the prospect is indeed dim.

Of course, there is a more general question: which aspects of the data should one's semantic theory aim to respect in the end? After all, the logical possibility of moral dilemmas and *obligation implies permission* are both intuitively plausible, and ordinary speakers may often behave as if both are true. Their actual linguistic practice exhibits inconsistencies in some places, and a formal semantic theory, if it purports to be consistent, will have to ignore some aspects of the data. What needs to be given up must be decided based on many considerations, and it could turn out that, all things considered, it is better to give up the logical possibility of moral dilemmas instead of *obligation implies permission* in one's semantic theory. This is a problem for every semanticist, not just for expressivists. The moral of the paper is that expressivists need to face an *extra* constraint in addressing this kind of issue: the mentalist assumption of their semantic project prevents them from flexibly dropping or adopting axioms in their theory to deal with technical problems. This, I believe, provides an explanation of why expressivists *in particular* will have hard time addressing technical issues such as the treatment of the logical possibility of moral dilemmas. And, importantly, this point would hold even if it turns out that expressivists are not required to make moral dilemmas possible in their theories, as I assumed. Regardless of whether they ultimately need to validate or invalidate *obligation implies permission* in their theories, the crucial point is that whatever theoretical moves necessary for arriving at the desired result will need to be justifiable in light of the mentalist assumption. The case study in this paper has shown that this is a significant burden that expressivists need to bear in pursuing their semantic project.[28]

[27] I thank an anonymous reviewer for suggesting to explore the implications of the case study from this angle.

[28] Does this mean that any expressivist semantic account is bound to collapse at some point due to its mentalist assumption? If one thinks so, then the result of the current paper might be taken as providing indirect support for attempts at exploring alternative, *non*-semantic ways of cashing out basic expressivist ideas (see Bar-On and Chrisman 2009; Bar-On et al. 2014).

Acknowledgments

I am grateful to Paul Bloomfield and two anonymous reviewers for the journal for their valuable comments on earlier versions of this paper.

Funding information: This work was supported by the Japan Society for the Promotion of Science (JSPS KAKENHI 21J00685 and 22KJ0522).

REFERENCES

Ayer, Alfred J. 1936. *Language, Truth, and Logic*. Dover.

Almeida, Michael J. 1990. "Deontic Logic and the Possibility of Moral Conflict." *Erkenntnis* 33 (1): 57–71. https://doi.org/10.1007/BF00634551.

Baker, Derek, and Jack Woods. 2015. "How Expressivists Can and Should Explain Inconsistency." *Ethics* 125 (2): 391–424. https://doi.org/10.1086/678371.

Bar-On, Dorit, and Matthew Chrisman. 2009. "Ethical Neo-Expressivism." *Oxford Studies in Metaethics* 4: 133–66.

Bar-On, Dorit, Matthew Chrisman, and James Sias. 2014. "(How) Is Ethical Neo-Expressivism a Hybrid View?" In *Having It Both Ways: Hybrid Theories and Modern Metaethics*, edited by Guy Fletcher and Michael Ridge, 223–47. Oxford University Press.

Bar-On, Dorit, and James Sias. 2013. "Varieties of Expressivism." *Philosophy Compass* 8 (8): 699–713. https://doi.org/10.1111/phc3.12051.

Blackburn, Simon. 1984. *Spreading the Word: Groundings in the Philosophy of Language*. Clarendon Press.

———. 1988. "Attitudes and Contents." *Ethics* 98 (3): 501–17. https://doi.org/10.1086/292968.

———. 1998. *Ruling Passions: A Theory of Practical Reasoning*. Oxford University Press.

Brink, David O. 1994. "Moral Conflict and Its Structure." *The Philosophical Review* 103 (2): 215. https://doi.org/10.2307/2185737.

Conee, Earl. 1982. "Against Moral Dilemmas." *Philosophical Review* 91 (1): 87–97. https://doi.org/10.2307/2184670.

Geach, Peter T. 1960. "Ascriptivism." *Philosophical Review* 69 (2): 221–25. https://doi.org/10.2307/2183506.

———. 1965. "Assertion." *Philosophical Review* 74 (4): 449–65. https://doi.org/10.2307/2183123.

Gibbard, Allan. 1990. *Wise Choices, Apt Feelings: A Theory of Normative Judgment*. Harvard University Press.

———. 2003. *Thinking How to Live*. Harvard University Press.

Goble, Lou. 2009. "Normative Conflicts and The Logic of 'Ought.'" *Noûs* 43 (3): 450–89. https://doi.org/10.1111/j.1468-0068.2009.00714.x.

Hansson, Sven Ove. 2019. "In Defence of Deontic Diversity." *Journal of Logic and Computation* 29 (3): 349–67. https://doi.org/10.1093/logcom/exv057.

Hare, R. M. 1952. *The Language of Morals*. Oxford University Press.

———. 1970. "Meaning and Speech Acts." *Philosophical Review* 79 (1): 3–24. https://doi.org/10.2307/2184066.

Holbo, John. 2002. "Moral Dilemmas and the Logic of Obligation." *American Philosophical Quarterly* 39 (3): 259–74.

Horgan, Terry, and Mark Timmons. 2006. "Cognitivist Expressivism." In *Metaethics after Moore*, edited by Terry Horgan and Mark Timmons, 255-98. Oxford University Press.

Köhler, Sebastian. 2017. "The Frege-Geach Objection to Expressivism, Structurally Answered." *Journal of Ethics and Social Philosophy* 6 (2): 1–7. https://doi.org/10.26556/jesp.v6i2.150.

Lemmon, E. J. 1962. "Moral Dilemmas." *Philosophical Review* 71 (2): 139–58. https://doi.org/10.2307/2182983.

Marcus, Ruth Barcan. 1980. "Moral Dilemmas and Consistency." *The Journal of Philosophy* 77 (3): 121-36. https://doi.org/10.2307/2025665.

McConnell, Terrance C. 1978. "Moral Dilemmas and Consistency in Ethics." *Canadian Journal of Philosophy* 8 (2): 269–87.

———. 2022. "Moral Dilemmas." In *The Stanford Encyclopedia of Philosophy* (Fall 2022 Edition), edited by Edward N. Zalta. Retrieved February 22, 2024. https://plato.stanford.edu/archives/fall2022/entries/moral-dilemmas/

McNamara, Paul, and Frederik Van De Putte. 2023. "Deontic Logic." In *The Stanford Encyclopedia of Philosophy* (Fall 2023 Edition), edited by Edward N. Zalta. Retrieved February 22, 2024. https://plato.stanford.edu/archives/fall2023/entries/logic-deontic/

Nair, Shyam. 2016. "Conflicting Reasons, Unconflicting 'Ought's." *Philosophical Studies* 173 (3): 629–63. https://doi.org/10.1007/s11098-015-0511-4.

Ross, W. D. 1930. *The Right and the Good*. Oxford University Press.

Schroeder, Mark. 2008a. *Being for: Evaluating the Semantic Program of Expressivism*. Clarendon Press.

———. 2008b. "What Is the Frege-Geach Problem?" *Philosophy Compass* 3 (4): 703–20. https://doi.org/10.1111/j.1747-9991.2008.00155.x.

———. 2008c. "How Expressivists Can and Should Solve Their Problem with Negation." *Noûs* 42 (4): 573–99. https://doi.org/10.1111/j.1468-0068.2008.00693.x.

———. 2010. *Noncognitivism in Ethics*. Routledge.

Sias, James. 2024. "Ethical Expressivism." *The Internet Encyclopedia of Philosophy*, edited by James Fieser and Bradley Dowden. Retrieved March 13, 2024. https://iep.utm.edu/eth-expr/

Sinnott-Armstrong, Walter. 1988. *Moral Dilemmas*. Blackwell.

Skorupski, J. 2012. "The Frege-Geach Objection to Expressivism: Still Unanswered." *Analysis* 72 (1): 9–18. https://doi.org/10.1093/analys/anr136.

Stevenson, Charles L. 1937. "The Emotive Meaning of Ethical Terms." *Mind* 46 (181):14-31.

Styron, William. 1979. *Sophie's Choice*. Random House.

Tessman, Lisa. 2015. *Moral Failure: On the Impossible Demands of Morality*. Oxford University Press.

Unwin, Nicholas. 2001. "Norms and Negation: A Problem for Gibbard's Logic." *The Philosophical Quarterly* 51 (202): 60–75.

Williams, Bernard. 1966. "Consistency and Realism." *Proceedings of the Aristotelian Society (Supplement)* 40: 1–22.

Vallentyne, Peter. 1989. "Two Types of Moral Dilemmas." *Erkenntnis* 30 (3): 301–18. https://doi.org/10.1007/BF00168283.

Van Fraassen, Bas C. 1973. "Values and the Heart's Command." *Journal of Philosophy* 70 (1): 5–19. https://doi.org/10.2307/2024762.

Zimmerman, Michael J. 1996. *The Concept of Moral Obligation*. Cambridge University Press.

European Journal of Analytic Philosophy EuJAP | 2024 | Vol. 20 | No. 1 | 87-112
https://doi.org/10.31820/ejap.20.1.4

TWO PROBLEMS ABOUT MORAL RESPONSIBILITY IN THE CONTEXT OF ADDICTION

Federico Burdman[1]

[1] Alberto Hurtado University, Chile

Original scientific paper – Received: 07/02/2024 Accepted: 18/04/2024

ABSTRACT

Can addiction be credibly invoked as an excuse for moral harms secondary to particular decisions to use drugs? This question raises two distinct sets of issues. First, there is the question of whether addiction is the sort of consideration that could, given suitable assumptions about the details of the case, excuse or mitigate moral blameworthiness. Most discussions of addiction and moral responsibility have focused on this question, and many have argued that addiction excuses. Here I articulate what I take to be the best argument for this view, based on the substantial difficulty that people with severe addiction experience in controlling drug-related behavior. This, I argue, may in some cases be sufficient to ground a mitigating excuse, given the way in which addiction undermines agents' responsiveness to relevant moral reasons to do otherwise. Much less attention has been devoted to a second set of issues that critically affect the possibility of applying this mitigating excuse in particular cases, derived from the ambivalent nature of agential control in addiction. In order to find a fitting response to moral harm, the person with the right standing to blame must make a judgment about the extent to which the agent possessed certain morally relevant capacities at the time of the act. In practice, this will often prove tremendously difficult to assess. The ethical challenge for the person with the right standing to blame is fundamentally one of making a judgment about matters that seem underdetermined by the available evidence.

Keywords: addiction; moral responsibility; behavioral control; mitigation; degrees of blameworthiness.

Correspondence: federicoburdman@gmail.com

1. Introduction

Imagine that Diego invited his new partner, Juan, over for dinner to meet his parents. During the afternoon, Juan gets heavily intoxicated, shows up at Diego's parents' house in a bad shape and behaves in inconvenient ways. As a result, Diego is hurt and disappointed. Prior to learning further details, it seems fair to assume that Juan is blameworthy for this. Now suppose that Juan suffers from severe addiction.[1] Does this mitigate his blameworthiness?

From the perspective of the person with the right standing to blame, the question raises two quite different sets of issues. The first concerns what I will call *the principle problem*: Is the fact that Juan suffers from addiction a consideration of moral import in assessing his degree of blameworthiness? Is addiction the sort of consideration that might, under appropriate conditions, mitigate moral blameworthiness? Most discussions of addiction and moral responsibility have focused on such questions, and many have argued for the view that addiction excuses. Here I articulate what I take to be the best argument for this view. The key consideration concerns the substantial difficulty that people with severe addiction experience in controlling drug-related behavior (section 3). This, I argue, may in some cases be sufficient to ground a mitigating excuse, given the way in which addiction undermines agents' responsiveness to relevant moral reasons to do otherwise (section 4).

Much less attention has been devoted to a second set of issues that crucially affect the possibility of applying this general principle to particular cases, which I will refer to as *the practical problem* (section 5). For the general principle that addiction excuses to have any bearing on the situation at hand, the person with the right standing to blame must make a judgment about the extent to which the agent possessed certain morally relevant capacities at the time of the act. Discussions of moral responsibility in the context of addiction have for the most part neglected the practical significance of the difficulties posed by the ambivalent nature of agential control in this context. Juan might be eligible for a mitigating excuse *if* his ability to control his behavior was sufficiently

[1] I will focus here on drug addiction, but I consider the view I put forward to be relevant to other sorts of addictions as well. As for the term 'drugs', I will use it liberally to refer to any substances that may be the target of addictive behavior, thus including alcohol, nicotine, and other substances not commonly referred to as drugs outside of the addiction literature. People with addiction are the target of a great deal of stigmatizing attitudes, and in everyday discourse, to label a person in this way often carries a negative connotation about her behavior or her character and may be taken to pick out an essential trait of the person being referred to. I intend my references to people with addiction to carry none such connotations.

impaired in a way that was relevant to the nature of the moral situation he faced. In practice, however, this will prove tremendously difficult to assess, given that the ambivalent nature of agential control makes for an evidentially underdetermined situation. From the point of view of the person with the right standing to blame who has to decide on a fitting response to moral harm, there is often no fully satisfactory way to navigate the intricacies of this situation. Diego must walk a narrow path between the risk of unfairly over-blaming and the risk of condescendingly under-blaming, with no definite guide to arriving at an appropriate response.

The principle problem is the natural focus for theories of moral responsibility. But it does not fully reflect the nature of the ethical challenge faced by affected parties that seek a fair and non-condescending way to respond to addiction-related moral harms. Diminished control may mitigate moral blameworthiness, but this provides only a rough general guide for resolving questions of moral responsibility in particular cases. For the person with the right standing to blame, the challenge of deciding on a fitting response to moral harm is fundamentally about making a judgment about matters that are underdetermined by the available evidence.

2. The principle problem: some preliminaries

It seems natural, to some extent, to think of people with addiction as morally responsible agents. Even in severe cases, addictive drug use remains an intentional action in a recognizable sense of the word. It is, or appears to be, explained in terms of motivation and decision-making, and it is typically performed with a reasonably adequate level of understanding of its consequences. Thus, it seems intuitively unlike paradigmatic cases where a full exemption or a full excuse is warranted.[2]

There is, however, another way for agents to be less than fully responsible for their actions, which involves mitigation. This obtains when there are grounds for *partial* rather than full exculpation. I submit

[2] Following Strawson (1962), the standard taxonomy of the ways in which ascriptions of moral responsibility can be defeated distinguishes between exemptions and excuses. Briefly put, *exemptions* obtain when a condition undermines an agent's relevant capacities so as to render her incapable of morally responsible agency. This may occur *globally*—when the condition affects the agent's capacities across the board—or *locally*—when it undermines only certain abilities, or does so only at certain times or under certain circumstances (King and May 2018). *Excuses*, on the other hand, apply when someone who is a morally responsible agent does wrong, but special circumstances block or undermine attributions of responsibility for her behavior (see Kozuch and McKenna 2016).

that the most intuitively appealing view on the principle problem is that the way in which addiction undermines agency may, in some cases, be sufficient to mitigate moral responsibility without fully exculpating agents from addiction-related moral faults. It speaks to its *prima facie* plausibility that many scholars have defended claims in the vicinity of such a view in the past.[3] And there is also some experimental evidence that folk intuition supports the view to some extent.[4] In connection with the principle problem, my aim will be to articulate a justification for this intuition.

(One tricky issue I will leave open along the way is whether this mitigation of blameworthiness is based on a localized imperfect fulfillment of the conditions for being a morally responsible agent (i.e., a mitigating local exemption), or on a localized difficulty encountered in the way of responding to relevant moral demands (i.e., a mitigating excuse). For the sake of brevity and simplicity, I will for the most part resort to the language of excuses, but the argument I develop in later sections is consistent with both possibilities. Deciding between them would require grappling with a difficult issue, namely, whether the source of the difficulty in controlling drug-related behavior experienced by people with addiction is more plausibly located in the agent's abilities or in the circumstances in which she acts. This is an issue I will not attempt to resolve here).

On what I take to be the intuitively appealing view, when we learn that Juan suffers from severe addiction, we see him, on that account, as less blameworthy than he might otherwise have been, even if we still think he is accountable for his behavior. To illustrate, consider two variations of the case. In both variations, every circumstance and aspect of the situation is exactly the same, except that in one Juan suffers from a severe addiction, while in the other, Twin Juan does not. My contention is that the intuitive view of the case is that Juan is a fitting target for blaming responses, even though he is, on account of his addiction, less blameworthy than non-addicted Twin Juan.

[3] Related claims have been defended by Matthews and Kennett (2019), Kennett, Vincent, and Snoek (2015), Levy (2011), McConnell (2022), Pickard (2017), T. Schroeder and Arpaly (2013), Sinnott-Armstrong (2013), Wallace (1999), Yaffe (2011), Watson (1999), and Henden (2023). David Brink (2021, ch. 13) and Stephen Morse (2000) accept that in some cases addiction may provide a basis for a partial excuse, but they suggest that a successful excusing argument will often be blocked by considerations of indirect responsibility.

[4] See Racine, Sattler, and Escande (2017), Rise and Halkjelsvik (2019), Taylor et al. (2021), Vonasch, Baumeister, and Mele (2018), and Vonasch et al. (2017).

Cashing out this intuition requires producing an excusing argument. There are two basic requirements that such an argument must meet: it must be based on an empirically defensible picture of addictive agency, and it must appeal to a sufficiently plausible theory of moral responsibility. To set the stage for the argument I present in the following sections, consider two ways of arguing for the addiction excuse that fail on these grounds.

The first is built on an analogy between addiction and duress (Husak 1999; Watson 1999). In this picture, a person with addiction may be acting under a sort of internal threat of harm in the form of withdrawal symptoms. If the pains contingent upon not using are severe enough, then—the argument goes—it would be unfair to demand from an agent that she suffers such pain, and so this may provide a (partial) excuse for moral wrongdoing suitably connected with decisions to use.

The analogy between addiction and duress is imperfect on several accounts.[5] But withdrawal cannot be what we are getting at if we think that addiction *in general* excuses—although it can certainly be relevant to morally appraise the actions of people who are experiencing such symptoms. One reason is that there are types of addiction that involve only mild withdrawal symptoms, and some that involve none at all (Emmelkamp and Vedel 2006, 4). And while withdrawal symptoms can be painful and extremely hard to endure in some cases, they are usually relatively short-lived. After some time, they begin to subside and eventually cease to be experienced (Emmelkamp and Vedel 2006, 5). However, addiction continues to have the potential to undermine agency in morally relevant ways, and thus to ground an excuse, long after withdrawal symptoms have ceased to be an issue. Furthermore, the argument portrays the avoidance of withdrawal pain as the primary reason why people with addiction choose to use. This may be true in some cases, but it is surely incorrect as a general explanation of addictive drug use. People with addiction may decide to use for a number of reasons, including but not limited to the need to avoid withdrawal. Other relevant reasons to use include seeking pleasurable experiences, coping with stress or other sources of psychological discomfort, or because it coheres with established self-narratives, among many other possibilities.

Now consider another popular idea: the view of addiction as a disease. It has sometimes been suggested that the exculpatory implications of such a view are one of the reasons for endorsing it. People with addiction are

[5] For discussion, see Brink (2021, 352–354), Morse (2000, 28–38), and Yaffe (2011, 115–118).

often burdened with feelings of shame and regret, as well as the targets of third-personal resentment and anger. Viewing addiction as a disease, it is argued, can do them a service by undermining such feelings (e.g., Volkow, Koob, and McLellan 2016, 368).

There is some appeal to the idea that someone can be excused for certain behaviors on account of suffering from a disease. For example, it may be that we are under a general obligation to show compassion to people who are unfortunate or suffering, and people who have a disease can fit that description. But it seems that the main consideration when it comes to moral responsibility has to do with agential capacities, and the consideration that someone has a disease serves at best as an imperfect indicator that some of their morally relevant capacities may be affected[6]—imperfect because some diseases do not seem to affect morally relevant capacities in a significant way. Furthermore, insufficient capacity need not issue from a disease-like cause to ground an exemption or an excuse—think, for instance, of standard cases of immaturity. In response, it may be argued that calling addiction a disease implies that addictive behavior is the result of mechanistic dysfunction, and thus indirectly speaks to the impairment of morally relevant capacities (Sisti and Caplan 2016; Wakefield 1992). But the disease view of addiction can be controversial in its own right, and some have found reason to doubt that it is correct (Field et al. 2019; Heather 2013; Lewis 2017; Pickard 2022; see Burdman 2024a, for an overview of the debate). Luckily, the fate of the addiction excuse does not hang on this controversy, and we need not resolve it here. The most promising place to look when thinking about the addiction excuse is the way the condition affects morally relevant capacities, whether or not it is properly called a disease.

3. Partially impaired behavioral control

By most scientific definitions, addiction involves an element of impairment of behavioral control over drug use—the sort of thing sometimes called ‘compulsion’.[7] This is the obvious candidate for an

[6] Many have made similar points in the past. See Jefferson and Sifferd (2018), Bortolotti, Broome, and Mameli (2014), among others.

[7] Although talk of compulsion is common in psychiatric contexts, the precise meaning of the term is often unclear. Highly influential institutional sources that endorse the view of addiction as somehow impairing behavioral control include the DSM-5-TR (American Psychiatric Association 2022), the ICD-11 (World Health Organization 2019), and the definition of addiction by the NIDA in the United States (NIDA, 2014), among many others. Impaired behavioral control is usually seen as related to another key feature of addiction: continued drug use despite negative consequences. For instance, the DSM-5-TR renders the “essential feature” of ‘substance use disorders’ as “a cluster of cognitive,

impairment of ability that could ground an excuse, since it directly concerns the volitional condition for moral responsibility. Not coincidentally, many classical pieces in the moral responsibility literature cite addiction as an example of an exempting/excusing condition (e.g., Fischer and Ravizza 1998, 35; Frankfurt 1971; Watson 1975, 325).
If addictive behavior were completely or literally compelled, this would allow for an easy solution to the principle problem. However, there is forceful evidence against the view of addiction as a condition that literally renders agents unable to refrain from drug use. The main challenge in producing an answer to the principle problem is to frame the basic insight that compulsion is incompatible with responsibility in terms of an empirically defensible view of addictive agency.

I will not rehearse here the full case against the view of addictive behavior as purely compulsive (for a summary of the evidence, see Pickard 2015, 2018; Sripada 2018; Heyman 2009). For present purposes, a few basic observations will suffice. The most important relates to the fact that people suffering from addiction are generally able to regulate drug use in a way that is responsive to relevant circumstances and conditions. Given the right kind of incentive structure, even severely addicted people can choose not to use, as both experimental (Hart et al. 2000) and clinical evidence (Petry et al. 2017) suggests. Also suggestive is the fact that many people who are correctly diagnosed with addiction at some point in their lives according to extant diagnostic criteria go on to recover without medical treatment (Sobell, Ellingstad, and Sobell 2000; Heyman 2009). Indeed, a survey of expert opinion on this issue, targeting both addiction therapists and experimental researchers, found that the view of addiction as a condition that makes people simply unable to abstain from using has little support among those who work in close contact with people with addiction (Carter et al. 2014).

The implication is not that addiction does not compromise agency at all. Rather, it is that the way in which addiction compromises agency needs to be understood in a different light than as a literal inability to abstain. Refraining from use remains an open possibility, even in severe cases. Using drugs is not a reflex-like occurrence that bypasses the agent’s will; it is intentional behavior explained in terms of motivation and decision-

behavioral, and physiological symptoms indicating that the individual continues using the substance despite significant substance-related problems” (p. 546). The eleven diagnostic criteria for substance use disorder are divided into four categories: impaired control, social impairment, risky use of the substance, and pharmacological criteria. On harm as a defining feature of addiction, see Heather (1998) and Sinnott-Armstrong and Pickard (2013).

making. To frame this as a literal inability to do otherwise is simply to misdescribe the nature of addictive agency.

To be clear, there is also compelling support for the claim that addiction compromises agency in relevant ways. This is reflected in the well-known fact that addiction may be extremely difficult to overcome. For those suffering from severe addiction, quitting is far from a simple matter, and many find themselves in the difficult position of continuing to use drugs despite being aware of significant harmful consequences of doing so. An indication of how difficult it can be to refrain from using is the fact that some people suffering from severe cases of alcoholism resort to medications that cause severe sickness when alcohol is consumed, as a self-imposed penalty to discourage future consumption (Banys 1988). Even knowing that such unpleasant consequences are guaranteed, many fail to abstain from drinking.

In sum, addictive drug use is not literally compelled, but neither is it the result of purely ordinary decision-making processes. The ability to control drug use is plausibly portrayed as *partially impaired* by addiction: it is undermined to some extent, without rendering people with addiction literally incapable of doing otherwise.

Another crucial consideration about addiction is that it is in many ways a highly heterogeneous condition (Pickard 2022). There are significant differences between the patterns of use associated with different substances, as well as between the individual characteristics of people suffering from it and their life circumstances. While a reduced ability to control drug use is a common feature of all cases of addiction, the precise nature of the control-undermining factors at play appears to be variable (Burdman 2022). Potential control-undermining factors include psychological anomalies, situational pressures, and challenging social-environmental conditions, with some of these playing a more prominent role in some cases than in others.

Consider social-environmental conditions first. A social context that offers very limited opportunities to pursue alternative drug-free life trajectories may negatively affect a person’s ability to control their drug use (Hart 2013). An environment that provides support, strong incentives, and realistically available alternatives to a drug-focused lifestyle enhances a persons’ ability to refrain from using. On the contrary, attempts to quit by people struggling with unemployment or housing instability are significantly less likely to succeed (Saloner and Cook 2013).

Situational factors also play a role. The degree to which people are sensitive to considerations relevant to their actions is a variable feature of agents that can be positively or negatively influenced by immediate situational pressures. It is, for instance, much more difficult for someone with addiction to refrain from using in settings rich in drug-related cues and opportunities for use, especially in the presence of drug-using companions (J. R. Schroeder et al. 2001).

In addition to these types of agent-external conditions, the explanation of addictive behavior typically includes a variety of different psychological factors, including anomalies in motivation, cognition, and decision-making processes. Addictive desires may be anomalous in some respects, persisting in a way that is unresponsive to desire-incongruent evaluative judgments and aversive past experiences (Burdman 2024b; Holton and Berridge 2013; Wallace 1999). Drug-related cognition may also be compromised in subtle ways. Evaluative judgments about drug use may become unstable, shifting over time without the acquisition of new evidence (Levy 2014), and drug-related belief formation may be biased toward use-congruent interpretations (Pickard 2016; Segal 2013), or otherwise distorted (Sripada 2022). In addition, addiction significantly affects the allocation of attention. This occurs both at the perceptual level, where drug-related perceptually available items tend to capture attention through bottom-up influences, and in the context of deliberation, where use-congruent considerations are more likely to remain within attentional focus (Cox, Klinger, and Fadardi 2016). In some cases, decision-making processes may be more generally skewed toward the pursuit of rewards that can be obtained sooner, leading to difficulties in appropriately weighing the value of rewards that are more distant in time (Ainslie 2000; Bickel et al. 2014; Bechara 2005).

The interpretation of available evidence is open to dispute and scientific knowledge is always subject to revision. For now, however, the tentative picture that emerges from the current state of knowledge is roughly as follows: people with addiction experience powerful motivation to use, they may have difficulty bearing in mind and appropriately weighing considerations that speak against drug use, and their attentional and belief formation processes may be tilted towards use-congruent outcomes. In some cases, these traits interact problematically with situational and social-environmental factors that contribute to undermining control. Crucially, all of these features are matters of degree. In important respects, addictive behavior remains voluntary and intentional; it is not necessitated. It is no accident, however, that many agents who fit the above description continue to use systematically, find it so hard to quit while they are users, and are so likely to relapse while in recovery.

4. From diminished ability to reduced blameworthiness

The idea that compulsive behavior precludes blameworthiness is treated as a data point by classical theories of moral responsibility. Incompatibilists of various stripes argue that any causal determination undermines moral responsibility, while compatibilists typically rely on intuitions about how compulsive behavior differs from ordinary cases of deterministic causation to argue for the conclusion that it is the former, not the latter, that is incompatible with responsibility. One thing on which all parties to the classical debate seem to agree is that addiction is a prime candidate for a condition that makes agents unfitting targets of responsibility demands. However, they usually do so by assuming that compulsion means that the agent is literally unable to do otherwise. Once we think of impaired control as a matter of degree, things look a little different.

For the purposes of this discussion, I will adopt what I consider to be the theory of moral responsibility best suited to graded distinctions, namely, a capacitarian account.[8] On this view, the basic requirement for morally responsible agency is the possession of certain morally relevant capacities (Fischer and Ravizza 1998; Vargas 2013; Brink 2021; Nelkin 2011; Sartorio 2016; Wallace 1994; Mckenna 2013). A useful way of unpacking this proposal is this: for an agent to be aptly held morally responsible for her actions, she must behave in a way that reflects a sufficient capacity to respond to relevant moral reasons pertaining to the situation at hand.[9] If an agent does not possess this capacity to a sufficient degree, she is not a fitting target of moral demands.

[8] I am inclined to think that the basic thrust of my argument could also be recast in the context of a Deep Self or a Quality of Will approach to moral responsibility. I cannot adequately defend this suggestion here, but the underlying idea is simple enough. Diminished control over drug use is relevant to an assessment of the extent to which an agent's behavior is a non-deviant expression of her deep evaluative commitments and cares. Similarly, it is a relevant consideration for a Quality of Will view, since partial impairment of behavioral control affects the extent to which morally wrongful behavior can be seen as expressing ill will toward wronged parties. I do not mean to suggest that these theories are extensionally equivalent, but I think they could all find a place for the intuition I am trying to articulate here.

[9] This is roughly put. In Fischer and Ravizza's formulation, the relevant condition is that the agent acts on a mechanism that is her own and that is moderately reasons-responsive, i.e., that it is regularly receptive and weakly reactive to moral reasons. Crucially, degrees of reasons-responsiveness are measured in terms of the set of possible worlds in which the agent successfully responds to potential or counterfactual sufficient moral reasons for doing otherwise (1998, ch. 3). In other accounts, reasons-responsiveness is pictured as a property of agents rather than of subpersonal mechanisms (e.g., Brink and Nelkin, 2013; McKenna 2013; Vargas 2013). Receptivity and reactivity are often seen as distinct components of normative competence, the former referring to the ability to detect the presence of relevant moral considerations and the latter to the ability to suitably govern one's behavior in light of such sensitivity. Cases of addiction typically involve some degree of impairment in both types of abilities.

Reasons-responsiveness is not a have-it-or-don't property of agents, but a scalar property falling along a continuous spectrum. Following Fischer and Ravizza, theorists of reasons-responsiveness typically think of responsibility as a threshold concept, meaning that there are minimum conditions that an agent must meet in order to be within the domain of morally responsible agency at all—there is some point along this gradient that determines the minimum degree of reasons-responsiveness that makes an agent an apt target of moral demands. Nonhuman animals and small infants are often cited as examples of agents that do not meet such minimum conditions. Their behavior is flexible enough to be modified by environmental circumstances, but it is not sufficiently responsive to the presence of moral reasons for them to be aptly held responsible when those reasons are overlooked.[10]

A crucial consideration is that the ability of morally responsible agents to track relevant reasons and to successfully respond to them comes in many shades, varying from person to person and within the same agent at different times or under different circumstances. The scalar nature of reasons-responsiveness implies that there will still be significant differences between agents who meet the relevant minimum requirements, i.e., those within the domain of morally responsible agency. This picture of degrees of reasons-responsiveness thus sits well with the intuition that moral responsibility is not an all-or-nothing affair. If we take reasons-responsiveness to be the agential capacity that grounds fitting ascriptions of moral responsibility, then it stands to reason that *partial ability* will lead to *partial responsibility*, provided that the minimum threshold conditions for morally responsible agency are met.

A natural development of the theory is then to think of degrees of moral responsibility as more or less directly tracking degrees of reasons-responsiveness (Coates and Swenson 2013; Nelkin 2016). This sort of approach can make sense of some intuitive cases. For instance, we tend to think of older children and adolescents as having an ambivalent standing when it comes to moral responsibility, with some demands on them seeming appropriate while others do not. The extent to which ascriptions of moral responsibility are appropriate seems to be plausibly captured by the extent to which we see maturing agents as having the capacity to suitably respond to the relevant moral reasons.

[10] I will side with the majority view here and speak of moral responsibility as a threshold concept. But this is not too important for the issue at hand, and the argument I outline is also consistent with the possibility of thinking of moral responsibility as fully scalar all the way down.

Now, consider again Juan's case. Diego reasonably expected him to show up in good shape when meeting his parents for the first time. Thus, there were reasons that, in those particular circumstances, spoke against the decision to use drugs at that time, as it was incompatible with the commitment he had made. The fact that Juan suffers from severe addiction is relevant for the assessment of his responsibility in failing to refrain, as it speaks to a partially undermined ability to respond to relevant moral reasons when decisions to use drugs are at issue. The fact that Juan suffers from severe addiction makes it much more difficult for him to respond to the presence of the relevant moral reasons, insofar as he experiences a substantial difficulty refraining from drug use. His degree of reasons-responsiveness seems sufficient for him to be aptly held responsible, i.e., he meets the minimum threshold conditions for moral responsibility. However, the fact that he enjoys the relevant ability to a lesser degree than non-addicted Twin Juan makes it the case that, all else being equal, he is less blameworthy for his behavior than Twin Juan.[11]

5. The practical problem

If the foregoing argument is correct, addiction excuses to the extent that it undermines agents' ability to respond to relevant moral reasons pertaining to the situation at hand. Thus, assessing the extent to which someone suffering from addiction is responsible for her behavior in a particular case involves making a judgment about the extent to which the agent enjoyed the relevant capacities. But making such judgment with any confidence will often prove to be an extremely difficult task. This is at the heart of the practical problem.

Of course, this problem is not unique to the addiction excuse. Moral theory is often concerned with general principles whose applicability in particular circumstances depends on further judgments about the nature of the case, including both matters of fact and normative appraisals.[12] But the problem takes a particularly dire form when it comes to addiction.[13] If

[11] For related arguments, see R. Jay Wallace (1999, 652-654) and Walter Sinnott-Armstrong (2013, 137-139). My concern here is with variables relevant to claims about *direct* responsibility in the context of addiction. Considerations of *indirect* responsibility are, of course, potentially relevant in this context, but I lack the space to adequately discuss them here.

[12] See Kelly (2018, 86-99) for an insightful discussion (not related to addiction) of some key situational variables that are critical to making fine-grained moral judgments.

[13] Some of the issues I discuss in this section probably arise with regard to other sorts of mental health illnesses as well (see Dings and Glas 2020). For present purposes, however, I will restrict the scope of the discussion to cases of addiction.

something approximating the argument laid out in the previous sections is correct, the addiction excuse is fundamentally grounded in the fact that people with addiction often lack full control over certain behaviors. And yet, behavioral control in the context of addiction is something of an elusive notion. On the most plausible view of addiction, control may be significantly reduced but is typically not eliminated—and the force of the addiction excuse depends on the correct assessment of the extent to which the actions in question were under the agent's control. This poses a significant challenge to blamers, who must make a particularly difficult call concerning the extent to which addiction has undermined the agent's control over the relevant actions. In practice, this is often difficult to determine given the available evidence. The most pressing ethical challenge for the person with the right standing to blame is how to navigate the epistemic precariousness of this situation.

Moreover, there are risks associated with getting the judgment wrong. Over-blaming is, of course, problematic. It is clearly unfair to blame someone more than is warranted by the extent of their actual responsibility. But under-blaming can also be problematic in its own way. On the one hand, there are instrumental reasons related to the function of blame in this context. As people with addiction struggle to gain a firmer grip on their agency, holding them accountable for their behavior can be a valuable way of supporting this effort by providing them with the right sort of feedback. In addition, there are other risks associated with under-blaming that are distinctively moral in nature. Withholding blame when blame is appropriate may convey the message that we see the other as less capable of moral agency. Thus, it amounts to denying an important form of recognition of the other's status as a moral agent: under-blaming risks sending the message that one does not see the other as a full member of one's moral community (Shoemaker 2022). I do not mean to suggest that a proper consideration of this issue should lead to the conclusion that the problems associated with over-blaming and under-blaming are symmetrical. It may be, for instance, that the harms that would result from over-blaming are somehow more serious. The important observation, in the present context, is that the risks associated with under-blaming are not insignificant and can be a subject of serious moral concern.[14]

[14] Insofar as one thinks of the harms of over-blaming as more serious than those of under-blaming, one might wonder, as two anonymous reviewers suggested, whether it follows that erring on the side of under-blaming is the preferable option given the fragility of our epistemic position with respect to blaming accurately. Blaming less may be a wise policy in these cases, though one should be aware that the solution is not optimal since, as noted in the main text, there are likely to be costs to erring on the side of under-blaming as well. In any case, my present aim is not to argue for a particular view on

For people with the right standing to blame—especially those in close relationships with people suffering from addiction—it is crucial to get this judgment right. And yet it is immensely difficult to do so.

5.1 How much control did the agent have?

Based on what we currently know about addiction, it is fair to say that the condition can, in some cases, significantly undermine agents' ability to refrain from using. From the point of view of the blamer, however, what needs to be determined is the extent to which a person with addiction was in control of some relevant action at the time of acting. In a sense, this involves the ordinary difficulty of making a judgment on a matter of fact based on incomplete evidence, compounded by the fact that there are ethical consequences to getting this judgment wrong. But when it comes to addiction, there are additional complications that make this assessment more difficult for the blamer.

One is that the very concept of partial or undermined control is unfamiliar and particularly difficult to grasp. Folk-psychological lore is not well equipped to deal with the ambivalent status of addictive agency when it comes to behavioral control. Complete lack of control is much easier to grasp. We seem to have no trouble picturing that there is a purely causal explanation for the sleepwalker's wandering or the seizure's victim erratic movements. These are not up to the agent in any relevant sense, and they seem to have nothing to do with what she has reason to do or her preferences, and so it is doubtful, at best, that these happenings belong in the realm of action. But it is much harder to grasp that a person can do something intentionally, at least in part because she wants to, and yet that her actions are not fully under her control. Moreover, commonsensical proxies for addictive motivation risk promoting a false sense of understanding. The predicament of the person with addiction who is trying to refrain is not, despite common metaphors, like the common difficulty of abstaining from eating too many chips or too much ice cream. There is something extraordinary about the difficulty that people with addiction face. This unordinary difficulty in refraining is, to put it bluntly, the main reason for thinking of addiction as a mental disorder.[15]

how someone with the right standing to blame should actually respond in the face of addiction-related moral harm, but to draw attention to the epistemic and normative challenges involved in assessing what the appropriate response might be, and in particular to how such challenges arise from some of the peculiar features of addictive agency.

[15] It is true that we are not unfamiliar with the idea that someone may be less than fully responsible for an action because they are in a particularly difficult situation that provides a partial or total excuse. We tend to cut people some slack when they are particularly stressed, suffering from difficult personal circumstances, or experiencing great pain or discomfort. On a natural reading, such

As suggested above, the philosophical toolkit can help with the thorny issue of how to make sense of the very notion of degrees of control. Thinking of degrees of control as degrees of reasons-responsiveness offers a way to capture both sides of the coin. On the one hand, it seems true that there are always sufficient reasons (actual or counterfactual) to refrain from using that even people with severe addiction would respond to. Thus, their inclination to use is not totally unresponsive to relevant considerations—they have some control. On the other hand, the set of actual or potential sufficient reasons to refrain to which they would successfully respond is plausibly smaller than the corresponding set for a non-addicted person under otherwise similar circumstances. Thus, there are some actual or counterfactual scenarios in which they have sufficient reason to refrain and yet fail to do so. In other words, they have less control than the non-addicted person, all else being equal.

And yet this will not get us very far when it comes to making the sort of judgment that is relevant to deciding particular cases. Did the person, at the moment of action, have sufficient capacity to respond to the moral reasons for doing otherwise that were overlooked from the point of view of the blamer? The graded nature of control in the context of addiction makes it particularly difficult to answer this question with any confidence. Whatever evidence we consider for the case, it will predictably be consistent with both the presence and the absence of the relevant sort of control. This kind of underdetermination follows once we accept that the person has the ability to respond to some relevant reasons, though possibly not to all of them. Ambivalent control implies that we should expect some evidence of preserved control as well as some evidence of diminished control.

A further difficulty is that the ability to control behavior is not a fixed property of agents, but one that varies across contexts and circumstances. The evidence suggests that control in addiction is highly sensitive to relevant contextual features. People with addiction seem to find it much more difficult to abstain when they are with certain people or in certain places. For example, the tendency to experience drug craving is known to be highly context sensitive (Skinner and Aubin 2010). Thus, it is not only

considerations concern situational factors rather than the more basic sort of normative competence that seems to be at stake when we focus on behavioral control (for the distinction between competence and situational factors, see Brink and Nelkin 2013). The idea that someone had, at the moment of acting, a diminished or impaired ability to control their own behavior is more unusual and difficult to grasp than the idea that someone is suffering from stress, in part because the latter, but not the former, is an experience to which virtually everyone can relate. This difficulty is not unique to addiction, though, as impaired behavioral control is plausibly involved in other psychiatric conditions as well. Thanks to an anonymous reviewer for pressing me on this point.

the intrinsic general ability to refrain that needs to be taken into account, but the specific ability that the person had in the particular circumstances under consideration. This is even more difficult to estimate.

A clinical assessment of the severity of the person's addiction may be helpful, but it is at best an imperfect proxy for the kind of assessment that is relevant to moral responsibility. Diagnostic criteria, imperfect as they are, are developed with a specific goal in mind, namely, to identify those cases in which clinical intervention might be beneficial. Thus, if the relevant goal is to appraise degrees of responsibility, there is a real possibility that the criteria that are useful to clinicians will turn out to be an imperfect guide.

The DSM-5-TR distinguishes between *mild*, *moderate*, and *severe* forms of substance use disorder (American Psychiatric Association 2022, 546). In practice, the distinction is operationalized in terms of the number of diagnostic criteria met by the patient, as judged by the clinician. The "severe" level is applied to cases in which the patient meets six or more of the eleven diagnostic criteria listed in the manual. This is intended to capture an observation that amounts to clinical common sense: within the domain of cases sensibly described as 'addiction,' some are more severe than others. But the quantity of symptoms is at best an imperfect measure of the intuitive notion of severity. The criteria that are useful for the purposes of diagnosis are not all equally relevant to responsibility judgments. For example, whether the person is using more than intended or is failing to fulfill social roles with which she identifies, seems prima facie more relevant to responsibility than whether she is showing signs of tolerance to the drug. Furthermore, the quantity of symptoms approach to assessing severity is intended to capture a graded notion, but it does so by adding up how many of the criteria are met, and each of the criteria is decided by a categorical assessment. The very fact that the manual proposes to measure severity in this fashion is a testament to the difficulty of assessing levels of ability. The availability of a diagnosis from a competent clinician can provide guidance in making the sort of judgment that is relevant to moral responsibility, but it will often not be enough to settle the issue.

5.2 How much reasons-responsiveness did the situation call for?

Estimating with confidence how much control the agent had at the moment of action is not the only challenge for the blamer. Another particularly tricky issue arises when we consider how much control *should* have been sufficient to avoid wrongdoing in the situation at hand. Or, to put it differently, just how compromised control must be in order for an agent to qualify for mitigation of responsibility under the relevant circumstances.

This too is, in a way, a general difficulty we encounter when we think of responsibility itself, and of the agential abilities that make someone responsible for her actions, as matters of degree. Moral reasons are not created equal: some are more salient than others. We assume, for instance, that it requires less moral understanding to see that it is wrong to murder someone than it does to realize that a particular joke might be offensive to someone with a different cultural background, even if we consider both to be morally required. And the same goes for reactivity. We expect some actions to be so aversive to a morally competent agent that it would take less self-control to refrain from doing them. Briefly put, avoiding certain morally criticizable behaviors requires less in the way of moral competence. Just how much moral understanding and what degree of ability to govern oneself are required to avoid engaging in morally criticizable activities are not fixed parameters, and vary along different moral situations. Thus, even if we assume that an agent possesses the abilities relevant to moral responsibility to an imperfect degree, it may be the case that she is sufficiently capable to warrant the normative expectation that she does not behave in certain ways when the situation at hand is less demanding of moral competence. Fully spelling out the rationale for this would involve resolving some difficult issues in the theory of moral responsibility. But the addiction excuse seems to be more forceful in some cases than in others, depending on how high the moral stakes are. The more the consequences of a decision to use drugs involve serious moral harms, the more the intuitive appeal of the addiction excuse appears to weaken.[16]

Of course, judging the degree of moral competence that a particular situation calls for is a difficulty that people with the right standing to blame face in many sorts of cases that do not involve addiction. This, too, is a difficulty we are bound to face once we think of responsibility and moral competence as matters of degree. What matters in the present context is that this further difficulty adds to the challenge faced by potential blamers in addiction cases. Assuming that the person had a diminished or imperfect ability to control her behavior in a given

[16] An addiction clinician recounted to me, in private conversation, a heartbreaking story about a former patient of his who, at one point, had become so desperate for money to buy drugs that she had forced her underage child to have sex with a stranger in exchange for money. Even assuming that her ability to respond to relevant moral reasons was compromised, as it surely was, it seems hard to envision how it is that whatever ability she retained was not sufficient to allow her to recognize that such behavior was utterly unacceptable.

situation, *should* that degree of control have been sufficient to prevent her from overlooking the relevant moral reasons for doing otherwise than she did?[17]

Think of Juan and Diego again. Diego rightly expected Juan to be kind and respectful to his parents who had invited him to their home, and Juan failed to respond to that expectation. When he decided to start drinking prior to their rendezvous, the prospect of letting Diego and his parents down did not exert sufficient pull on his deliberation to make him exercise his ability to refrain. However, if he had learned that there was a fire in the building he was in when he was about to pour his glass, he would probably have chosen to run from the fire instead of having a drink. And if he had believed that having that drink would, through some intricate causal chain, lead to Diego's death, he would likely not have done it either. He had some control that he could have exercised if the situation had been dire enough. The prospect of letting Diego down and hurting his feelings, which he likely contemplated that afternoon, gave him a less salient and less compelling reason to refrain than the possibility of precipitating his death would have given him. *Should* that have been enough?

Note that *post hoc* expressions of remorse, apologies, and other attempts at relationship repair are relevant to the moral situation, albeit in a different way. Juan may use these means to demonstrate that he cares, that he recognizes the legitimacy of Diego's expectations, and that he values their relationship. Despite the relevance of all this to their ongoing moral conversation, it speaks to a different kind of concern. *Backward-looking* responsibility is particularly grounded in the amount of agential control the person had at the time of action. Thus, whether Juan cares about past harms now does not substitute the need to assess whether he should have been able to respond to the relevant moral reasons at the time of action.

As before, there is no general solution to this difficulty. A person may be fully convinced that the imperfect ability to control behavior that we see in addiction can, in principle, provide a mitigating excuse for overlooking

[17] Duress cases can present a similar conundrum. Suppose that the rationale for a duress excuse is that it would be unfair to demand from someone that she confronted a credible threat, or that she suffered the threatened harms, if a person of reasonable firmness would not be expected to do so (Watson 1999). The extent to which such a principle provides an excuse in particular cases arguably depends, among other things, on the nature of the consequences that would follow from giving in. A threat to someone's life might, given suitable assumptions, excuse that person from driving the getaway car in a robbery, or from failing to alert the police. But it seems intuitively insufficient to excuse a person for, say, participating in mass murder.

certain moral reasons in some situations. Notwithstanding this, the person with the right standing to blame will still be faced with the need to navigate the intricacies of ambivalent agential control in order to resolve what her response should be. Of course, how to respond to addiction-related moral harm is a complex question that depends on factors other than whether backward-looking blame is appropriate. For example, Diego could, perhaps should, consider which responsibility response would be more useful for them and for the relationship going forward. However, to the extent that the question is whether blame is deserved, as opposed to whether it would serve some forward-looking purpose, the search for an answer will lead one to ponder the difficulties arising from ambivalent behavioral control. This puts the potential blamer in a particularly difficult position: how much control the agent had, or how compromised her ability to respond to relevant moral reasons pertaining to the situation at hand was, is underdetermined by the available evidence. And whether the degree of control the agent had should have sufficed to respond to the relevant moral reasons calls for another challenging normative appraisal that is difficult to make with any confidence.

6. Conclusion

Can addiction be the basis for a mitigation of responsibility for addiction-related moral faults? In some cases, I have argued, it can. The reason is that addiction may partially impair the ability to control drug-related behavior. As a result, an addicted person’s responsiveness to moral reasons may be diminished when decisions to use drugs are at issue. On a plausible theory of moral responsibility, such a decrease in reasons-responsiveness affords not a full exemption or excuse, but a mitigation of responsibility for moral faults that are suitably connected with decisions to use drugs.

However, this offers only limited guidance when it comes to assessing degrees of blameworthiness in particular cases. Applying the addiction excuse involves further factual and normative appraisals that are particularly difficult to make. Moreover, getting these assessments right is often important to avoid the risks involved in both over- and under-blaming. For people with the right standing to blame, the need to navigate the intricacies of the ambivalent agential control that we see in addiction poses a significant ethical challenge.

Acknowledgments

I thank Migdalia Arcila-Valenzuela, Gerald Lang, Diego Lawler, Diana Pérez, and Tobías Schleider for their insightful comments on earlier versions of this paper. Thanks also to two anonymous reviewers for this journal. Research for this paper began while I was a Visiting Fellow at the School of Philosophy, Religion and History of Science, University of Leeds, and was completed while I was a Research Resident at the Brocher Foundation in Geneva (www.brocher.ch).

REFERENCES

Ainslie, George. 2000. "A Research-Based Theory of Addictive Motivation." *Law & Philosophy* 19: 77–115. https://doi.org/10.1023/A:1006349204560.

American Psychiatric Association. 2022. *Diagnostic and Statistical Manual of Mental Disorders: DSM-5-TR*. American Psychiatric Association.

Banys, Peter. 1988. "The Clinical Use of Disulfiram (Antabuse®): A Review." *Journal of Psychoactive Drugs* 20 (3): 243–61. https://doi.org/10.1080/02791072.1988.10472495.

Bechara, Antoine. 2005. "Decision Making, Impulse Control and Loss of Willpower to Resist Drugs: A Neurocognitive Perspective." *Nature Neuroscience* 8 (11): 1458–63. https://doi.org/10.1038/nn1584.

Bickel, Warren K., Mikhail N. Koffarnus, Lara Moody, and A. George Wilson. 2014. "The Behavioral- and Neuro-Economic Process of Temporal Discounting: A Candidate Behavioral Marker of Addiction." *Neuropharmacology* 76 (PART B): 518–27. https://doi.org/10.1016/j.neuropharm.2013.06.013.

Bortolotti, Lisa, Matthew R. Broome, and Matteo Mameli. 2014. "Delusions and Responsibility for Action: Insights from the Breivik Case." *Neuroethics* 7 (3): 377–82. https://doi.org/10.1007/s12152-013-9198-4.

Burdman, Federico. 2022. "A Pluralistic Account of Degrees of Control in Addiction." *Philosophical Studies* 179 (1): 197–221. https://doi.org/10.1007/s11098-021-01656-7.

———. 2024a. "Is Addiction a Disease?" *Análisis Filosófico*. Forthcoming.

———. 2024b. "Recalcitrant Desires in Addiction." In *Oxford Studies in Agency and Responsibility, Vol. 8*, edited by Santiago Amaya, David Shoemaker, and Manuel Vargas. Oxford University Press. https://doi.org/10.1093/oso/9780198910114.003.0004

Brink, David. 2021. *Fair Opportunity and Responsibility*. Oxford: Oxford University Press.

Carter, Adrian, Rebecca Mathews, Stephanie Bell, Jayne Lucke, and Wayne Hall. 2014. "Control and Responsibility in Addicted Individuals: What Do Addiction Neuroscientists and Clinicians Think?" *Neuroethics* 7 (2): 205–14. https://doi.org/10.1007/s12152-013-9196-6.

Coates, D. Justin, and Philip Swenson. 2013. "Reasons-Responsiveness and Degrees of Responsibility." *Philosophical Studies* 165 (2): 629–45. https://doi.org/10.1007/s11098-012-9969-5.

Cox, W. Miles, Eric Klinger, and Javad S. Fadardi. 2016. "Nonconscious Motivational Influences on Cognitive Processes in Addictive Behaviors." In *Addiction and Choice*, edited by Nick Heather and Gabriel Segal, 259–85. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198727224.003.0015.

Dings, Roy, and Gerrit Glas. 2020. "Self-Management in Psychiatry as Reducing Self-Illness Ambiguity." *Philosophy, Psychiatry, & Psychology* 27 (4): 333–47. https://doi.org/10.1353/ppp.2020.0043.

Emmelkamp, Paul M. G., and Ellen Vedel. 2006. *Evidence-Based Treatment for Alcohol and Drug Abuse. A Practitioner's Guide to Theory, Methods, and Practice*. New York: Routledge.

Field, Matt, Nick Heather, and Reinout W. Wiers. 2019. "Indeed, Not Really a Brain Disorder: Implications for Reductionist Accounts of Addiction." *Behavioral and Brain Sciences* 42 (March): e9. https://doi.org/10.1017/S0140525X18001024.

Fischer, John M., and Mark Ravizza. 1998. *Responsibility and Control. A Theory of Moral Responsibility*. Cambridge: Cambridge University Press.

Frankfurt, Harry. 1971. "Freedom of the Will and the Concept of a Person." *The Journal of Philosophy* 68 (1): 5–20.

Hart, Carl. 2013. *High Price*. Harper Perennial.

Hart, Carl, M. Haney, R. W. Foltin, and M. W. Fischman. 2000. "Alternative Reinforcers Differentially Modify Cocaine Self-Administration by Humans." *Behavioural Pharmacology* 11 (1): 87–91. https://doi.org/10.1097/00008877-200002000-00010.

Heather, Nick. 1998. "A Conceptual Framework for Explaining Drug Addiction." *Journal of Psychopharmacology* 12 (1): 3–7. https://doi.org/10.1177/026988119801200101.

———. 2013. "Is Alcohol Addiction Usefully Called a Disease?" *Philosophy, Psychiatry, & Psychology* 20 (4): 321–24. https://doi.org/10.1353/ppp.2013.0050.

Henden, Edmund. 2023. "Addiction and Autonomy: Why Emotional Dysregulation in Addiction Impairs Autonomy and Why It Matters." *Frontiers in Psychology* 14 (February). https://doi.org/10.3389/fpsyg.2023.1081810.

Heyman, Gene M. 2009. *Addiction: A Disorder of Choice*. Cambridge, Mass.: Harvard University Press.

Holton, Richard, and K. Berridge. 2013. "Addiction Between Compulsion and Choice." In *Addiction and Self-Control*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199862580.003.0012.

Husak, Douglas N. 1999. "Addiction and Criminal Liability." *Law and Philosophy* 18 (6): 655. https://doi.org/10.2307/3505096.

Jefferson, Anneli, and Katrina Sifferd. 2018. "Are Psychopaths Legally Insane?" *European Journal of Analytic Philosophy* 14 (1): 79–96. https://doi.org/10.31820/ejap.14.1.5.

Kennett, Jeanette, Nicole A. Vincent, and Anke Snoek. 2015. "Drug Addiction and Criminal Responsibility." In *Handbook of Neuroethics*, edited by J. Clausen and N. Levy, 1065–83. Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-94-007-4707-4_71.

King, Matt, and Joshua May. 2018. "Moral Responsibility and Mental Illness: A Call for Nuance." *Neuroethics* 11 (1): 11–22. https://doi.org/10.1007/s12152-017-9345-4.

Kozuch, Benjamin, and Michael McKenna. 2016. "Free Will, Moral Responsibility, and Mental Illness." In *Philosophy and Psychiatry. Problems, Intersections, and New Perspectives*, edited by Daniel D. Moseley and Gary Gala. New York: Routledge.

Levy, Neil. 2011. "Addiction, Responsibility, and Ego Depletion." In *Addiction and Responsibility*, edited by Jeffrey Poland and George Graham, 89–112. The MIT Press. https://doi.org/10.7551/mitpress/9780262015509.003.0004.

———. 2014. "Addiction as a Disorder of Belief." *Biology and Philosophy* 29 (3): 337–55. https://doi.org/10.1007/s10539-014-9434-2.

Lewis, Marc. 2017. "Addiction and the Brain: Development, Not Disease." *Neuroethics* 10 (1): 7–18. https://doi.org/10.1007/s12152-016-9293-4.

Matthews, Steve, and Jeanette Kennett. 2019. "Diminished Autonomy: Consent and Chronic Addiction." In *Beyond Autonomy. Limits and Alternatives to Informed Consent in Research Ethics and Law*, edited by David G. Kirchhoffer and Bernadette J. Richards, 48–62. Cambridge University Press. https://doi.org/10.1017/9781108649247.004.

McConnell, Doug. 2022. “Moral Responsibility in the Context of Addiction.” In *The Oxford Handbook of Moral Responsibility*, edited by Dana Nelkin and Derk Pereboom. Oxford University Press.

Mckenna, Michael. 2013. “Reasons-Responsiveness, Agents, and Mechanisms.” In *Oxford Studies in Agency and Responsibility*, edited by David Shoemaker, 1:151–84. Oxford: Oxford University Press.

Morse, Stephen. 2000. “Hooked on Hype: Addiction and Responsibility.” *Law and Philosophy* 19: 3–49.

National Institute on Drug Abuse. 2014. *Drugs, Brains, and Behavior. The Science of Addiction*. www.humanconnectomeproject.org.

Nelkin, Dana Kay. 2011. *Making Sense of Freedom and Responsibility*. Oxford: Oxford University Press.

Petry, Nancy M., Sheila M. Alessi, Todd A. Olmstead, Carla J. Rash, and Kristyn Zajac. 2017. “Contingency Management Treatment for Substance Use Disorders: How Far Has It Come, and Where Does It Need to Go?” *Psychology of Addictive Behaviors* 31 (8): 897–906. https://doi.org/10.1037/adb0000287.

Pickard, Hanna. 2015. “Psychopathology and the Ability to Do Otherwise.” *Philosophy and Phenomenological Research* 90 (1): 135–63. https://doi.org/10.1111/phpr.12025.

———. 2016. “Denial in Addiction.” *Mind and Language* 31 (3): 277–99. https://doi.org/10.1111/mila.12106.

———. 2017. “Responsibility without Blame for Addiction.” *Neuroethics* 10 (1): 169–80. https://doi.org/10.1007/s12152-016-9295-2.

———. 2018. “The Puzzle of Addiction.” In *The Routledge Handbook of Philosophy and Science of Addiction*, edited by H. Pickard and S. Ahmed, 9–22. New York: Routledge.

———. 2022. “Is Addiction a Brain Disease? A Plea for Agnosticism and Heterogeneity.” *Psychopharmacology* 239 (4): 993–1007. https://doi.org/10.1007/s00213-021-06013-4.

Racine, Eric, Sebastian Sattler, and Alice Escande. 2017. “Free Will and the Brain Disease Model of Addiction: The Not So Seductive Allure of Neuroscience and Its Modest Impact on the Attribution of Free Will to People with an Addiction.” *Frontiers in Psychology* 8 (November). https://doi.org/10.3389/fpsyg.2017.01850.

Rise, Jostein, and Torleif Halkjelsvik. 2019. “Conceptualizations of Addiction and Moral Responsibility.” *Frontiers in Psychology* 10 (June): 1483. https://doi.org/10.3389/fpsyg.2019.01483.

Saloner, Brendan, and Benjamin Lê Cook. 2013. “Blacks and Hispanics Are Less Likely Than Whites to Complete Addiction Treatment, Largely Due to Socioeconomic Factors." *Health Affairs* 32 (1): 135–45. https://doi.org/10.1377/hlthaff.2011.0983.

Sartorio, Carolina. 2016. *Causation and Free Will*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198746799.001.0001.

Schroeder, Jennifer R., Carl A. Latkin, Donald R. Hoover, Aaron D. Curry, Amy R. Knowlton, and David D. Celentano. 2001. "Illicit Drug Use in One's Social Network and in One's Neighborhood Predicts Individual Heroin and Cocaine Use." *Annals of Epidemiology* 11 (6): 389–94. https://doi.org/10.1016/S1047-2797(01)00225-3.

Schroeder, Timothy, and Nomy Arpaly. 2013. "Addiction and Blameworthiness." In *Addiction and Self-Control,* edited by Neil Levy, 214–38. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199862580.003.0011.

Segal, Gabriel M. A. 2013. "Alcoholism, Disease, and Insanity." *Philosophy, Psychiatry, & Psychology* 20 (4): 297–315. https://doi.org/10.1353/ppp.2013.0059.

Sinnott-Armstrong, Walter. 2013. "Are Addicts Responsible?" In *Addiction and Self-Control,* edited by Neil Levy, 122–43. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199862580.003.0007.

Sinnott-Armstrong, Walter, and Hanna Pickard. 2013. "What Is Addiction?" In *The Oxford Handbook of Philosophy and Psychiatry,* edited by K.W.M. Fulford, Martin Davies, Richard G. T. Gipps, George Graham, John Z. Sadler, Giovanni Stanghellini, and Tim Thornton. Vol. 1. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199579563.013.0050.

Sisti, Dominic, and Arthur Caplan. 2016. "The Concept of Disease." In *The Routledge Companion to Philosophy of Medicine*, edited by Miriam Solomon, Jeremy R. Simon, and Harold Kincaid, 5–15. New York: Routledge.

Skinner, Marilyn D., and Henri Jean Aubin. 2010. "Craving's Place in Addiction Theory: Contributions of the Major Models." *Neuroscience and Biobehavioral Reviews* 34 (4): 606–23. https://doi.org/10.1016/j.neubiorev.2009.11.024.

Sobell, Linda C., Timothy P. Ellingstad, and Mark B. Sobell. 2000. "Natural Recovery from Alcohol and Drug Problems: Methodological Review of the Research with Suggestions for Future Directions." *Addiction* 95 (5): 749–64. https://doi.org/10.1046/j.1360-0443.2000.95574911.x.

Sripada, Chandra. 2018. "Addiction and Fallibility." *The Journal of Philosophy* 115 (11): 569–87. https://doi.org/10.5840/jphil20181151133.

———. 2022. "Impaired Control in Addiction Involves Cognitive Distortions and Unreliable Self-Control, Not Compulsive Desires and Overwhelmed Self-Control." *Behavioural Brain Research* 418 (February): 113639. https://doi.org/10.1016/j.bbr.2021.113639.

Strawson, Peter. 1962. "Freedom and Resentment." *Proceedings of the British Academy* 48 (January): 1–25. https://doi.org/10.1073/pnas.48.1.1.

Taylor, Matthew, Heather M. Maranges, Susan K. Chen, and Andrew J. Vonasch. 2021. "Direct and Indirect Freedom in Addiction: Folk Free Will and Blame Judgments Are Sensitive to the Choice History of Drug Users." *Consciousness and Cognition* 94 (September): 103170. https://doi.org/10.1016/j.concog.2021.103170.

Vargas, Manuel. 2013. *Building Better Beings. A Theory of Moral Responsibility*. Oxford: Oxford University Press.

Volkow, Nora D., George F. Koob, and A. Thomas McLellan. 2016. "Neurobiologic Advances from the Brain Disease Model of Addiction." *New England Journal of Medicine* 374 (4): 363–71. https://doi.org/10.1056/nejmra1511480.

Vonasch, Andrew J., Roy F. Baumeister, and Alfred R. Mele. 2018. "Ordinary People Think Free Will Is a Lack of Constraint, Not the Presence of a Soul." *Consciousness and Cognition* 60 (April): 133–51. https://doi.org/10.1016/j.concog.2018.03.002.

Vonasch, Andrew J., Cory J. Clark, Stephan Lau, Kathleen D. Vohs, and Roy F. Baumeister. 2017. "Ordinary People Associate Addiction with Loss of Free Will." *Addictive Behaviors Reports* 5 (June): 56–66. https://doi.org/10.1016/j.abrep.2017.01.002.

Wakefield, Jerome C. 1992. "The Concept of Mental Disorder: On the Boundary between Biological Facts and Social Values." *American Psychologist* 47 (3): 373–88. https://doi.org/10.1037/0003-066X.47.3.373.

Wallace, R. Jay. 1994. *Responsibility and the Moral Sentiments*. Harvard University Press.

———. 1999. "Addiction as a Defect of the Will." *Law and Philosophy* 18 (6): 621–54.

Watson, Gary. 1975. "Free Agency." *The Journal of Philosophy* 72 (8): 205–20. https://doi.org/10.2307/2024703.

———. 1999. "Excusing Addiction." *Law & Philosophy* 18: 589–619.

World Health Organization. 2019. *International Classification of Diseases, Eleventh Revision (ICD-11)*. https://icd.who.int/browse11.

Yaffe, Gideon. 2011. “Lowering the Bar for Addicts.” In *Addiction and Responsibility*, edited by Jeffrey Poland and George Graham, 113–38. Cambridge, Mass.: MIT Press.

European Journal of Analytic Philosophy EuJAP | 2024 | Vol. 20 | No. 1 | 113-131
https://doi.org/10.31820/ejap.20.1.5

NONTRIVIAL EXISTENCE IN TRANSPARENT INTENSIONAL LOGIC

Miloš Kosterec[1]

[1] Institute of Philosophy, Slovak Academy of Sciences, Slovakia

Original scientific paper – Received: 19/11/2023 Accepted: 18/04/2024

ABSTRACT

The paper analyses the validity of arguments supporting the assumption of a constant universe of individuals over all possible worlds within Transparent Intensional Logic. These arguments, proposed by Tichý, enjoy widespread acceptance among researchers working within the system. However, upon closer examination, this paper demonstrates several weaknesses in the argumentation, suggesting that there is an open possibility to incorporate a variable universe of individuals even in models within this system.

Keywords: individual; existence; non-trivial property; existence test.

Correspondence: milos.kosterec@gmail.com

1. Introduction

The constant universe of individuals of discourse is a fundamental concept within Transparent Intensional logic (TIL). [1] From the perspective of those well-versed in TIL, assigning existence to an individual holds little to no value, as every individual possesses it in a trivial manner. This foundational assumption governs the manner in which certain data is explained within the framework of TIL. Notably, existence, when considered as a value with *informative* content, is not ascribed to individuals but rather to what is known as 'offices' (positions that, at most, one can occupy at a given moment, such as the president of the USA) or to properties. An office is said to exist, for instance, when there is currently an occupant in that position—e.g., the president of the USA exists.[2]

There is a shared stance in TIL that when we ascribe existence to an individual, we claim something trivial, as it is presumed to be a property an individual cannot lack. However, ascribing existence to an office or a property can instantiate a non-trivial claim. When we ascribe existence to the office (e.g., 'The current president of the USA exists.'), we claim that there is currently an individual occupying that office. This claim should not be confused with *another* claim stating that the office itself exists (e.g., 'The office of the president of the USA exists.'). Offices can be vacant, but that does not mean they are non-existent. A vacant office does not venture into obscurity or 'non-existence'. The widely discussed examples of 'the king of France' or 'the first female president of the USA' are completely graspable offices belonging to the ramified hierarchy of objects over the standard base in TIL.[3] This line of inquiry then extends into areas of intensional logic or philosophy of fiction.[4]

[1] The reader interested in TIL should consult e.g., Tichý (1988), Duží, Jespersen, and Materna (2010), and Raclavský (2020).

[2] This needs some clarification. The offices do exist even if they are vacant. This statement is considered trivial and non-informative, however. These functions are members of the ontology of the universe of objects defined within TIL. For example, the office 'the first female president of the USA' does exist. Technically, according to TIL, it is a function from possible worlds into chronologies of individuals, which does not have a value (i.e., the occupant) in the present world and time. So, even if an office is currently vacant, it does not lose its existence—it is a function we can still talk about. This statement should not be confused with the statement about there being an occupant of the office. This is a non-trivial statement that can differ in its truth value over time.

[3] This, of course, depends on some presuppositions being satisfied (e.g., USA, France being somehow part of the ramified hierarchy too). That is usually handled by the correct selection of the base of the ramified hierarchy.

[4] For discussions about intensional logics, see e.g., Jespersen (2015) and Duží (2017), for philosophy of fiction, consult Glavaničová (2018) and Duží, Jespersen, and Glavaničová (2021).

The assumption of triviality of existence, when considered as a property of individuals, is not merely a baseless postulation within TIL. Instead, it forms a pivotal point in argumentation consistently endorsed by TIL proponents. It is not just an assumption; rather, it is a defended one, supported by a line of reasoning. The core objective of this paper is to meticulously analyse these arguments and expose their shortcomings. It may well be the case that modelling existence as a property of offices, properties, and concepts is more fruitful. However, Tichý's line of argumentation was aimed at supporting the idea that existence, when used as a property of individuals, is only trivial. My argumentation in this paper directly addresses this point.

Tichý argued from a logical perspective rather than a pragmatic or methodological one. Therefore, even if the choice to model existence non-trivially as a property of offices, properties and concepts seems more promising, that is not the line of Tichý's argumentation. He appears to posit that stances assuming existence as a non-trivial individual property are based on conceptual confusion, and my argumentation challenges this view.

The paper is structured as follows: The second section provides a standard presentation of TIL with its foundational definitions. The third section outlines TIL's position within the debate about existence and non-existent objects. The fourth section presents the arguments for the constant universe of individuals over possible worlds, as stated in TIL. The fifth, core section of the paper, delves into an in-depth investigation of these arguments to highlight associated problems. The paper concludes with a discussion of relevance of the provided results.

2. TIL in brief

TIL is a system of explications designed to elucidate (natural) language phenomena, developed over a framework of abstract entities referred to as 'constructions' and their associated properties.[5] This system equips us

[5] Usually, the term 'procedure' is deemed more general than the notion of construction, which, although also having non-formal interpretations, is the one originally defined within the formal definitions of TIL—there is a definition of *construction, constructing according to a valuation,* there are collection of *constructions of order n* within the ramified type hierarchy. I am aware of the use of the term procedure as well. Several of the primary TIL based texts published within the last decade or so contain the notion of construction and it is still used quite often. What is, perhaps, the most important thing, is that the formal definitions of the two terms are identical. The term construction is (or at least was) widely used, for example, in Duží, Jespersen, and Materna (2010), Duží and

with the necessary tools to precisely distinguish the meanings of linguistic terms, particularly when they are considered within hyperintensional contexts. The objects used for explication within this system are defined inductively and are rooted in a foundational level known as the ‘base’. While TIL allows for any finite set of disjoint non-empty sets to be considered as a base, it typically assumes a base consisting of sets of individuals, truth values, possible worlds, and real numbers, the latter being employed for modelling time and numbers. In TIL, individual properties are represented as characteristic functions defined on individuals.[6] Individual offices are modelled as partial functions mapping worlds to time chronologies of individuals (if any), and propositions are represented as partial functions from possible worlds into the chronologies of truth values (if any). This foundational structure serves as a basis for expanding the system with constructions, which provide a model for hyperintensions. These constructions introduce new types of objects posited in interesting logical relationships with classical entities such as individuals and classical intensional entities, e.g., individual properties.

Before focusing on the model of existence, let’s present the relevant foundational definitions of TIL.[7] Tichý’s canonical version of TIL presents the notion of valuation first:

> Thus, where R^1, R^2, R^3, R^4, ... is an enumeration (without repetition) of all the types, a valuation is an array of the form
>
> (*v*)
> $$\begin{array}{l} X^1{}_1, X^{\mathit{1}}{}_{\mathit{2}}, X^1{}_3, X^1{}_4, \ldots \\ X^2{}_1, X^2{}_2, X^2{}_3, X^2{}_4, \ldots \\ X^3{}_1, X^3{}_2, X^3{}_3, X^3{}_4, \ldots \\ X^4{}_1, X^4{}_2, X^4{}_3, X^4{}_4, \ldots \\ \ldots \end{array}$$
>
> where $X^{\mathrm{i}}{}_1$, $X^{\mathrm{i}}{}_2$, $X^{\mathrm{i}}{}_3$, $X^{\mathrm{i}}{}_4$, … is an R^{i}-sequence. (Tichý 1988, 61)

I agree this could appear alien to the standard notions of valuation known from classical logics. For TIL, valuations are infinite arrays of countably infinite sequences of objects. These arrays contain exactly one such

Jespersen (2015), Duží (2019), and Kosterec (2020). The term procedure is primarily used in, e.g., Jespersen (2019) and Jespersen and Duží (2022).

[6] Which is equivalent to modelling them as sets of individuals.

[7] Those who are already familiar with the foundational definitions can safely skip this section.

sequence for each type. Consequently, variables are assigned an object with respect to such a valuation according to their position index and type index (this is usually not presented into technical details, but it is understood in TIL).

Let's continue with the main notion within TIL:
Def. *construction*

(i) *Variables* x, y, … are *constructions* that construct objects (elements of their respective ranges), dependent on a valuation v; they v-construct.

(ii) Where X is any object whatsoever (even a construction), ${}^0X$ is the *construction Trivialization* that constructs X without any change.

(iii) Let X, $Y_1,\ldots,Y_n$ be arbitrary constructions. Then *Composition* $[X\ Y_1\ldots Y_n]$ is the following *construction*. For any v, the Composition $[X\ Y_1\ldots Y_n]$ is *v-improper* if at least one of the constructions X, $Y_1,\ldots,Y_n$ is v-improper or if X does not v-construct a function that is defined at the n-tuple of objects v-constructed by $Y_1,\ldots,Y_n$. If X does v-construct such a function, then $[X\ Y_1\ldots Y_n]$ v-constructs the value of this function at the n-tuple.

(iv) (λ-)*Closure* $[\lambda x_1\ldots x_m\ Y]$ is the following *construction*. Let x_1, x_2, …, x_m be pairwise distinct variables and Y a construction. Then $[\lambda x_1\ldots x_m\ Y]$ *v-constructs* the function f that takes any members B_1, …, B_m of the respective ranges of the variables x_1, …, x_m into the object (if any) that is $v(B_1/x_1,\ldots,B_m/x_m)$-constructed by Y, where $v(B_1/x_1,\ldots,B_m/x_m)$ is like v, except that it assigns B_1 to x_1, …, B_m to x_m.

(v) Where X is any object whatsoever, ${}^1X$ is the *construction Execution* that v-constructs what X v-constructs. Thus, if X is a v-improper construction or not a construction at all, ${}^1X$ is v-improper.

(vi) Where X is any object whatsoever, ${}^2X$ is the *construction Double Execution*. If X is not itself a construction, or if X does not v-construct a construction, or if X v-constructs a v-improper construction, then ${}^2X$ is v-improper. Otherwise, ${}^2X$ v-constructs what is v-constructed by the construction v-constructed by X.

(vii) Nothing is a *construction* unless it so follows from (i) through (vi).

Examples:

- ${}^0+$, ${}^0Paul$, $[\ {}^0+\ {}^01\ x]$ are constructions

The notion of a *construction* is a fundamental concept defined within TIL. The notion tends to be informally explicated using connotations with procedures. The most important aspect is that a construction is different from its results, and many different constructions can lead to the same result. This is grasped by construction's ability to construct an object (if any) with respect to a valuation.[8]

The ontology of TIL, providing models for natural language phenomena, includes constructions as well as non-constructional objects. Constructions are defined with the assumption of other kinds of objects.[9] One, therefore, needs to be careful to avoid potential vicious circles. TIL employs a type system for this matter. This type system is inductive and assumes there is a foundation: *base*. Here is the precise formulation:

Definition 2 (*ramified hierarchy of types*). Let B be a *base,* where a base is a collection of pair-wise disjoint, non-empty sets. Then:

T_1 *(types of order 1)*

- i) Every member of B is an elementary *type of order 1 over B.*
- ii) Let α, β_1,..., β_m ($m > 0$) be types of order 1 over B. Then the collection (α β_1 ... β_m) of all m-ary partial mappings from β_1 ,..., β_m into α is a functional *type of* order 1 *over B.*
- iii) Nothing is a *type of order 1 over B* unless it so follows from (i) and (ii).

C_n *(constructions of order n)*

- i) Let x be a variable ranging over a type of order n. Then x is a *construction of order n over B.*
- ii) Let X be a member of a type of order n. Then ${}^0X$, ${}^1X$, ${}^2X$ are *constructions of order n over B.*
- iii) Let $X, X_1,..., X_m$ ($m > 0$) be constructions of order n over B. Then $[X\ X_1... X_m]$ is a *construction of order n over B.*
- iv) Let $x_1,..., x_m$, X ($m > 0$) be constructions of order n over B. Then $[\lambda x_1...x_m\ X]$ is a *construction of order n over B.*
- v) Nothing is a *construction of order n over B* unless it so follows from C_n (i)-(iv).

T_{n+1} (*types of order n* + 1)

Let $*_n$ be the collection of all constructions of order n over B. Then

- i) $*_n$ and every type of order n are *types of order n* + 1.

[8] The iv) point of the definition of construction is where the construction Closure explicitly *v*-constructs a function. Closures always *v*-construct functions. However, Closures are not *the only* constructions that can *v*-construct a function. A variable can *v*-construct a function, a Trivialization can, etc. Therefore, there is no conceptual dependence of the point iii) on the point iv) of the definition.

[9] As the main topic of the paper is the notion of the existence of individuals, I was a bit concerned about its usage when considering other kinds of objects. To be precise, we do not just assume objects—we assume their existence, at least as far as the system is concerned.

ii) If $m > 0$ and α, $\beta_1, \ldots, \beta_m$ are types of order $n + 1$ over B, then $(\alpha\ \beta_1 \ldots \beta_m)$ (see T_1 ii) is a *type of order* $n + 1$ *over B*.

iii) Nothing is a *type of order n + 1 over B* unless it so follows from (i) and (ii).

The standard epistemic base assumed for the wide majority of models provided in TIL is as follows:

ο: the set of truth-values {T, F};
ι: the set of individuals (the universe of discourse);
τ: the set of real numbers (doubling as times);[10]
ω: the set of logically possible worlds (the logical space).

For any type τ, a set of objects of type τ is usually modelled by its characteristic function, which is assigned (οτ) as its type. Standard intensional entities (individual properties, offices, …) are modelled as follows: if α is a type, then ((ατ)ω) is an intension (abbreviated as $\alpha_{\tau\omega}$)—a function from possible worlds ω to chronologies of objects of a particular type (ατ). Propositions—as intensions into the truth values—are assigned a type $ο_{\tau\omega}$.

The specification of the standard epistemic base within TIL includes the basic type of individuals.[11] Technically, when TIL provides models over the standard epistemic base, it does not analyse or explicate the members of the base over that base. "The elements of the members of *B*[ase] serve as arguments for intensions, and cannot be analysed within TIL without incurring circularity" (Duží, Jespersen, and Materna 2010, 59). The

[10] TIL does allow for infinite domains. Moreover, TIL does not prescribe cardinality on basic types in general. As Tichý stated: "Any domain of initially given objects can serve as a base of infinite hierarchy of *types* of entity, (…)" (1988, 65). The ramified hierarchy of typed objects within TIL is built upon a *base* that is a collection of non-empty and pairwise disjoint collections. The standard epistemic base of TIL contains at least one uncountable basic type—real numbers doubling for times and numbers. The cardinality of individuals is usually not discussed, although nothing seems to be blocking it from being uncountable, too. Nevertheless, TIL does not require that the language has a constant for every object in the domain. The inductive definition of the ramified type hierarchy does not assume this. There is also a particular caution when modelling relations of an agent to intensions or hyperintensions in TIL not to expect, prescribe, or presuppose any grasp on the actual infinite. This is not a problem, however, as investigations into the properties of objects in and defined over uncountable domains are usually done using, at most, a countable language. One needs to be careful when specifying the semantic models of such a language.

[11] The notion of 'base' is a technical term from the definition of the ramified type hierarchy within TIL—a *base* is a collection of pairwise disjoint, non-empty sets. As such, a base is whatever fulfils this condition. An epistemic base, as the term is standardly used in the TIL literature, is a base *accompanied with an explication of the members of the base*—so it is not just a set of collections of objects ι, ο, ω, τ, but these are explicated as sets of individuals, truth-values, possible worlds and real numbers. The term "epistemic" emphasizes the added explication of what the members of the set stand for.

standard model of a sense of a proper name in TIL is a Trivialization of an individual.[12] Duží et al. further characterise the conditions on the use of proper names by competent language users: "(…) the understanding a sense of a name is what enables a language-user to intellectually identify or select the bearer of a name" (Duží, Jespersen, and Materna 2010, 285). Of course, the other standard means for identifying an individual is by the use of a determiner. Following Duží et al., we can present the competence to identify and discern among the individuals within the domain as a condition for linguistic competence (of a speaker), which we model.[13]

3. TIL and (non)-existence

In this paper, I focus on how the property of existence is represented within the system. Within this context, existence, as a property of individuals, is modelled as a trivial property—specifically, a property inherent to all individuals for all possible worlds, at all times. Essentially, every individual is attributed with existence; that is the extent of it. Consequently, it becomes implausible to assert the non-existence of an individual based on its representation in TIL. This characteristic of TIL prompts us to reevaluate the well-recognized challenges associated with negative existential claims—statements such as 'The king of France does not exist.' or 'Sherlock Holmes does not exist.'. These challenges are presented as not being fundamentally about the existence of a particular individual but rather pertaining to the status of some office and the occupancy state thereof. From a technical standpoint, the existence of individuals is captured by a constant function, which assigns the truth value *True* to every individual. This representation conforms to the standard model of individual existence within TIL.[14]

Popular stances within philosophical logics and analytic philosophy have been devised to discuss and analyse arguments containing non-existent

[12] This was discussed in some depth e.g., in Duží, Jespersen, and Materna (2010, ch. 3.2).

[13] Although TIL includes the term 'logic' in its name, it is not typically regarded as a logic according to the conventional understanding of the term. Classical logic typically involves a definition of language, interpretation, and models. In contrast, Tichý and his followers present their framework and provide semantic models within it. TIL is better understood as a theory of abstract objects and their relations. However, there has been a recent trend in presenting TIL in a format resembling standard presentations of formal theories, see Raclavský (2020). I decided not to present TIL in this form in the paper, as the focus is more on the philosophical motivations behind certain decisions within the system rather than its formal properties.

[14] There has been some discussion about how to model some kinds of non-trivial existence, but then it was understood rather as properties like *having a mass, being positioned in space and time,* etc. See, e.g., Raclavský (2010).

individuals. Various approaches exist for handling this problem. Let's mention a few. Meinong and his followers present a position according to which "there is indeed an object for every mental state whatsoever—if not an existent object, then at least a nonexistent one" (Reicher 2022, sec. 2). Another popular line of investigations is based on employing the notion of impossible worlds (see, e.g., Berto 2008), and at least some of these presumably contain impossible, non-existent individuals.[15] These lines of investigations can be considered a 'bottom-up' approach, as they present an enrichment of the domain in one way or another.

TIL adopts a robust 'top-down' approach when considering the notion of existence. Apparent examples of non-existent individuals are usually analysed as (hidden) individual offices (e.g., *Pegasus* is really the winged horse, Vulcan is really the particular planet in an orbit between Mercury and the Sun, etc.). This way, TIL does not need to posit a particular metaphysics involving non-existent particulars. It elucidates the semantics of language contexts seemingly dependent on such concepts by utilising notions already available in its conceptual framework (office, hyper-office, etc.).[16] This approach enables TIL to circumvent the need to posit problematic features such as two primitive kinds of predication or a distinction between nuclear and extranuclear properties.[17]

4. Why existence *has to be* a trivial property

This section is dedicated to presenting the argumentation in favour of the proposed model of individual existence as a trivial property within TIL. The primary argumentation line, articulated by Tichý, encapsulates the core of this perspective. Below, I highlight the essential elements of this

[15] Philosophical analysis of fictional contexts presents another wide domain of stances dealing with the apparent existence of fictional characters—non-existent objects par excellence, see e.g. Zalta (2003).

[16] See, e.g., Duží, Jespersen, and Glavaničová (2021).

[17] TIL enables us to distinguish between a predication *de dicto* and a predication *de re*. This distinction can be nicely seen in the analysis of the meanings of sentences concerning predication to the offices, in contrast to the sentences with predication to the occupants (if any) of the offices. We predicate de dicto when we assign a property (of offices) to the office itself, e.g., 'The president of the USA is an elected office.'. We predicate de re when we assign a property (of individuals) to the occupant of the office, e.g., 'The president of the USA is a white male.' From a technical standpoint, both predications are grasped by the use of Composition, which presents an application of a function to an argument. The difference between de dicto and de re predication is then modelled by different functions applied to different objects within these models—there is usually an extensionalization process when we predicate de re. This poses no issues since it does not introduce new primitive notions; it only enables us to grasp the semantic difference using notions already in place. On the other hand, Zalta's theory assumes the introduction of two distinct primitive types of predication: *exemplifying* and *encoding*.

argumentation along with some additional commentary.[18] Tichý's central argument aims to refute the notion of a fluctuating universe of individuals, specifically countering the idea that the same set of individuals does not belong to each world:

> Indeed the most widespread view of possible worlds is to the effect that although worlds do share objects, they do so on a selective basis: the universe of discourse, it is assumed, may expand and/or contract from world to world.
> (…)
> An individual which is present in the actual world may, on this view, be missing from some alternative worlds, and conversely, an individual which is to be found in some alternative worlds may be missing from the actual world.
> (…)
> This view is popular, but not easy to defend. (Tichý 1988, 180)

Tichý proceeds to challenge the conception of 'possibilia', referring to objects that do not exist in the actual world but are posited to exist in some other world.[19] He urges the proponents of this position to provide precise specifications for such objects, contending that such precision is unattainable.[20] Tichý articulates his argument concisely. He challenges anyone to provide at least one example of an individual missing from the actual world. He begins with the standard example of Pegasus. Which particular individual is Pegasus? The standard reply—it is the winged horse—fails to designate an individual in the actual world and presumably does not uniquely specify a *sole* candidate in a world in which it exists. However, for the sake of the argument, let's consider a particular world that contains the unique winged horse. How can one be certain that the winged horse there is not one of the wingless horses in the

[18] This argumentation is accepted by virtually all researchers working in TIL.

[19] Tichý (1988, ch. 36) utilised the distinction between actual and alternative/possible worlds, at least when presenting his arguments against the idea of varying domains of individuals. He did not, however, base these arguments concerning the existence of individuals on any particular logic. Instead, he focused on presenting the limitations of certain positions, following some basic assumptions. I do not presume any particular logic behind my counterargument either. Although I am well aware that Tichý does provide an explication of his notion of worlds by introducing the notion of 'determination system' (see Tichý 1988, 197ff). Nevertheless, he presents his arguments against the varying universe of individuals without the reference to this explication, which only followed several pages after this particular line of argumentation. TIL is developed with a strong 'anti-actualist' stance (see Duží, Jespersen, and Materna 2010, ch. 2.4.1). I do not presume, however, that either Tichý's argumentation or my counter-argumentation depends on a particular notion of an actual world. Both can be reformulated without the need to use this particular term. I acknowledge that the specification of the actual world would amount to omniscience.

[20] It is presented in a concise way in Tichý (1988, 179ff).

actual world—possessing wings is presumably a contingent property. So, something more is needed. Specifically, a claim that the unique winged horse in the considered world is numerically distinct from any individual in this world. But presumably, there is more than one non-existent individual (if we do not want to beg the question). According to Tichý "[t]o be able to exploit the determiner in pinpointing such an individual, one has to have an epistemic handle on the individual's numerical identity in the first place" (1988, 181). I concede that his reasoning up to this point is sound.

Now, let's focus on the point that Tichý does not stop here—i.e., he is not satisfied by dismantling a position about the possibility of an individual not existing in the actual world but existing in some other possible world. He goes on and, in his words, *per impossibile*, grants that "we have managed to focus on a specific non-existent individual" (Tichý 1988, 182). He continues:

> (...) what evidence could we possibly have that it indeed fails to exist? If existence is something that an individual may have or lack, then the question whether it lacks it is a factual one and cannot be answered *a priori*. Just as one cannot be sure that an individual fails to be golden without subjecting it to a goldeness test, so (on the view under consideration) one cannot be sure that it fails to exist without subjecting it to an *existence test*. Yet the idea of testing an individual for existence is grotesquely absurd. (Tichý 1988, 182)

This is a famous argument of the test, respected and repeated on many occasions in TIL literature. However, claims of absurdity can be seen as suspicious, as what is absurd for one can be the basis for a career for another.

5. Devil is in the details

Having presented the arguments against the position advocating varying domains of individuals in TIL, let's now delve deeper into the intricacies of these arguments. This section aims to shed light on certain problematic aspects within the argumentation.

Tichý initially agrees that the concept of a fluctuating universe of individuals suggests that some individuals, not existing in this world, do exist in some other possible worlds, and *conversely*, some individuals existing in this world do not exist in some other worlds. However, the initial part of his argument primarily addresses just one aspect of this possibility. Specifically, he argues against the feasibility of specifying an individual that doesn't exist in the actual world but exists in some other possible world—he argues against possibilia. It is crucial to note that this argument does not inherently challenge the alternative possibility: an individual existing in the actual world but not existing in some other possible world. This aspect is *not* directly addressed in the initial part of Tichý's argument (i.e., in his argumentation against possibilia).

It's worth noting that Tichý, seemingly recognizing the potential limitations of his initial argument, proceeded to present another, ostensibly more robust, argument against the concept of non-trivial existence considered as a property of individuals. This subsequent argument, if valid, would effectively eliminate the possibility of such a property within the TIL framework.

However, it is essential to emphasize that, even in the case of this second argument, there remain questions regarding its validity.[21] In the following, I outline my reasons for asserting the second argument's potential shortcomings and invite a critical examination of its claims.

Here's my reasons. Let's, once again, present the argument of the test in full:

> If existence is something that an individual may have or lack, then the question whether it has or lacks it is a factual one and cannot be answered *a priori*. Just as one cannot be sure that an individual fails to be golden without subjecting it to a goldenness test, so (on the view under consideration) one cannot be sure that it fails to exist without subjecting it to an *existence test*. Yet the idea of testing an individual for existence is grotesquely absurd. If the individual does not exist, it is simply not available for testing; and if it *is* available then it is entirely futile to proceed with the test, because it is clear already that it exists. An existence test for individuals, whatever it might consist in, would have to be one which cannot possibly yield a negative result. (Tichý 1988, 182)

It is my contention that Tichý, in his argument, takes a logical step that lacks sufficient substantiation. Specifically, he makes a critical move from the assumption that existence is a property an individual may have

[21] Even if it is respected by virtually all within the TIL community.

or lack to an intermediary conclusion that it is a factual property, hence rendering it unanswerable *a priori*. This logical step is crucial for Tichý's subsequent argumentation, wherein he posits the absurdity of empirically testing such a property. However, Tichý did not adequately support this logical transition. Tichý appears to consider two following concepts as co-extensional: *non-trivial individual property* and *empirically testable individual property*.[22] He relies on the assumption that for us to claim that an individual possesses a non-trivial property, it must necessarily undergo a factual testing. However, this is not a universally applicable principle. There exist non-trivial individual properties that can be assigned to an individual without the requirement of empirical testing. I do not contend that numerous trivial properties do not warrant empirical testing, but I posit that not all non-trivial individual properties follow this pattern. Essentially, even if an individual property, assignable to an individual only after factual testing and hence, modelled as a non-trivial property within the system, exists, it does not automatically imply that any non-trivial individual property must be empirically testable. In simpler terms, while we may model empirical properties using non-trivial ones, this does not establish a one-to-one correspondence (or a subsumption), wherein every non-trivial property must be empirical in nature.

Let's demonstrate this. Let's assume that we have several possible worlds in our domain—say w_1, w_2, etc. By having these in the domain we can mention them explicitly in the linguistic statements (the same way we do with individuals). Now, let's specify this property: 'being identical to oneself and being such that the world is w_1'. Although it could sound strange, it is along with the properties like 'being such that it's raining' or 'being such that one plus one equals two'. [23]

Employing the notion of construction as well as the definition of the ramified type hierarchy, we can specify the Closure, which *v*-constructs such an individual property. [24] Let's present the standard type assignment:[25] $w/{*}_1 \rightarrow \omega$, $x/{*}_1 \rightarrow \iota$, w_1/ω, $\&/(ooo)$, $=_\iota /(o\iota\iota)$, $=_\omega / (o\omega\omega)$, then

$$\lambda w[\ \lambda x\ [^0\&\ [^0{=_\iota}\ x\ x]\ [^0{=_\omega}\ w\ {^0w_1}]]]$$

[22] Even this may be too strong a claim. Tichý's argument assumes that any non-trivial individual property is an empirical one. He does not need the converse to be true.

[23] Similar kinds of individual properties (individual is such that …) were discussed within modern debate over connections of essentialism and modality. See, e.g., Fine (1994) and Wildman (2016).

[24] I do not use a temporal index for simplicity in the paper.

[25] The type assignments to the parts of the semantic model presented do not deviate in any way from the usual type assignments. The example does not rely on any deviation in typing.

v-constructs an individual property, which all individuals possess in world w_1 and no individual possesses in any other possible world.[26] This is an example of a construction of an individual property that is non-trivial, but we do not need an empirical testing to acknowledge it is so.[27] This property is possessed by all individuals *only* in the world w_1.[28] No individual possesses this property in any other world. We know this *a priori*, without testing. And it is an example of a non-trivial individual property. This is therefore an example demonstrating that Tichý's argument of test is based on an unsubstantiated assumption about the subsumption of extension of the notion of non-trivial property under the extension of the notion of an empirical property.

The proponents of TIL do not assign any priority to the actual world. So, it is much in line with the suggestion that the actuality is only a contingent property of a possible world. Consider that the world w_b happens to be actual (or that w_b is actual from the viewpoint of w_b). Then Tichý's argumentation does not block the possibility of there being another possible world that has even fewer individuals than those that occupy w_b. I acknowledge that the world w_a is probably not graspable from the viewpoint of w_b—as, by assumption, w_a is occupied by more individuals than w_b. However, this is not a concern, as the epistemic and conceptual possibilities, *as far as the individuals within that world are concerned*, can and do vary across possible worlds. Tichý's argumentation was against conceivability of the exact specification of a particular individual not existing in the actual world (whichever world being actual). My counter-argumentation does not face this challenge—from any world that happens to be the actual, we *can* consider worlds that comprise even less individuals than that world—the problem of specification does not appear in that scenario.

Section 1.4.2.1 in Duží, Jespersen, and Materna (2010) provides a detailed analysis of various kinds of individual properties. What is important for the purposes of the paper and for the specification of non-trivial properties is the class of trivial properties, *Triv*/$(o(o\iota)_{\tau\omega})$, as defined: "To sum up, a property *P* belongs to the class *Triv* iff *P* has a

[26] In more detail, this is a Closure, which *v*-constructs a function from possible worlds into objects *v*-constructed by $\lambda x[\ {}^0\& \ [{}^0{=}_\iota \ x \ x] \ [{}^0{=}_\omega \ w \ {}^0w_1]]$. This second Closure *v*-constructs a function from individuals into truth values *v*-constructed by $[\ {}^0\& \ [{}^0{=}_\iota \ x \ x] \ [{}^0{=}_\omega \ w \ {}^0w_1]]$—which depicts a conjunction of the statement 'individual is identical to itself AND the world is identical to w_1'. As such, this condition is fulfilled by all individuals with respect to the particular possible world w_1 and nowhere else (as the second condition: the world is identical to w_1, is fulfilled only with respect to w_1).

[27] It also does not belong to the class *Triv* discussed by Duží, Jespersen, and Materna (2010, sec. 1.4.2.1).

[28] I am not using the temporal index for simplicity.

non-empty essential core *EC*. Individuals belonging to *EC* have *P* necessarily" (Duží, Jespersen, and Materna 2010, 68). Now, is the example of an individual property used in my counterargument to Tichý a case of a trivial individual property in this manner? No, it is not, because it does not have a non-empty essential core. There is no individual that possesses this property in every possible world. Duží, Jespersen, and Materna (2010) adopt the concept of 'essential core' as introduced by (Cmorej 1996). The essential core of a property refers to a subset that exists in every possible extension of the property. In the context of individual properties, the essential core consists of individuals who possess the property in every possible world. It follows straightforwardly from this definition that the individual property in my counter-example above lacks a non-empty essential core. This is because it is a property with an empty extension in all possible worlds except w_1.

This counter-argument seems to be relying on a world-indexing 'trick', like 'the US President at world w_1'. Within TIL, one can create an artificial property that no individual possesses except in one particular world (thanks to world-indexing) and which is nontrivial. The idea is as follows: with a non-empty collection of possible worlds within the base, multiple constructions construct these worlds. For instance, for every world within the type ω, there is a Trivialization of the world, as defined by the notion of construction and ramified type hierarchy. Consequently, having a particular world within the type implies the existence of its Trivialization within the ramified hierarchy. As a result, there are more complex constructions containing this Trivialization as a constituent.

I want to emphasize the artificiality of the example. Nevertheless, Tichý's argumentation was not exclusively aimed at 'non-artificial' individual properties but rather at all of them. Therefore, the argument of the test is susceptible to critique even with these kinds of examples. Once we establish that there are non-trivial individual properties, the extensions of which we can establish with respect to particular worlds without the need of empirical testing, the logical relation that Tichý's argument of test presumes no longer holds. These kinds of intensional entities, as well as constructions *v*-constructing these, do exist over the standard epistemic base of TIL. Therefore, we must consider them. If we leave them out, we are compelled to provide some arguments for this omission. Tichý's argumentation did not address these aspects.

I should add that it is not a standard practice to include Trivializations of particular possible worlds within the models usually presented in TIL-based research. Duží et al. explicitly emphasize this point in their methodology: "However, as we prefer to understand explicit intensionalization, the method is restricted to *variables* ranging over possible worlds, which may then be bound in a variety of ways" (Duží, Jespersen, and Materna 2010, 179). This is a *preference* rather than an inevitable route. Perhaps the simplest way to strengthen Tichý's argument concerning the analysis of existence is to limit the area of applicability of his arguments to the individual properties graspable via these kinds of constructions (i.e., including at most variables for possible worlds, not Trivializations). However, such a move would require further argumentation to avoid being *ad hoc*, especially considering the argumentation about triviality of existence as an individual property.

We can even agree with Tichý that if existence is to be modelled by an empirical property, it runs into absurdities. But the idea of varying domains is not identical to the claim that individual existence needs to be a factually testable property. A logician trying to analyse logics over such kinds of logical spaces need not to employ this assumption.

One could nevertheless ask whether the notion of a possible world, as implemented in TIL, consequently forces the individual existence to be a trivial property. Not really. Even if we begin with the pre-theoretical assumption that a possible world is understood as maximally consistent totality of facts, we need not model existence as a trivial property. TIL is based over partial functions and it is quite possible to model the statements containing individual names with respect to a world in which it does not exist, e.g., by partial propositions.

6. Conclusion

This paper evaluated the arguments supporting the assumed constant universe of individuals for all possible worlds within the framework of TIL and the models provided within it. The analysis delves into the core steps of these arguments and finds them lacking. The upshot is that the assumption need not be considered unalterable within the framework, even though it appeared as such for so long.

I do not intend to assert this as my definitive stance, however. Instead, I present it as a position that was not entirely refuted by Tichý's argumentation, even though it is widely assumed to be so by virtually all researchers in TIL. It is plausible that such a model of individual existence could lead to unwelcome consequences.

The notion of a constant domain of the universe is pivotal in choosing particular models within the ramified hierarchy of TIL. If the domain was not constant it could potentially necessitate changes in the models of several crucial notions, such as requisite. This could be undesirable, given that much research has been conducted under the presupposition of a constant domain. This paper is not a call for revision, but rather an invitation to provide additional arguments or bolster the existing ones to reinforce the assumption of a constant domain of universe for the semantic models of natural language phenomena in TIL.

Acknowledgments

I am grateful to the two anonymous reviewers for EuJAP for their valuable comments on the earlier versions of this paper.

Funding Information: This research was supported by the University of Oxford project 'New Horizons for Science and Religion in Central and Eastern Europe', funded by the John Templeton Foundation (subgrant 'The necessity of God's Existence: Divine Names and Modality'). The opinions expressed in this publication are those of the author, and do not necessarily reflect the view of the John Templeton Foundation.

REFERENCES

Berto, Francesco. 2008. "Modal Meinongianism for Fictional Objects." *Metaphysica* 9: 205-218. https://doi.org/10.1007/s12133-008-0033-z.

Cmorej, Pavel. 1996."Empirické esenciálne vlastnosti (eng. "Empirical Essential Properties"). *Organon F* 3 (3): 239-261.

Duží, Marie. 2017. "Property Modifiers and Intensional Essentialism." *Computación y Sistemas* 21 (4): 601-613. https://doi.org/10.13053/CyS-21-4-2811.

———. 2019. "If Structured Propositions are Logical Procedures then How are Procedures Individuated?" *Synthese, special issue on the Unity of propositions* 196 (4): 1249-1283. https://doi.org/10.1007/s11229-017-1595-5.

Duží, Marie, and Bjørn Jespersen. 2015. "Transparent Quantification into Hyperintensional Objectual Attitudes." *Synthese* 192 (3): 635-677. https://doi.org/10.1007/s11229-014-0578-z.

Duží Marie, Bjørn Jespersen, and Daniela Glavaničová. 2021. "Impossible Individuals as Necessarily Empty Individual Concepts." In *Logic in High Definition. Trends in Logic (Studia Logica Library)* 56, edited by Alessandro Giordani and Jacek Malinowski, 177-202. Cham: Springer. https://doi.org/10.1007/978-3-030-53487-5_9

Duží Marie, Bjørn Jespersen, and Pavel Materna. 2010. *Procedural Semantics for Hyperintensional Logic. Foundations and Applications of Transparent Intensional Logic*. First edition. Berlin: Springer.

Fine, Kit. 1994. "Essence and Modality: The Second Philosophical Perspectives Lecture." *Philosophical Perspectives* 8: 1-16. https://doi.org/10.2307/2214160.

Glavaničová, Daniela. 2018. "Fictional Names and Semantics: Towards a Hybrid View." In *Objects of Inquiry in Philosophy of Language and Literature. Studies in Philosophy of Language and Linguistics,* edited by Piotr Stalmaszczyk, 59-73. Berlin: Peter Lang. https://doi.org/10.3726/b14249.

Jespersen, Bjørn. 2015. "Structured Lexical Concepts, Property Modifiers, and Transparent Intensional Logic." *Philosophical Studies* 172: 321-345. https://doi./org/10.1007/s11098-014-0305-0.

———. 2019. "Anatomy of a Proposition." *Synthese* 196: 1285-1324. https://doi.org/10.1007/s11229-017-1512-y.

Jespersen, Bjørn, and Marie Duží. 2022. "Transparent Quantification into Hyperpropositional Attitudes De Dicto." *Linguistics and Philosophy* 45: 1119-1164. https://doi.org/10.1007/s10988-021-09344-9.

Kosterec, Miloš. 2020. "Substitution Contradiction, Its Resolution and the Church-Rosser Theorem in TIL." *Journal of Philosophical Logic* 49: 121-133. https://doi.org/10.1007/s10992-019-09514-y.

Raclavský, Jiří. 2010. "Co obnáší kontingentní existence individuí." *Organon F* 17 (3): 374-387.

———. 2020. *Belief Attitudes, Fine-Grained Hyperintensionality and Type-Theoretic Logic*. Studies in Logic 88. London: College Publications.

Reicher, Maria. 2022. Nonexistent Objects. *The Stanford Encyclopedia of Philosophy* (Winter 2022 Edition). Edited by Edward N. Zalta & Uri Nodelman. Accessed April 2024. https://plato.stanford.edu/archives/win2022/entries/nonexistent-objects/.

Tichý, Pavel. 1988. *The Foundations of Frege's Logic*. Berlin and Boston: De Gruyter. https://doi.org/10.1515/9783110849264.

Wildman, Nathan. 2016. “How (not) to be a Modalist about Essence.” In *Reality Making*, edited by Mark Jago, 177-196. Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198755722.003.0009.

Zalta, Edward N. 2003. “Referring to fictional characters.” *Dialectica* 57 (2): 243-254. https://doi.org/10.1111/j.1746-8361.2003.tb00269.

European Journal of Analytic Philosophy EuJAP | 2024 | Vol. 20 | No. 1 | 133-156
https://doi.org/10.31820/ejap.20.1.6

CAN WE DEFEND NORMATIVE ERROR THEORY?

Joshua Taccolini[1]

[1] Saint Louis University, USA

Original scientific paper – Received: 13/09/2023 Accepted: 05/06/2024

ABSTRACT

Normative error theorists aim to defend an error theory which says that normative judgments ascribe normative properties, and such properties, including reasons for belief, are never instantiated. Many philosophers have raised objections to defending a theory which entails that we cannot have reason to believe it. Spencer Case objects that error theorists simply cannot avoid self-defeat. Alternatively, Bart Streumer argues that we cannot believe normative error theory but that, surprisingly, this helps its advocates defend it against these objections. I think that if Streumer's argument is successful, it provides error theorists an escape from Case's self-defeat objection. However, I build upon and improve Case's argument to show that we could never even successfully defend normative error theory whether we can believe it or not. So, self-defeat remains. I close by offering some reasons for thinking our inability to defend normative error theory means that we should reject it, which, in turn, would mean that it's false.

Keywords: Normative Error Theory; self-defeat; theory defense.

Correspondence: joshua.taccolini@slu.edu

Introduction

An error is a mistake. According to normative error theory, we make a systematic mistake when making normative judgments such as "murder is wrong" because these judgments ascribe normative properties such as the property of being wrong, and such properties are never instantiated.[1] Normative error theory (hereafter, NET) is a global error theory about *all* normative properties, not just the moral kind.[2] Its proponents (hereafter, error theorists) deny the instantiation of both moral *and* epistemic normative properties, but they are split on whether reasons for belief carry normative content and consequently whether we can rationally believe NET.

Many philosophers have tried to undermine this theory by arguing that its defenders, in believing it, argue from a self-defeating position.[3] From my view, these objections all rely on talk about our ability to believe NET. I think this allows error theorists to escape self-defeat by adopting the cognitive attitude of non-belief toward the theory they defend. However, I want to argue that we could never even *successfully defend* NET, and so it won't matter whether we can believe it. I qualify "defend" with "successfully" to leave open various ways we might attempt to defend what we could never successfully defend and still call that "defense". You might think, for example, that a poorly constructed theory defense, even if doomed to fail, still fulfils the action description "defending a theory". My aim is to eliminate the possibility of ever finding success in defending NET. Showing NET indefensible by any plausible metric of success would be a significant and surprising result in its own right. However, it could be that some theories we cannot successfully defend are nonetheless true. I close, therefore, by offering some initial reasons for thinking that our inability to defend NET is very bad for normative error theorists since it gives us good reason to think NET is false. A full defense of these consequences, however, I leave for future work. My principal aim in this paper is to show that we could never successfully defend NET.

[1] Or such properties do not exist at all. This won't matter to my argument. I will target epistemic and meta-ethical notions of "reasons" and "normativity" and leave metaphysical commitments about such things aside.

[2] For example, Jonas Olson (2014) and Bart Streumer (2017). NET is an alternative to realist, non-cognitive, and reductionist views about normative properties, which, according to error theorists, each have fatal flaws of their own. For examples of non-cognitive views see Simon Blackburn (1993) and (2000). For an example of a reductionist view see Frank Jackson (2000), and for a non-reductive realist view, see Derek Parfit (1997) and Russ Shafer-Landau (2003). NET is historically about exclusively moral judgments (see Mackie 1977/1990).

[3] Bart Streumer (2013) thinks we cannot believe NET, while Stan Husi (2013), Olson (2014), and Christopher Cowie (2016) think we can.

Before advancing my argument, we should first consider what we require to successfully defend a theory. In lieu of a complete theory of theory defense and conditions for its success, I will propose a working definition here and a *necessary* condition for any successful theory defense later. It seems to me that what we mean when we say that someone has defended a theory T (implying minimal success) is that, they have (at least) provided an epistemic reason, relevant to the question of T's being true or false, which counts as a consideration against rejecting T.[4] Let's stipulate, then, that to *successfully* defend a theory minimally requires offering a reason which counts in favor of believing that T is true and works against believing that it's false. This definition means to exclude arbitrary, merely pragmatic, preferential, or crazy "reasons" for belief. One of my opponents aiming to successfully defend NET also excludes such "reasons". On this, more later.

In offering my working definition, I don't arbitrarily assign normative status to reasons for belief which count as reasons relevant to successfully defending a theory (hereafter, theory defense reasons); that is, I leave open whether theory defense reasons weigh normatively on belief. I think they do; but I arrive at that conclusion only on consideration of consequences following its denial. I will argue that without epistemic norms, one's theory defending position is self-defeating, that is, it provides opponents no theory defense reasons which is the aim of a successful theory defense.

To illustrate in a general way what I have in mind, consider Socrates defending some theory T. He first considers various arguments for and against T. He then offers reasons which constitute considerations in favor of T. Finally, he considers objections to his argument and devises replies which undermine these objections. In all this, I understand Socrates to be successfully defending T, where "defending" minimally involves providing reasons favoring the truth of T. As such, it seems to me that to successfully defend T, we should be able to perform at least one of the following actions:

> Providing a reason which constitutes a consideration in favor of T.
> Offering arguments or other evidence which favors believing T.
> Offering reasons against objections to T.

[4] My arguments will assume theory defense *success* or *failure* in terms of meeting this condition.

Therefore, in this paper I will understand ability to perform at least one of these actions as constituting a necessary condition for successful theory defense. I assume that all error theorists, whether the believing or unbelieving type, have *attempted* to perform at least some of these actions while defending NET. However, they must *successfully* perform at least some of these actions to *successfully* defend this theory.[5]

Surprisingly, I think that we *cannot* successfully perform these actions relative to the successful defense of NET. I think this because I think that theory defense requires that theory defense reasons weigh normatively on belief. Again, I do not assume *in advance* that theory defense reasons weigh normatively on belief and therefore theory defense is by definition a normativity-discharging enterprise. However, I do think that *on reflection* theory defense reasons *turn out* to weigh normatively on belief. And in this paper, I aim to show that NET strips its advocates (but not the rest of us) of access to normativity *even if they do not believe this theory*. If I'm right, we cannot successfully defend NET. And if we cannot successfully defend it, I think this a serious problem for it.[6]

This paper consists of five sections. In section I, I consider a recent objection from unavoidable self-defeat levied against error theorists. I do this to introduce error theorists to an escape route from self-defeat objections but also because my project will build and improve on this argumentative strategy. In section II, I analyze an argument for error theory's unbelievability to show how error theorists might use it to escape self-defeat but also to showcase the trouble with defending NET. With these two arguments considered, I shift in section III to constructing my own argument that we could never successfully defend NET. In IV, I suggest some initial reasons for thinking that that consequently we should reject this theory, which in turn would prove it false.

1. Why error theorists face self-defeat in arguing for NET

If reasons for belief carry normative content, then were NET true, there would be no reasons for belief, including reasons to believe this theory. Many philosophers, including some error theorists, have noted the paradoxical position of believing a theory according to which there are no

[5] *Believing* that I am providing reasons for belief when I am actually not providing any will not satisfy successful theory defense. As before, I exclude such reasons from my definition.

[6] I introduce what I find problematic in Section 4.

reasons for belief. For example, Terence Cuneo (2007), opposing this theory, argues that

> If they [error theorists] say that there are reasons to believe NET, their view is self-defeating. For the property of being a reason is a normative property, which does not exist if NET is true. But if error theorists say that there is no reason to believe NET, their view is polemically toothless. For if there is no reason to believe NET, it is not a rational mistake to reject this theory. (Cuneo 2007, 117–18; As quoted by Streumer 2013a, 203–4)

Cuneo takes reasons for belief to carry normative content. This allows him to formulate an objection from *self-defeat* in the sense of a performative contradiction error theorists commit while arguing for their position. [7] Error theorists are giving reasons to believe their view according to which there are no reasons for belief. Stan Husi, an error theorist, concedes this worry observing that skepticism about all normative reasons "appears to be cutting off the very justificatory branch it sits upon, seeking to engage [in] a dialectical enterprise while denying its currency" (2013, 429). Husi, along with Jonas Olson and Chris Cowie, instead proposes reasons of a different sort for believing NET which don't smuggle in normative content. [8] If their strategy succeeds, then in supplying these non-normative reasons for their position, error theorists are free from Cuneo's self-defeat objection.

However, Spencer Case (2020) argues that no matter how error theorists construe *reasons for belief*, they cannot avoid self-defeat.[9] Even *indicator evidence*—evidence for a proposition which does not count as a consideration in favor of believing it—such as premises logically entailing their conclusion won't, the following argument shows, save error theorists from self-defeat.[10] Case takes this to be sufficient reason to reject this theory. I disagree. I think error theorists can avoid self-defeat while defending NET. However, my project to show that NET cannot be

[7] I'm understanding "self-defeating" to refer to performative contradictions such as writing that I'm not writing, and "self-refuting" to refer to propositions and arguments which contradict themselves such as "there are no universal truths" (see Mackie 1964).

[8] For example, see Olson (2016, 461–73).

[9] Mustafa Khuramy and Erik Schulz (2024) disagree, but as will be clear in what follows, their objection from the ambiguity of self-defeat attribution does not affect my arguments (nor, in fact, the crux of Case's as I present it below). I do not have space to discuss.

[10] Streumer (2017b, 172, n. 3) replies by enlisting indicator evidence taken not to count as a consideration in favor of belief.

successfully defended crucially adopts elements of Case's argument. So, his argument is worth reproducing at the start. Here it is in two steps.

Self-Defeat Argument

Step 1

(1) Error theorists are committed to the self-defeating proposition, "NET is true, but I have no reason to believe that".
(2) If adopting any philosophical position commits us to a self-defeating proposition, then we should reject that position.

We should reject NET.

Step 2

(3) If we should reject NET, then NET is false.

Therefore, NET is false.

Self-Defeat Argument doesn't stop at **Step 1** because NET could still be true even if we should reject it. For example, a utilitarian might have practical reasons to reject an epistemically justified philosophical position (Case 2020, 3). However, **Step 2** capitalizes on the normative property ascribed by (2), namely, the property of being obligatory to reject theories entailing self-defeating positions. If we *should* reject NET, then there is at least one instantiated normative property, but NET eliminates normative properties, so it's false.

(1) and (2) need support. In support of (2), Case argues that if error theorists are willing to bite the bullet and accept that their position is self-defeating, they should be willing in principle to accept other equally counterintuitive positions that, say, reject a proscription against killing and eating our own children or a proscription against holding contradictory beliefs, provided that such positions are less counter-intuitive than accepting a self-defeating position. After all, that a self-defeating theory correctly represents the world is already highly counterintuitive, so there is no reason, in principle, that the proponent of such a theory should reject comparably counterintuitive commitments.

The whole argument turns on (1). If there are reasons for belief of a sort which do not carry any normative content implicit or otherwise, (1) is false. In support of (1), Case offers the following:

Weak Normativity Argument

(4) The normative error theorist’s partisanship toward the epistemic domain either makes a normative difference to him or it does not.

(5) If it does not, then the error theorist remains committed to self-defeating propositions.

(6) If it does, then NET is false.

Therefore, NET is either self-defeating or false.

It is here that I find a resource for my own case against NET. The dichotomy between reasons which make a normative difference to us and those that don’t is, by my lights, crucial to seeing the problem for NET. How does Case put this distinction to use? Case contends that if error theorists can offer only reasons for believing NET which make no normative difference to them, then they can offer only reasons which need not make any difference to opponents in the debate. If reasons for belief are not considerations in favor of belief, considerations, that is, which obligate one to at least refrain from unreflectively dismissing them before rational deliberation, error theorists are once again polemically toothless. With Cuneo (2007), Case thinks that without considerations which weigh normatively on believing NET, error theorists are polemically toothless, which is to say that they’re in a self-defeating position (premise 5). Alternatively, if error theorists can offer reasons for believing NET which *do* make a normative difference to them, they now re-introduce normativity into discussion, which is inconsistent with NET (premise 6). Either way, error theorists cannot avoid self-defeat.

In this respect, Case notes that if error theorists want to insist on entitlement to reasons for believing NET—where “reasons” are understood non-normatively—self-defeat persists. Error theorists are here committed to saying, “Error theory is true, but there is no reason—of a kind that anyone need take *the least bit seriously*, all things considered — for anyone to believe it” (2020, 8; emphasis mine). Stripping reasons for belief from any kind of binding authority might save our ability to *believe* NET (contra Cuneo 2007), but it will not save error theorists from self-defeat.

However, the problem with *Weak-Normativity Argument* is that it leaves open an escape route for error theorists. Both Cuneo’s objection and *Weak-Normativity Argument* assume that error theorists are *committed* to believing NET, that is, they assume that error theorists always believe the theory they defend. The “self” in “self-defeat” refers to a problematic relationship to believing NET. But what if we *cannot* believe this theory? If we cannot believe NET, then error theorists do not defend it from a place of commitment to it. Error theorists can then avoid self-defeat altogether by adopting a cognitive attitude of *non-belief* in this theory. No performative contradiction arises from defending a theory which eliminates reasons for belief if I don’t believe what I’m defending.

This is how I understand Bart Streumer’s recent arguments for NET. Streumer (2013a; 2017a) argues that we cannot believe NET, but that, surprisingly, our inability to believe it fortifies error theorists against self-defeat and other *reductio ad absurdum* objections. After all, our inability to believe a theory does not make it false. Case reads Streumer’s position as biting the self-defeat bullet, but we don’t need to construe Streumer’s position this way. If I’m right, and if Streumer’s argument for NET’s unbelievability is successful, error theorists can ward off self-defeat objections without appealing to alternative reasons for believing it. Instead, they can claim to successfully defend it by insisting that we *cannot* believe it. To show all attempts at NET defense futile, I must therefore show that not even NET’s unbelievability can restore its polemical force in the debate. I next introduce Streumer’s argument as a potential escape from self-defeat objections, but in so doing, I observe what I consider a worse problem for error theorists.

2. How error theorists escape self-defeat only to face theory defense futility

Streumer’s argument for NET’s unbelievability (hereafter, *Unbelievability Argument*) provides an escape from self-defeat and many other objections. But I contend that this argument also betrays the necessity of normativity for successfully defending a theory. In this section, I analyze *Unbelievability Argument* and demonstrate its force in blocking objections to NET. However, I end by observing its proponents’ unintended application of normative reasons for belief.

The relevant terms in *Unbelievability Argument* are *reasons for belief*, *belief*, *normative judgments*, and *normative properties*.[11] Streumer qualifies *belief* to mean full, confident, non-compulsory, *rational* belief which

[11] Streumer (2011) argues that normative properties (if they existed) are irreducible to descriptive properties. Ontological commitments regarding properties are irrelevant here. “Favoring relation”, e.g., can replace “property” without affecting my argument.

excludes partial, somewhat confident, compulsory, or crazy belief (Streumer 2013a, 197; 2017a, 7).[12] By *rational*, Streumer only means closed under believed entailment (believing what I believe is entailed by my beliefs), which he takes to be a descriptive property with no normative bearing on belief.[13] By *reasons for belief,* Streumer means any consideration in favor of a belief, and he takes considerations in favor of a belief to weigh normatively on belief.[14] In support of this, he says that "reasons for belief are considerations that we base our beliefs on, and we cannot base a belief on a consideration without making at least an implicit normative judgment" (2013a, 198). *Normative judgments* are beliefs which aim to represent the world. So, when NET says that "normative judgments are beliefs which ascribe normative properties", this is a cognitivist position about normativity such that our normative judgments aim to represent instantiated normative properties (Streumer 2013b). An example of such a judgment may simply be that we ought to believe in light of the supporting evidence. In what follows, I take these terms just in the sense Streumer takes them.

NET can be construed as the conjunction of the following two propositions:

> (J) Normative judgments are beliefs which ascribe normative properties.
> (P) Normative properties are never instantiated.

Streumer argues for three claims about this theory. He argues that NET is unbelievable, that NET's unbelievability undermines objections which have been made against it, and that we can come close to believing this theory and it may be a rational mistake not to. With these claims in hand, error theorists can argue that though NET cannot be believed, this does not make it false, and competing normative theories such as normative realism (including reductive realism) and normative non-cognitivism *are* false, which makes NET more likely true.

[12] For Streumer (2013a), partial belief differs from coming close to believing NET. This will be made clear in what follows.

[13] Says Streumer: "belief is rational in a certain sense: it is closed under believed entailment, since the person who has this belief believes what he or she believes to be entailed by this belief, and it is not believed to be unsupported, since the person who has this belief does not believe that there is no reason for this belief. But that is no objection to my argument (…). Being closed under believed entailment and not being believed to be unsupported are descriptive properties" (2017a, 7f).

[14] "The property of being a reason for belief, in the sense of a consideration that counts in favour of this belief, is a normative property" (Streumer 2013a, 197).

Streumer begins his argument by proposing two claims about belief:

> (B1) We cannot fail to believe what we believe is entailed by our own beliefs.
> (B2) We cannot have a belief while believing that there is no reason for this belief.

If these claims are true, error theorists can then argue as follows. Anyone who believes NET believes that there are no normative properties. Reasons for belief are normative properties, so if NET is true, there are no reasons for belief.[15] So, by (B1), anyone who believes NET believes that there are no reasons for belief. But by (B2), we cannot have a belief while believing that there is no reason for this belief. Therefore, NET is unbelievable.

As it stands, *Unbelievability Argument* does not show that there are no reasons to believe NET. The conditional claim "if NET is true, there are no reasons for belief" does not (alone) entail that there are no reasons to believe NET. So instead, Streumer (2013a, 199–200) offers the following two claims about reasons:

> (R1) There cannot be a reason for someone to do *x* if this person cannot do *x*.
> (R2) There cannot be a reason for someone to believe that *p* if this person cannot believe that *p*.[16]

If these claims are true, error theorists can then argue as follows. We take reasons for a belief to count in favor of that belief just as reasons for an action count in favor of that action. So, if (R1) is true of actions, then it follows that (R2) is true of beliefs. But if (B1) and (B2) about beliefs are true, then we cannot believe NET. By (R2), we cannot have a reason to believe what we cannot believe. Therefore, there are no reasons to believe NET.

In summary, we can construct *Unbelievability Argument* as follows:

(P1) According to NET, normative properties are never instantiated.
(P2) Reasons for belief are normative properties.
C1 Therefore, if NET is true, there are no reasons for belief.

[15] Streumer (2016; 2017a, §51) argues that reasons for belief are normative properties.
[16] I do not have space to give Streumer's defense of these claims. In what follows, I grant them for the sake of argument.

(P3) Anyone who believes NET believes C1.
(P4) We cannot have a belief while believing that there is no reason for this belief.
C2 Therefore, we cannot believe NET.
(P5) We cannot have a reason to believe what we cannot believe.
C3 Therefore, there are no reasons to believe NET.

The argument does not stop at C2 because knowing that <if the error theory is true, there are no reasons to believe it> cannot make us believe that there are no reasons to believe it *if we cannot believe the antecedent of this conditional claim.* So instead, we need another reason to believe there are no reasons to believe NET. (P5) provides this reason.

If *Unbelievability Argument* is successful, error theorists are in an improved position in the dialectic. Normally, demonstrating a theory's unbelievability would count against that theory, but in this case, it *supports* normative error by protecting error theorists from objections directed at believing error theorists (e.g., Olson 2014). If error theorists cannot believe NET, these objections miss the mark.

Finally, Streumer contends that we can come close to believing NET so long as *coming close to believing a theory* is less than full belief in that theory, meaning that this claim does not contradict (B2). But coming close to belief is not merely partial or weak belief in NET; rather, it is to be convinced that these arguments together *seem* to show that NET is true (2013a, 203).[17] Streumer argues that we can *come close to* believing this theory by believing arguments in favor of (J) (that normative judgments are cognitive) without explicitly believing (P) (that normative properties do not exist) and by, at a later time, believing arguments in favor of (P) without explicitly believing (J). We can also come close to believing NET by believing arguments against alternative theories; for example, we can believe that, contra irrealist, theories normative judgments really do aim to represent the world, and we can believe that contra realist theories there really are no normative properties.

The strength of *Unbelievability Argument* lies in its ability to block objections. First, recall Cuneo's (2007) observation that error theorists are either arguing from a self-defeating position if there are reasons to believe their theory or are polemically toothless in the debate if there are no reasons for belief. Says Streumer in reply:

[17] Streumer does not think that there *seems* to be sound arguments that show that NET is true but that there *are* sound arguments which together seem to show that NET is true.

> [This] only shows that *if NET is true,* there is no reason to believe NET. And the belief that this conditional claim is true will only make us believe that there is no reason to believe NET if we already believe NET, which I have argued we cannot do. (Streumer 2013a, 204)

NET's unbelievability blunts the force of Cuneo's (2007) objection.[18] Error theorists here argue for NET without believing it, thus retaining polemical teeth in the fight and avoiding self-defeat.

Does this mean error theorists are guilty of a form of bad faith in defending a theory they don't believe there is any reason to defend? No. Our ability to *come close to* believing NET at least partially returns error theorists' dog to the fight. Says Streumer:

> Since we can come close to believing the [normative] error theory in these ways, there can be reasons for us to come close to believing it in these ways, and it can be a *rational mistake* if we do not come close to believing it in these ways. (Streumer 2013a, 204; emphasis mine)

If there are reasons for coming close to believing NET, then error theorists have reason to argue for this theory. And if so, they are saved from bad faith objections.[19] In this way, *Unbelievability Argument* is a powerful strategy for error theorists: it provides an escape from the self-defeat which afflicts every card-carrying error theorist by denying everyone a card, but it also preserves reasons for taking NET seriously since it may be true and it may be a mistake to fail to come close to believing it.

However, we are now beginning to see the trouble for error theorists with attempting to successfully defend NET.[20] NET eliminates normative properties. But talk about the *strength* of *Unbelievability Argument* in

[18] For example, Shah (2010) argues that if NET is true, there are no beliefs. Streumer replies: "Of course, it then remains the case that if [NET] is true, there are no beliefs. But if my arguments are sound, this cannot make us think that there are no beliefs, since we cannot think that the antecedent of this conditional claim is true" (2013a, 201).

[19] Says Streumer: "If my arguments are sound, however, no one can believe [normative] error theory, not even those who defend this theory. And to be in bad faith is to close one's eyes to the truth, not because one *cannot* believe it, but because one does not *want* to believe it. If defenders of [normative] error theory come close to believing it in the ways I have described, they are as far from being in bad faith as it is possible to be" (2017a, 177–78).

[20] For other objections to which Streumer has replied, see Marianna Bergamaschi Ganapini (2016) and Alexander Hyun and Eric Sampson (2014).

blocking objections seems to be an appeal to this argument's polemical force in the debate; that is, it seems to appeal to its *normative* difference to us. Just so, talk about "reasons for coming close to believing" and the "rational mistake" we commit in failing to do so pack no punch in the dialectic if such talk is stripped of anything which weighs normatively on belief. In what follows, I develop this observation into an argument for the indefensibility of NET.

3. Why we could never succeed in defending normative error theory

Can we successfully defend a theory without presenting any reason for believing it? What about a theory according to which there are *no* reasons for believing it? Can we sincerely do these things? If you are a normative error theorist, you might think that we can. After all, if we *cannot* believe NET then we don't, and if we don't believe it, perhaps we are entitled to marshal normative reasons for belief in its defense.

However, even if we concede that error theorists' belief in reasons for belief remains safe[21] while defending NET (and I will make this concession), this concession won't do enough to ensure the possibility of a successful defense. This is because to successfully defend NET, we must meet the abovementioned necessary conditions for successful theory defense which we can consolidate into the following:

> **Theory Defense Condition:** We can successfully defend a theory T only if it is possible for us to offer at least one theory defense reason which counts as a consideration in favor of T.[22]

And I contend that with respect to NET we cannot meet this condition (hereafter, **TDC**). Before examining this claim, first notice how intuitive this condition is for successful theory defense. Theory defense is a communicative act wherein we express theory defense reasons to

[21] Here and throughout *safety* refers to immunity from charges of incoherence, inconsistency, or polemical toothlessness, such as: "You believe in (or utilize) normative reasons for belief while defending NET, but you also claim that no such things exist".

[22] Possibility (and necessity) referring here only to what is practically and epistemically possible *for us*. To accept this condition, we need not hold *ontological* commitment to the *existence* of a theory defense reason, but we must at least hold *epistemic* commitment to the belief that included in the set of all reasons for belief is at least one which is in principle epistemically accessible to us such that someone could practically offer it in the course of performing the action of theory defense. **TDC** is a constraint on theory defense not on the existence of properties.

interlocutors. And it won't be just any reasons which count toward success but reasons which actually favor T's being true. So, if I *cannot* even in principle offer at least one theory defense reason which favors T, any reason communicated will be *no* reason against rejecting it, and in that case, I will never have successfully defended T, no matter how many alternative (theory defense irrelevant) reasons I offer. Now suppose S thinks it is possible that someone *could* successfully defend T. Even if S remains unaware or unable to express a theory defense reason favoring T herself, if she thinks that T could be successfully defended by someone, surely we should take her to think that at least one such reason is available to some other potential defender of T. By contrast, if S correctly believes that *no* theory defense reason favors T such that no one could ever advocate for T by providing reason against believing it's false, we will naturally say that S correctly believes defending T a necessarily futile exercise.

Even so, it might look to error theorists like I'm smuggling normativity into a necessary condition for successful theory defense. After all, *being correct to believe* looks like a normative property.[23] If it is, then it looks like I beg the question against error theorists by introducing in advance a condition for theory defense which requires that we *correctly believe* it is impossible to offer at least one theory defense reason in favor of T. The same is true if I assume that *theory defense reasons*, here required for theory defense, are themselves normative properties.

As before, I am here only considering theory defense reasons in the sense of reasons which motivate, on pain of being irrational were they thoughtlessly dismissed, to avoid rejecting T. And I continue to remain neutral about their normative status while advancing my argument. The same goes for *being correct to believe.* If it is correct to believe that two and two make four, and I believe this, then it is irrational for me to reject this claim, whether or not I believe being correct weighs normatively on belief. *Later* I will propose that we actually do have independent reason to accept the normativity of theory defense reasons, but *here* I rely only on what I take to be acceptable to error theorists. So, I do not beg the question against NET.

Would NET fail to meet **TDC**, *Unbelievability Argument* would be of no use for a successful defense of it. There would be no epistemically relevant reason for opponents to believe the premises of the argument nor to reconsider NET in light of it. *Unbelievability Argument* would give us

[23] Streumer (2011) considers it normative.

no such reason to even *come close to* believing NET nor help us see why it would be a rational mistake to fail to do so since a theory's supporting evidence just is a theory defense reason favoring it, and we cannot provide evidence for a theory which no theory defense reason supports. Theory defense reasons here extend far beyond reasons for believing the theory itself. They extend to reasons for believing at least one reason favors accepting or disfavors rejecting a theory T, reasons for believing the premises of arguments whose conclusion advances T in some way, and reasons for believing that objections marshalled against T fail. If belief comes in degrees, then reasons for belief extend even to those reasons which raise our credence level in T to any degree.

Yet NET *does* fail to meet **TDC**. Clearly, if theory defense reasons are normative properties, then NET is false. But as before, if they are epistemically non-normative, then opponents can safely ignore them. Earlier, I introduced Case's *Weak-Normativity Argument* which puts a dilemma to believing error theorists who in support of NET either offer reasons for belief which render them inconsistent or reasons for belief which make no difference to opponents. We saw how error theorists can escape this dilemma. However, I have now re-directed this dilemma toward theory defense reasons, and I no longer see an escape route for error theorists. Epistemic norms are what provide polemical force to an argument. What we *ought* to believe compels us on pain of epistemic vice to follow the imperative. We ought not hold contradictory beliefs, for example, on pain of being irrational. Redefining "irrational" in merely psychologically descriptive terms strips it of its argumentative force in philosophical discussion. Just so, error theorists might offer alternative weapons of defense such as reasons of personal preference or pragmatic reasons for advancing NET. But if such reasons are normatively bankrupt, unless I share this preference or those practical goals which render NET useful to me, I can safely ignore these reasons in the debate. Reasons which we can safely ignore fail to count as considerations in favor of a theory's being true. So, NET fails to meet a necessary condition for rendering even possible a successful defense of it.

It might be objected that so long as error theorists present arguments whose premises, if true, guarantee the truth of NET, and evidence that makes these premises likely to be true, they adequately defend NET. **TDC** appears to problematically sunder truth from normativity, however, since if NET is true, there are no theory defense reasons, so defendants here fail to meet **TDC**. Yet, there is nothing stopping error theorists from offering evidence that supports the truth of NET by way of premises and a conclusion or by evidence against objections to NET. So, if NET is true, error theorists—despite offering valid arguments with likely premises which conclude that NET is true—have not successfully defended NET which seems absurd.

In response, we should first note that, as before, whether theory defense reasons are normative properties will be a matter of disagreement between error theorists. Error theorists who reject their normative status will, therefore, read **TDC** as void of normative commitments in which case NET's defendant has *not* failed to meet **TDC** in the above objection. However, this view faces the *Weak-Normativity Argument* as we've already seen since opponents can safely ignore normatively bankrupt reasons offered in defense of NET.

On the other hand, error theorists who accept the normativity of theory defense reasons which include reasons for belief remain consistent only if they also accept the normative status of the relevant notions of *evidence* and *truth* which, as before, are *also* theory defense reasons. After all, what epistemic value would these notions have in relation to theory defense if we have no epistemic obligation to prefer *rational, evidentially supported, true* beliefs over *irrational, evidentially unsupported, false* ones? And to say that we *should* prefer the former over the latter is to say that these notions carry normative content. To be sure, *if NET is true*, all judgments deploying theory defense reasons would here be false since there would be no epistemic norms. But can we believe the antecedent of that conditional claim? After all, it was the normative status of considerations which count as favoring NET which *Unbelievability Argument granted* so as to show that we *cannot* believe the antecedent of these conditional claims which suppose the truth of NET. The purpose was to ward off *reductio ad absurdum* objections. By that line of reasoning, NET's being true cannot make us *believe* that no theory has ever been successfully defended since we *cannot* believe NET. Yet all of this shows only that *even if NET is true*, error theorists despite all appearances have *not* successfully defended it—not because my arguments divorce truth from normativity but because NET entails the unbelievable consequence that no one has ever successfully defended a theory.

Finally, you might still think that we simply do observe successful defenses of what the defender believes is indefensible. Suppose, for example, that a professor is teaching Kantian ethics to undergraduates. Suppose the professor presents Kant's main reasons in support of his ethical theory and subsequently answers every students' objection just as she thinks Kant would (or should) answer it. Suppose further that this professor doesn't find her students' objections convincing; rather, she thinks that a Kantian could easily dispense with them. And yet, let's imagine this professor to be strictly committed to a form of act consequentialism, and that she believes that there simply are no theory defense relevant reasons in favor of being a Kantian at all. Since she has replied to her students' objections, can we not say that she has defended what she personally believes there are no good reasons to believe, and is therefore a counterexample to **TDC**?[24]

The objection clarifies the difference between failing **TDC** and less disastrous ways a theory might lack support. **TDC** does not speak to theories we personally believe lack even one theory defense reason favoring it. Theory defenders in such cases can remain open to the possibility of being surprised by an objection not considered. By contrast, if, in the above case, the professor *correctly* believes it *impossible* for *any* potential Kantian ethics theory defender to offer any theory defense reason favoring it, she could not *also* believe (and remain consistent) that responding to objections *counts* as a reason in favor of this theory's being true; that second belief of hers would just be false. If the professor correctly believed that no theory defense reason could ever become available to any Kantian ethics defender, she should therefore say that not even her replies to her students' easily dispensable objections give her or them *any* reason favoring Kantian ethics; otherwise, her replies *themselves* would work *against* Kantian ethics (if indefensible) by instantiating those very properties (theory defense reasons) she denies are available to any defender of Kantian ethics. If she claims to be defending Kantian ethics, she surely is not, on these suppositions, *successfully* defending it. This is precisely what makes NET so unusual. We do not normally rule out *in advance* the possibility of any theory defense reason supporting belief in a theory. Yet, unlike Kantian or consequentialist moral theories, NET *itself* rules out the possibility of any attempts (including responding to weak objections) counting as considerations in favor of its being true. If I, for example, were to defend NET against my students' easily dispensable objections, while correctly believing that there are no theory defense reasons in favor of NET—correctly believing that NET fails to meet **TDC**—I would have to concede to these students that insofar as responding to objections counts toward successfully defending NET, my replies were, in fact, utterly futile toward its defense, *even while successfully responding to their objections*.

[24] I thank an anonymous referee for this objection.

4. Theory defense failure is not safe

If my arguments preventing successful defense of NET are sound, what does this mean for error theorists? I take our inability to successfully defend NET a serious concern for error theorists, but I leave a complete exploration of the problems for later work. Instead, in this section I offer some initial suggestions highlighting the sort of problems which lurk behind NET's failure to meet a necessary condition for successfully defending theories. I think that the arguments sketched below give us, at the very least, good reason to warn error theorists not to completely ignore our inability to successfully defend NET.

You might think that our inability to successfully defend NET is no problem for error theorists. Recall that many objections to NET target error theorists who appear to argue from a self-defeating position. As was shown, this does not entail that NET is false. Like global skeptics, error theorists could insist that NET might be true and that, in support, good arguments show competing normativity theories to be false. If so, then we might think of error theorists as normativity messengers with a skeptical message. Shooting the messenger won't absolve us of the skeptical problem. You might think, for example, that some claims are not successfully defensible yet just as conceivably true as conceivably false. Consider the claim that the number of stars in the Andromeda galaxy is an even number. Suppose we lack sufficient information to favor odd or even. If we cannot present information which counts as evidence favoring an even number, it looks like we cannot successfully defend this claim. However, this does not mean that we *should* reject the claim. After all, the number of stars may in fact be even. Since theories consist of claims, it might be that some theories are true, yet, we cannot successfully defend them on grounds of insufficient information. After all, some philosophers think that skeptical arguments are valuable not because anyone believes their conclusions but instead because they teach us important lessons for our epistemologies and because it is not obvious where these arguments go wrong.[25] If that's right, then can we not provide along these same lines some safety for NET from my objections?

Before responding, it is worth noting that error theorists do not take themselves to be offering a skeptical puzzle for us to solve collectively.[26] Streumer only tells us we can't believe NET enroute to defending it, and

[25] John Greco (2000, 3) argues that for these reasons skeptical arguments should not be dismissed even if skepticism is self-defeating for anyone who accepts it since the skeptic claims to know that no one knows. See, also, David Enoch (2006, 183–84).

[26] Some (e.g., Joyce 2014, 843) take themselves to be "card-carrying proponents" of NET.

he wants us to join him in coming close to believing it by rejecting opposing views. All the same, while arguments for NET are still valuable for our moral epistemologies, and we shouldn't reject a theory on rhetorical grounds alone, I think our inability to successfully defend NET gives us *philosophical* reasons to reject it, even while conceding the rhetorical point that problems for the messenger don't disprove the message. That is, I think that if we know in advance about a theory, T, not merely that we *lack sufficient information* to mount a convincing case in favor of T (as in the odd or even case above), but that T rules out *tout court* the possibility of any theory defense reason *ever* becoming available to *anyone*, then it seems like the rational thing to do is to reject that theory. What I am suggesting here is that if a theory fails to meet **TDC**, that fact alone seems to give us good reason to reject it. The problem for NET in failing to meet **TDC**, isn't just that it lacks favoring evidence "in hand" to present to opponents (evidence which may yet come), but that *nothing could ever count as evidence* since the theory itself either guts favoring relations of normative force or flatly eliminates them. **TDC** failure means no evidence is *possibly* available to us to offer on a theory's behalf not that we *currently* suffer some access limitation which may one day be overcome. As such, it is hard to see any serious reason to consider it a possibility any longer even if we fall short of disproving it. Is this not an *ad hoc* move against error theorists? To be sure, it is difficult to think of any theory like this other than NET (as far as I know, no other theory eliminates *all* normative properties or at least all instantiated ones). All the same, it seems to me an independently plausible principle that demand at least one theory defense reason be possibly accessible to us to offer in favor of some theory to ensure the possibility for us of successfully defending it. If my intuition is correct, we won't need any other theory to justify application of the principle toward rejecting NET.

In suggesting that we should reject NET, I am not antecedently ruling out the possibility that NET is true. Yet, if theory defense reasons carry normative weight, then we are closer to knowing that it's false. And it looks like they are. As before, *denying* defense reasons' normativity results in a self-defeating position for error theorists, and while *accepting* their normativity is a problem for defenders of NET, it is no problem for anyone else, who, like me, thinks that we are justified in rejecting it. Consider that if I'm wrong, and it would be a rational mistake to reject NET, then despite no theory defense reason possibly counting in favor of it, we would still not enjoy justification in rejecting it. This would mean that, even were it true, the truth of NET would itself be no reason to believe it! This is absurd.[27] Likewise, any amount of evidence—such as reasons for belief, arguments, responding to objections (etc.)—marshalled against NET would not, by supposition, justify rejecting it.

Still, you might recall that Streumer advocates adopting the cognitive stance of *coming close to* belief. You might therefore think, following Streumer, that error theorists could offer reasons to "celieve" NET sufficient for theory defense where "celieving" is between rejecting and believing a theory. After all, we have seen that Streumer thinks there are enough reasons favoring NET such that it is a rational mistake to reject it. To hold that defending NET requires giving reasons for *belief* and not *celief* in NET is question-begging. Therefore, NET remains defensible. Note, in response, that *Unbelievability Argument* might run by parity just as well on "reasons for celief" as "reasons for belief". If so, then we cannot celieve NET either. Of course, if the parity argument fails, then we *can* celieve NET. But the determining factor remains the same: are reasons for celief normative on celief or not? Our Case-style dilemma returns: if no, NET is indefensible for weak-normativity reasons; if yes, NET is false. The same problem faces error theorists' definition of "theory defense". If error theorists insist that they "defend" NET—where "defend" smuggles no normativity into play—opponents can safely ignore whatever "theory defense reasons" they offer, and otherwise, error theorists rely on the instantiation of what NET denies is instantiated. In either case, the claim that we're justified in rejecting NET is not affected by introducing normatively deflated definitions of these terms.

The plausibility of my rejection proposal might come to light in the following analogy. Suppose you're told about a product called MoneySucker©. The only function of this product is to suck money and give nothing in return. It would be silly for a consumer to buy this product. However, suppose it turns out that *no one* can buy this product. It's not for sale and never will be. Perhaps we're not justified in rejecting the product out of hand. After all, we can't buy it, so perhaps we can't ever be sure that it would be a bad purchase. Now suppose you encounter a street salesman promoting MoneySucker©. He yells to passersby: "End all spending", "Purchases are evil", "MoneySucker© is the only worthy product remaining because we have no reason to buy things!" You ask him why he's selling MoneySucker© if it can't be purchased and we have no reason to buy things? He replies: "For a small sum, I'll tell you why". But why should you spend to learn why spending is evil and that a product which is not for sale whose function is to suck

[27] And, like before, favors the view that truth and evidence weigh normatively on belief.

dry all our spending power is the only worthy product remaining? You should reject that offer. Similarly, you should not "buy" the arguments of proponents of a theory that "sucks dry" the "currency" of theory defense reasons needed to successfully defend it, which places some of its "sellers" without opponent "purchase", and which cannot be "bought into" leaving its sellers unable to "make the sale", that is, unable to successfully defend it. In other words, the reasonable response when presented with a theory (even if NET is the only one) according to which no theory defense reasons could ever count as considerations against rejecting it is to reject it.

If we should reject NET, then it would follow straightforwardly that NET is false. To be sure, showing that we *should* reject a theory does not always mean that this theory is false. Plausibly, there are claims which we should reject without knowing that they are false. You might think, for example, that we cannot know that our reason is reliable. Even so, it would be rational to reject the claim that our reason is entirely unreliable even if we cannot be certain that this claim is false. However, things are different for NET. Recall that if any normative property is instantiated, NET is false. Now consider the following argument.

Indefensibility Argument

Step 1

(7) We cannot successfully defend NET because it fails to meet Theory Defense Condition (TDC).
(8) If a theory fails to meet TDC, we should reject it.

We should reject NET.

Step 2

(9) If we should reject NET, NET is false.

Therefore, NET is false.

If (8) is true, it's easy to see how this argument would succeed. I've already shown that (7) is true. In this section, I've begun to motivate (8) by offering reasons for thinking that we are justified in rejecting NET on grounds that it fails **TDC**. Unlike the self-defeat argument, (8) does not depend on anyone adopting the attitude of belief in NET while defending it. And as in the Self-Defeat Argument, rejecting (8) looks at first blush more problematic than rejecting other widely held intuitions such as the

Law of Non-Contradiction or the claim that it is impermissible to torture innocent children for fun. Rejecting the requirement for possibly offering even one theory defense reason in favor of the theory we defend looks close to saying that in philosophical discourse, there is no difference between good arguments and bad ones. In other words, the intuitive costs of rejecting (8) seem to me so high that it clearly looks like a rational mistake to do so.

You may disagree. Yet even if you're right, supposing (8) is true, what clearly follows is that we *should* reject NET because there would be at least one normative property instantiated, the property of being right to reject a theory we know we could never possibly successfully defend; and if so, then NET is false. The argument would succeed regardless of whether or not error theorists are committed to a self-defeating proposition. According to NET, no normative properties are ever instantiated, so (9) would be true by definition. The *prima facie* plausibility of such an argument, even if not yet completely convincing, I think, already shows that the consequences for error theorists following from our inability to successfully defend NET are *not* benign, that is, they are the kind that (never-successful) defenders of NET cannot safely ignore.

5. Conclusion

I have argued that an argument for NET's unbelievability provides an escape to a self-defeat objection to this theory. But it's a pyrrhic victory, since from these arguments, we can now clearly see that any attempt to defend NET is futile. At first, it might seem crazy to argue that there is a theory which we cannot successfully defend. But when we consider how strange it is to try to defend a theory which entails that there are no normative reasons to believe it, we realize that our inability to succeed in defending this theory is no less strange. I concluded by suggesting that as a result we should reject NET. And if we *should* reject NET, then there is at least one normative property instantiated, and NET is false. I have not here provided a complete defense of these two consequences following from my central objection to NET, but their initial plausibility strikes me sufficient to make theory defense failure a significant concern in this case.

Acknowledgments

Thanks especially to Bart Streumer, Spencer Case, Christian Lee, Hud Hudson, Joe Milburn, John Greco, Logan Gage, and Mark K. Spencer for comments on early drafts. Thanks also to Anthony Sloan, Lukas Rowland, and Jonathan Taccolini for helpful discussion, and to two anonymous referees for helpful comments on earlier drafts.

REFERENCES

Bergamaschi Ganapini, Marianna. 2016. “Why We Can Still Believe the Error Theory.” *International Journal of Philosophical Studies* 24 (4): 523–36. https://doi.org/10.1080/09672559.2016.1203978.

Blackburn, Simon. 1993. *Essays in Quasi-Realism*. New York: Oxford University Press.

———. 2000. *Ruling Passions: A Theory of Practical Reasoning*. Oxford: Oxford University Press.

Case, Spencer. 2020. “The Normative Error Theorist Cannot Avoid Self-Defeat.” *Australasian Journal of Philosophy* 98 (1): 92–104. https://doi.org/10.1080/00048402.2019.1582079.

Cowie, Christopher. 2016. “Good News for Moral Error Theorists: A Master Argument Against Companions in Guilt Strategies.” *Australasian Journal of Philosophy* 94 (1): 115–30. https://doi.org/10.1080/00048402.2015.1026269.

Cuneo, Terence. 2007. *The Normative Web: An Argument for Moral Realism*. Oxford: Oxford University Press.

Enoch, David. 2006. “Agency, Shmagency: Why Normativity Won’t Come from What is Constitutive of Action.” *The Philosophical Review* 115 (2): 169–98. https://doi.org/10.1215/00318108-2005-014.

Greco, John. 2000. *Putting Skeptics in Their Place: The Nature of Skeptical Arguments and Their Role in Philosophical Inquiry*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511527418.

Husi, Stan. 2013. “Why Reasons Skepticism Is Not Self-Defeating.” *European Journal of Philosophy* 21 (3): 424–49. https://doi.org/10.1111/j.1468-0378.2011.00454.x.

Hyun, Alexander, and Eric Sampson. 2014. “On Believing the Error Theory.” *Journal of Philosophy* 111 (11): 631–40. https://doi.org/10.5840/jphil20141111140.

Jackson, Frank. 2000. *From Metaphysics to Ethics: A Defence of Conceptual Analysis*. Oxford: Clarendon.

Joyce, Richard. 2014. "Taking Moral Skepticism Seriously." *Philosophical Studies* 168 (3): 843–51. https://doi.org/10.1007/s11098-013-0213-8.

Khuramy, Mustafa, and Erik Schulz. 2024. "Normative Error Theory and No Self-Defeat: A Reply to Case." *Philosophia* 52 (1): 135–40. https://doi.org/10.1007/s11406-024-00718-4.

Mackie, John L. 1964. "Self-Refutation—A Formal Analysis." *The Philosophical Quarterly* 14 (56): 193. https://doi.org/10.2307/2955461.

———. 1977/1990. *Ethics: Inventing Right and Wrong*. Reprinted. Penguin Book Philosophy. London: Penguin Books.

Olson, Jonas. 2014. *Moral Error Theory: History, Critique, Defence*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198701934.001.0001.

———. 2016. "On the Defensibility and Believability of Moral Error Theory." *Journal of Moral Philosophy* 13 (4): 461–73. https://doi.org/10.1163/17455243-01304005.

Parfit, Derek. 1997. "Reasons and Motivation." *Aristotelian Society Supplementary Volume* 71 (1): 99–130. https://doi.org/10.1111/1467-8349.00021.

Shafer-Landau, Russ. 2003. *Moral Realism: A Defence*. 1st ed. Oxford University Press. https://doi.org/10.1093/0199259755.001.0001.

Shah, Nishi. 2010. "The Limits of Normative Detachment." *Proceedings of the Aristotelian Society* 110: 347–71. https://doi.org/10.1111/j.1467-9264.2010.00290.x.

Streumer, Bart. 2011. "Are Normative Properties Descriptive Properties?" *Philosophical Studies* 154 (3): 325–48. https://doi.org/10.1007/s11098-010-9534-z.

———. 2013a. "Can We Believe the Error Theory?" *Journal of Philosophy* 110 (4): 194–212. https://doi.org/10.5840/jphil2013110431.

———. 2013b. "Do Normative Judgements Aim to Represent the World?" *Ratio* 26 (4): 450–70. https://doi.org/10.1111/rati.12035.

———. 2016. "Why Jonas Olson Cannot Believe the Error Theory Either." *Journal of Moral Philosophy* 13 (4): 419–36. https://doi.org/10.1163/17455243-01304003.

———. 2017a. *Unbelievable Errors: An Error Theory about All Normative Judgements*. Oxford: Oxford University Press.

———. 2017b. "Why We Really Cannot Believe the Error Theory." In *Moral Skepticism: New Essays*, edited by Diego E. Machuca. London: Routledge.

European Journal of Analytic Philosophy EuJAP | 2024 | Vol. 20 | No. 1 | 157-181
https://doi.org/10.31820/ejap.20.1.7

BETTER-MAKING PROPERTIES AND THE OBJECTIVITY OF VALUE DISAGREEMENT

Erich H. Rast[1]

[1] Nova University Lisbon, Portugal

Original scientific paper – Received: 20/10/2023 Accepted: 09/07/2024

ABSTRACT

A light form of value realism is defended according to which objective properties of comparison objects make value comparisons true or false. If one object has such a better-making property and another lacks it, this is sufficient for the truth of a corresponding value comparison. However, better-making properties are only necessary and usually not sufficient parts of the justifications of value comparisons. The account is not reductionist; it remains consistent with error-theoretic positions and the view that there are normative facts.

Keywords: values; axiology; better than; the good; objectivity; value disagreement.

Correspondence: erast@fcsh.unl.pt

1. Introduction

This article defends a version of value realism, according to which many, if not most, value disagreements are objective and factual. When we rightly value something, it must have one or more distinctive properties that provide reasons why we value it more than other things. Value-based debates frequently revolve around whether or not comparison objects possess such “better-making” properties and which properties fall into this category. Addressing these questions is a factual inquiry.

Unlike metaphysical accounts of value realism like McDowell (1985), the argument presented in this article does not claim that all aspects of our value judgments are objective; the thesis is rather that a substantial part of our value judgments is objective. To “rightly value” is meant in an epistemic, not in a moral sense, in the above formulation. We may call the position defended in this article a light value realism because it remains compatible with the moral skepticism of John Mackie (1977) as well as the moral relativism and contextualism of authors such as Gilbert Harman (1975, 1996), David Wong (1984), and Brit Brogaard (2008, 2012). The objectivity of better-making properties invalidates purely subjectivist takes on value, however, and may therefore serve as a stepping stone towards a more encompassing value realism.

What are values, and what are facts? Providing a definition would be equivalent to solving the fact/value problem, and it is doubtful that this problem has a general solution. Instead, our prior grasp of these notions can serve as a starting point. Ordinary speakers can identify certain adjectives as evaluative. Competent speakers of English, for example, understand that “good” and “brilliant” are evaluative adjectives. The following statements will serve as examples:

(1) Friendship is good.
(2) a. Democracy is good.
b. Democracy is better than oligarchy.
(3) a. This knife is good.
b. Knife *a* is better than knife *b*.
(4) Alice: Chocolate ice cream is better than vanilla ice cream.

Based on prior understanding, we can identify (1), (2), and (3) as value statements. In contrast, I argue in Section 3.1 that the statements in (4) are not value statements, albeit being evaluative in a more general sense. They are based on subjective preferences and do not give rise to direct disagreements about the content of the utterance.

This article assumes that values can be identified with the comparison structure that represents the abstract truth conditions of statements containing the comparative form of a corresponding value predicate. This assumption is prevalent in publications on value structure such as Hansson (2001), Carlson (2018), and Chang (2002). I will also follow Rast (2022a) in presuming that overall value can be calculated by aggregating a finite number of value relations that are regarded sub-values and represent characteristics of the overall value.[1] These relations will be abbreviated by '$\succeq$' for weak betterness reading *x is better than or equal to y*, and corresponding relations '$\succ$' for *x is (strictly) better than y* and '$\sim$' standing for *x and y are equally good*. What counts as good can be defined based on a value relation in this setting, where the exact definition hinges on whether value neutrality is allowed and whether there can be incomparability.[2] In this view, statements like (2a) and (3a) are true or false relative to a more specific value structure that the uses of the comparatives in (2b) and (3b) partly constitute. I will argue in Section 2.2 that examples of intrinsic value attributions such as (1) remain compatible with such a conception of value.

Before continuing, the risk of trivializing the fact/value problem must be mentioned. Cognitivists believe that value statements are either true or non-true (false or lacking a truth-value).[3] If a value statement turns out to be true, it will be true due to a specific fact. So there are trivially only facts in this view. To avoid this deflationary take on the fact/value problem, the following sections focus on "narrow" facts, and a dependence on broader facts will only be addressed in Section 3.2. Narrow facts are either empirical facts, that is, facts that can be confirmed by empirical evidence and are principally testable by experiment, or abstract mathematical and logical truths.

The remainder of this article is structured as follows. Section 2 details better-making properties and briefly addresses Moorean objections. In Section 3, reasons are laid out why better-making properties are objective

[1] In what follows, the term "value" will be used for value relations, sub-values, features contributing to value, and aggregated value. A gain in brevity outweighs the imprecision of this usage, as the details of multidimensional value representations are not relevant for the following arguments.

[2] See Chisholm and Sosa (1966), Dalen (1974), Hansson (1990), Gustafsson (2013, 2015), and Carlson (2014) about "good" in terms of "better than". See Hansson (2018, 512-514) for the opposite direction of deriving value orderings from classificatory value concepts. Rast (2022a, 74–94) provides an overview of value aggregation methods.

[3] According to Oddie (2013), the position may be more aptly named "propositionalism". However, the term "cognitivism" is more common. Importantly, a cognitivist could subscribe to an error-theory, according to which all value statements lack a truth value, but most cognitivists are not error-theorists in that sense.

and constitute sufficient conditions for the truth of value statements although they are typically only necessary and not sufficient parts of their justification. Arguing for this position involves several steps. First, if an apparent disagreement rests on subjective preferences, it does not concern values. Second, Section 3.2 details why better-making properties are objective properties of comparison objects. Having such a property or lacking it are narrow facts. Finally, one might dispute what constitutes such properties and how different betterness judgments ought to be combined. According to Section 3.3, answers to these questions require an evaluation of theories according to their merits. This process is epistemic and the values in it are epistemic values.

2. Better-making properties

A property P is *better-making* for a value ordering $\succeq$ and comparison objects a and b if and only if $P(a)$ & $\neg P(b) \supset a \succ b$.[4] Since the rule that connects the better-making property to the value comparison uses a conditional, it expresses a sufficient condition. If a is not better than b, then a cannot have a better-making property for that value.

Why should anyone accept this rule? Suppose that $a \succ b$, and there is no better-making property. Object a would have no properties that might be cited as to why it is better than b. This position is absurd. A given comparison object must have *some* property that makes it better than another object, whatever that property may be. For example, it would be ludicrous to assert that knife a in (3b) possesses no properties that someone could use to justify why it is better than b. On the contrary, several properties may make it better; it may be sharper than the other, have a better handle than the other, have better steel, and so on. People rarely run out of possible candidates for better-making properties in evaluative practice.

2.1 Complex better-making properties

Multiple features complicate matters. Comparing the knives in (3b), a might turn out to be sharper *and* have a better handle than b, and therefore it may be more suitable than b as a kitchen knife. It is well-known that there are many ways to combine features in such a multidimensional scenario. If the features can be expressed quantitatively, one might sum them up, provided that intuitions about

[4] As a rule for parenthesis elimination, '&' binds stronger than '⊃' in this notation.

overall comparisons remain compensatory and consistent with additive models. For example, a knife with handle quality 3 and blade sharpness 5 must be equal in value to a knife with handle quality 5 and blade sharpness 3 in an additive model. Sometimes, such additive models do not suffice and more complicated aggregation methods are called for. We may put aside many of these details, however, because better-making property are allowed to be arbitrarily complex. Better-making properties are decisive for a comparison if all other relevant comparison features are equal, no matter how complex they are. Some such comparisons may be between hypothetical objects that only differ in one aspect.

Some cases deserve special attention, though. Several properties may be decisive only when they are present together in a sub-additive or a super-additive way. Super-additivity means, in this context, that if those features could be quantified, then their combined presence would have a higher value than the sum of the values of each of the features taken individually. Going back to Moore (1903), this view is often discussed under the label “organic unity”.[5] In contrast, in a sub-additive value combination the combined presence of the features may have a lower value than the sum of the values of each feature taken individually. The holistic assumption behind sub- and super-additive value aggregation can be reformulated as the thesis that the complex better-making property emerges as a qualitatively new property. Claiming that such properties exist ought not pose more problems than the appeal to the holistic assumption.

Better-making properties cannot be contradictory under the same value. If there are two properties, P and P', and two items a and b such that $P(a)$ & $\neg P(b)$ & $\neg P'(a)$ & $P'(b)$, then P and P' cannot be better-making properties belonging to the same value. This constraint is more of a methodological requirement than one concerning value philosophy. Methodologically, it makes sense to specify that conflicting better-making properties belong to separate (sub)values, because methods for aggregating multiple value relations into an overall assessment already allow for dealing with such value conflicts. Otherwise, the underlying value representations would have to be paraconsistent, allowing for the truth of $a \succ b$ & $b \succ a$, rather than the unproblematic case $a \succeq b$ & $b \succeq a$ commonly used to define $a \sim b$. Paraconsistent logics of value can represent moral dilemmas. Still, an account with multiple dimensions has enough expressive power without additional paraconsistency if it allows aggregation failures to represent

[5] See Moore (1903, 28) and Carlson (1997, 2020). Notice that super-additivity can be defined abstractly without assigning numbers to features first.

incomparability. Assume the knife a is better than b in terms of sharpness and b is better than a in handle quality. Various value aggregation algorithms provide solutions to this problem. If the sharpness aspect weighs more than or outranks the handle quality, a might be better than b. If the two values have the same weight or rank, then $a \sim b$ would be an acceptable aggregation. Finally, it is feasible to have two values in a conflict so that aggregating them fails in a specific case.

There are additional technical requirements on the rules for better-making properties. They must generally cohere with the properties of the value relations they indirectly constitute when assembled from piece-wise comparisons. Strict betterness $\succ$ is often considered transitive.[6] If this is the case, then the following rule must hold: For any three objects x, y, z, if there is a better-making property $P1$ that implies $x \succ y$, and there is a better making property $P2$ that implies $y \succ z$, then there is a better-making property $P3$ such that $x \succ z$. Without further ado, the above rule also complies with the irreflexivity of strict betterness since $P(a)$ & $\neg P(a)$ is already excluded as a contradiction when the base logic is not paraconsistent. The standard account of "better than" does not require other rules, but when using nonstandard value relations like semiorders, additional rules must ensure that better-making properties comply with those alternative base relations. For instance, semiorders have the "Ferrer's property".[7]

Finally, we should avoid trivial positions. A better-making property for value comparison $a \succ b$ may not be circular. We should not allow properties whose comprehensive characterization would amount to restating the value comparison in the subsequent of the rule. For instance, this condition prohibits the *property of being better than b*. Although a better-making property can be relational, it may not be relational in the trivial sense of repeating the same or a similar value relation that represents the value under discussion.

2.2 Better-making properties and final value

Better-making properties seem to be hard to square with intrinsic and final value. Since there is widespread agreement in the Moorean tradition of axiology that final value exists, this criticism would at least severely

[6] For counter-arguments to the transitivity of strict betterness, see Temkin (1987, 2012) and Rachels (1998, 2001).

[7] See Luce (1956) and Vincke and Pirlot (1997) for more information about semiorders.

limit the usefulness of the above definition. The purpose of this section is to show that better-making properties are compatible with final value.

Something has a final value when it is valuable for its own sake, without having to take into account other values and consequences of having the value. For example, if friendship in (1) has final value, it is not valuable because having friends provides pleasure or other advantages, it is valuable for its own sake. Some philosophers, such as Korsgaard (1983), consider what is valuable for its own sake final value and oppose it to instrumental value, whereas intrinsic value is opposed to extrinsic value and based on intrinsic properties. This terminology makes final value more important than intrinsic value because there are compelling examples of things with final value not based on an intrinsic property (see Beardsley 1965; O'Neill 1992; Kagan 1998; Rabinowicz and Rønnow-Rasmussen 2000, 2005). Even authors like Zimmerman (2001), who prefers the label "intrinsic", agree that intrinsic value cannot always be based on intrinsic properties of comparison objects in the narrow sense.

For example, according to Beardsley (1965) rare stamps may have a value on their own, and being rare is not an intrinsic property of a stamp. Zimmerman solves this problem by delineating an ontology of states of affairs with basic intrinsic value, but we need not enter the (mostly terminological) debate about intrinsic versus final value. It suffices for current purposes to acknowledge that among arbitrary comparison objects, not all final value is based on intrinsic properties.[8] Likewise, it need not concern us that some authors like Zimmerman (2001) and Perrine (2018) argue that the basic objects of comparisons are states of affairs, whereas others such as Rabinowicz and Rønnow-Rasmussen (2000) argue against this view. The following discussion is neutral about the nature of the comparison objects.

The criticism is this: A better-making property provides the reason why one comparison object is better than another; that is a comparative definition. In contrast, final value does not seem to be comparative at all. To say that friendship in (1) has final value is to say that it is valuable on its own and not relative to other concepts. Hedonists consider pleasure a final value not because it is better than pain but because it is intrinsically good from their point of view. A painting might be valuable in its own right, being so unique that it would be hard even to compare it to other

[8] This is not to say that it is not possible to develop a mereology like Zimmerman's in which the basic value bearers (states of affairs akin to situations) are individuated in just the right way to allow them to have intrinsic value because they have an intrinsic value-providing property. I wish to remain neutral about such mereological approaches in this article.

paintings. Such examples seem to indicate that better-making properties cannot provide a final value and, therefore, cannot be the sole reason why we attribute value in general if final value exists, although they may be useful for reasoning about the instrumental and extrinsic value of objects. As I will argue, this criticism rests on a misunderstanding. Any kind of value, including final value, must allow for comparisons, and better-making properties provide reasons for specific comparisons. There is no incompatibility in the first place.

My counter-argument relies on the choice-guiding nature of values. A necessary, though not sufficient condition for being a value is to potentially guide someone's choices. That is to say, a particular value might never guide anyone's choices in practice, but *if* someone has to choose between several alternatives, then the value must be able to guide the choice provided it is applicable and relevant. I consider this an analytic aspect of what it means to be a value. There are no values that cannot possibly be choice-guiding.[9] The person in need of guidance must somehow be able to apply or use that value to evaluate alternatives and figure out, based on that value, whether one alternative is better than another, they are equally good under that value, they are on a par in the sense of Chang (2002), or the comparison fails for some reason. In all cases except the last one, the properties that provide intrinsic value to a comparison object must play an integral role in the comparison since they are the reasons why these objects have value relative to the other object, and these reasons should guide choices rather than something else.

Thus, when something has a final value, the properties that give it this value must allow for comparisons. When comparing, a better-making property may be identical to the property or relation that lends the comparison object its final value. Nevertheless, the fact that a comparison is made need not be constitutive of the value. For example, suppose that two states of affairs *a* and *b* containing John and Mary are compared. Suppose John and Mary are good friends in *a* and no friends in *b*. If friendship has intrinsic value, then one might say that *a* is better than *b* because *a* has the property of containing two friends that *b* lacks. This property is the better-making property in this example. Despite this, the fact that *a* and *b* are compared is not itself constitutive of the intrinsic value of friendship.

[9] Values must also allow for comparison for reasons not directly related to choices. For example, according to the positivity of goodness, if *a* is good and *b* is better than *a*, then *b* must also be good (Hansson 2018, 509). This principle cannot be formulated without comparisons.

For Moore (1922, 260-261), intrinsic value can come to a specific degree, which trivially enables multiple comparisons. Zimmerman (2001, 159-180) expands on this and even argues that value can be summed up. These fairly strong assumptions about value allow one to use utility functions to represent value. I will address some problems with such representations in Section 3.2 when discussing desire. For now, it suffices to show that better-making properties remain compatible with such views on intrinsic value.

Suppose *a* in the above example has the intrinsic value of friendship to degree 0.8 on a normalized scale between 0 and 1, and *b* has this value to degree 0 because there is no friendship at all in this state of affairs. The better-making property is the property of containing friendship to a normalized degree 0.8 (whatever that means). The same better-making property would also serve as a reason for the comparison to a third state of affairs with two more superficial friends of degree 0.4 only, yielding the judgments $a \succ c \succ b$. The better-making properties include the particular degrees or amounts of the intrinsic value in such cases. Although it is doubtful that such an account of intrinsic value would be adequate for examples like (1) and (2), and one might argue instead that such examples only involve ordinal value comparisons, better-making properties are perfectly compatible with stronger value conceptions according to which intrinsic value comes at a degree.

In summary, better-making properties neither implicitly nor explicitly presume that comparisons are value-constitutive. Value must be able to guide someone's choices under the right circumstances and allow for comparisons, yet the reason why something has value may still be that it has value for its own sake.

3. The role of better-making properties in value disagreement

A better-making property is sufficient for the truth of a "better than" comparison by some value. If object *a* has a better-making property and *b* has not, then *a* is better than *b*. However, the same property cannot make all "better than" comparisons under some value true. If $a \succ b$ and $b \succ c$ hold, then there must be two different better-making properties P and P' such that $P(a)$ & $\neg P(b)$ & $P'(b)$ & $\neg P'(c)$. Hence, better-making properties do not permit a more compact value representation.

The presence of a better-making property in one thing and its absence in another implies an individual value comparison, but this regularity does not necessarily *justify* the comparison. In general, justifications go beyond the mere mention of an isolated condition. Suppose a customer buys a new phone, and battery life is crucial to them. Then, a phone with a battery life of 24 hours is superior to one with an 8-hour battery life, but merely presenting such an attribute as a rationale for the value judgment is likely insufficient. Such a flimsy rationale is only admissible when it is clear that the relevant feature is the most important factor and no other reasons are expected. Generally, justifications need to be more detailed. Why is battery life so critical? How does it relate to other potential better-making properties such as price, camera, and reception quality? How complete a justification needs to be hinges on the context and the goal of the value assessment, but at some point, it must resort to a better-making property. There is no way to argue that *a* is better than *b* without pointing out at least one property of *a* that *b* lacks and that makes *a* better than *b*. A better-making property is a necessary component of justifying a value comparison, though not always sufficient.

Justifications are typically broad and concern all value comparisons by a specific value instead of just one. They can be thought of as theories (in a broad sense) that comparison objects can instantiate. Let $T\,[a, b]$ be the outcome of instantiating such a theory T by objects a and b. For T to be a theory of value $\succeq$, $T\,[a,b]$ must entail the statements $P(a)\ \&\ \neg P(b) \supset a \succ b$ and $P(a)\ \&\ \neg P(b)$ for some better-making property P.

This characterization remains compatible with textbook definitions of necessary and sufficient conditions. According to these definitions, α is a necessary condition for β whenever $\beta \supset \alpha$ holds, and α is a sufficient condition for β whenever $\alpha \supset \beta$ holds. The presence of a better-making property P in a and its absence in b is a necessary condition for the theory to provide a proper justification of the value comparison because $T\,[a, b] \supset P(a)\&\neg P(b)$ holds and, at the same time, it is a sufficient condition for the truth of the value comparison itself since $P(a)\ \&\ \neg P(b) \supset a \succ b$ also holds.

Even when they are relational, better-making properties can be objective. In example (4), the better-making property of chocolate ice cream for Alice is that it tastes like chocolate. Tasting in a particular way is a relation between the object and the taster; thus, the property is relational and the supposed value is agent-relative. The property is also objective, or at the very least, intersubjective. Anyone with a functioning sense of smell will recognize chocolate ice cream. Nevertheless, it is important to note that the justification of an evaluative comparison statement can be subjective even though the better-making property is objective. In this example, Alice may state that ice cream *a* is better than *b* because she prefers chocolate over vanilla taste, whereas Bob may disagree. He prefers the flavor of vanilla to that of chocolate. The taste of the ice cream is mostly objective, but the evaluation of the taste is subjective.[10]

The following sections aim to show that such examples of subjective justification are not the basis of value comparison by arguing for the following theses:

1. If justifications of value comparisons are subjective, we cannot speak of value comparisons. When Alice states that chocolate is better than vanilla ice cream, she ought not be taken literally.
2. Better-making properties are always objective, or at the very least, intersubjective.
3. Because better-making properties provide sufficient conditions for individual value comparisons, many value disagreements concern what constitutes the better-making properties of a value comparison and whether the comparison items have or lack these properties.

3.1 Lack of disagreement about matters of personal taste

This section aims to show that apparent disagreements about personal taste are not value disagreements since they are no disagreements. This idea is not new; it has been discussed quite extensively in recent literature on relativism versus contextualism of predicates of personal taste.

Consider a disagreement in the ice cream scenario. As Lasersohn (2005, 2008) argues, disputes involving uses of predicates of personal taste may be cases of faultless disagreement. Alice might truthfully state (4), and Bob might truthfully state the negation of this sentence. Both assertions may be true, respectively, in relation to the assessors Alice and Bob. According to Lasersohn, in such a case the disagreement is faultless; both of them are right. Other people may also assess the statements in one or the other way in this version of relativism.

[10] As Smith (2007) lays out about wine tasting, "[t]astes are properties a wine has that give rise to certain experiences in us; and they cannot be reduced to, or equated with, those experiences". The circumstances and abilities of the taster need to be appropriate to identify tastes properly, and the possibility of error requires distinguishing more subjective experiences from how things taste. However, there are variations of smelling and tasting abilities among people, so the senses of taste and smell are not *fully* intersubjective. For example, according to a meta-study by Sorokowski et al. (2019), women tend to have better olfaction than men. Training also likely makes a difference. Master perfumers are expected to be able to identify hundreds of notes and accords blindly, a level of expertise laypersons can hardly reach without equivalent training.

It is controversial whether such statements are true relative to an assessor (assessor-relativism) or whether their truth-value varies only because their semantic content varies (contextualism).[11] We do not have to decide on these issues, as both accounts share the same idea: If a comparison is based on preferences of personal taste, it is subjective because people's tastes differ. What is questionable about these cases is whether these cases count as instances of disagreement.[12] As long as Alice in (4) provides as a reason that this is her preference, there need not be any disagreement between Alice and Bob precisely because subjective justifications are deemed appropriate in matters of personal taste. Suppose Bob prefers vanilla over chocolate ice cream. In that case, *his* preference is compatible with Alice's preference, and he can agree with Alice if he agrees that (4) is based on *her* preferences. Strictly speaking, it is incorrect to call such cases subjective disagreements because they are no disagreements in the first place.

This is not to say disagreements over such issues cannot occur at all. A dispute might concern whether someone has a particular preference. Although there is some first-person authority about preferences, this authority is not absolute. Bob may know Alice's preferences better than her. People only sometimes know what they want and can be mistaken or confused about their preferences. Moreover, people may signal disagreement in a conversation, even when there is no disagreement about the underlying subjective aspect of an evaluation. A dispute might concern something else, such as presupposed content or social inferences drawn from the belief that someone has a specific taste. For example, Bob may disagree with Alice because he believes that people who prefer chocolate ice cream over vanilla ice cream are tasteless brutes. As ridiculous as this may sound about ice cream, disputes about musical preferences are often of this sort.[13] It is common in the personal, social, and political realms to have disagreements about something other than the content of a particular utterance the disagreement seems to be about. In these indirect disputes, the utterance content only serves as fuel for other persistent disagreements in the background.

[11] A contextualist might claim that *better than* is a shortcut for *better than for + AGENT*, for example.

[12] This concern was first voiced by Stojanovic (2007), and later refined by Stojanovic (2015) and McNally and Stojanovic (2017). The criticism is also at the heart of Dworkin's "semantic sting" argument in Dworkin (1986).

[13] To mention a famous example (out of many), there were violent clashes between "rockers" and "mods" in Southern England in 1964-66. Cohen (2002) analyzes the media coverage of these incidents and the reactions it caused.

There may also be disagreement over whether the justifications can be subjective. For example, one person may believe that there are objective criteria for determining if one painting is better than another, yet another may be a subjectivist about art. People may also dispute what constitutes a better-making property and whether objects have the property in question. However, once we identify a disagreement as one about taste, we know it will involve primarily subjective justifications. In the other examples mentioned, the disagreement concerns something else, such as social norms and functions. Such additional disagreements may be legitimate, but they are not direct disagreements about the evaluative statement in question. They concern the better-making properties, or a standpoint or social issue hidden behind the evaluative statements seemingly under dispute.

Considering all this, I suggest distinguishing between more broadly conceived evaluative comparisons and value comparisons in the narrow sense. Value comparisons, in the narrow sense, are not based on subjective preferences, although the underlying value relations may look similar to these from a modeling perspective. Value statements are meant to be intersubjective or objective. In contrast, apparent taste disagreements concern evaluative comparisons that reveal subjective preferences, but they involve no disagreement; if there is disagreement, it is not directly about the evaluative statement.

3.2 Better-making properties are objective

In this section, I argue that better-making properties are objective. As previously stated, agent-relative and relational properties can be objective. But what does *objective* mean? Although this question may be hard to answer in general, the following distinctions suffice for the purpose of this article. A subjective property is one that an object can only have if one particular person has a belief or a similar non-factive, truth-upholding attitude about the object and if the property cannot be reduced to a property that does not entail that attitude.[14] In contrast, characterizing an objective property does either not involve any reference to attitudes at all or it involves factive attitudes like knowledge.

A property may also be intersubjective. If a property *P* is such that having *P* presupposes that rational persons within a given community with

[14] We may speak of a truth-upholding attitude whenever an attitude holder takes an embedded proposition more likely to be true than false. For example, certainty and belief are truth-upholding, whereas entertaining a thought and considering a proposition are not.

common knowledge about the world can be expected to hold certain attitudes dispositionally, or upon sincere reflection, about objects that have the property, then P is intersubjective.

To exemplify these distinctions, consider monetary cost. Being believed by Bob to cost $50 is a subjective property. So is being believed by Alice to cost $12. In contrast, the property of costing $50 is an intersubjective property. Monetary systems hinge on people's attitudes about money and its worth, the governing institutions, and markets. In the case of fiat money, those beliefs partially constitute the property of costing $50. Nevertheless, the property of costing $50 is not constituted by any *particular* person's belief about the object, not even the seller's, and therefore is not subjective. Finally, being known by Bob to cost $50 is an objective property because knowledge is factive; everything with this property also has the property of costing $50, which does not require a specific person to hold a belief about it.[15]

Suppose a better-making property P was subjective. According to the definition of a better-making property, $P(a)\&\neg P(b)$ implies the value comparison $a \succ b$. Since a person needs to hold a non-factive attitude about an object for that object to have a subjective property, the rule states in this case that it is a sufficient condition for the truth of a value comparison that a particular person holds an attitude about the object. This position is absurd if the attitude in question is belief or another truth-upholding attitude. The mere fact that someone believes something about a and does not believe the same about b does not warrant that a is better than b; there must be some property in which a and b differ that allows for that conclusion regardless of what a particular person believes about them.

Consider the monetary value of two comparison items, for instance. Just because Alice believes that a is cheaper than b and therefore better in terms of cost does not warrant the conclusion that a is better than b in terms of cost; a is only better than b under this value when it is cheaper. Under normal circumstances, it is not enough for someone to believe that the comparison items have or lack a particular property; they must

[15] Although objective and intersubjective properties need not be mind-independent, they presuppose properties that supervene on mind-independent facts. Such a notion of objectivity evades a recent attack on the inherent value judgments of realism by Dasgupta (2018); see Sider (2022, 196), who does not endorse this notion of objectivity and proposes a metasemantic account instead. However, the debate ranges back to Goodman (1955) and Putnam (1980), and in my opinion a proper response to Dasgupta needs to go back to Putnam's original model-theoretic argument and the role of measurement and combinatorial restrictions imposed by theories, as these theories evolve over time. However, this topic needs to be left for another occasion for lack of space.

actually have the property or lack it. If Alice happens to find out that her belief was false and *b* is cheaper than *a*, she would not say that her values (or, in this case, subjective evaluation) have changed. She would rather say that she misjudged the value of *a* in terms of costs and concede, insofar as she acts rationally, that *b* was better than *a* in terms of costs in the first place.

Only matters of personal taste might be an exception to this rule. Maybe Alice's belief that some ice cream tastes like chocolate is good enough for her evaluation, even if her senses are confused and the ice cream does not actually taste of chocolate. However, as I have argued above, such examples do not illustrate value comparisons because they do not give rise to disagreement. A subjectivist may call these subjective evaluations values, of course. However, this is merely a terminological choice; the point is that subjective evaluations based on personal preferences differ substantially from value comparisons that constitute what one might call real or "genuine" values because the latter give rise to disagreements, whereas the former do not.

Properties involving attitudes that are directly about comparison items fare better. Could the property of being desired by someone be a better-making property? Such an account might seem plausible for Humeans who consider desire a basis for choice. However, there are compelling arguments against the idea that the property of being desired by someone makes something better.

To begin with, being desired does not suffice. To conclude that *a* is better than *b*, the desire for *a* must be greater than the desire for *b*. So degrees or intensities of desire are needed. If these exist, then it is indeed possible to formulate a rule stating that whenever the amount of *X*'s desire for *a* is larger than the amount of *X*'s desire for *b*, then *a* is better than *b* for *X*.

However, such conceptions of "better than" as desire get the direction of justification wrong. We desire *a* more than *b because* it is better (for us, to stay within the agent-relative realm for the sake of argument). The converse is not valid. It is not generally true that whatever we desire more than something else is better (for us).[16] The reason to reject desire as a basis for goodness is not potential psychological confusion, as is sometimes argued against subjectivists, but rather a temporal dimension

[16] Broome (1999, 3) mentions a related principle in terms of preferences, the Preference Satisfaction Principle: the principle that humans always prefer what is better for them. He also considers this principle implausible.

of desire that goodness does not have. We desire something episodically, at a particular time, when the consequences of fulfilling that desire are not yet fully known. If the consequences turn out to be negative in the future, the person still had the desire in the past.

In contrast, suppose we say that something is better than something else for someone. If the consequences turn out to be negative, the initial betterness statement is retracted and considered false. It is not the case that the option for that person was good and is now no longer good; rather, it was bad from the start. This asymmetry in the temporal dimension of the two notions makes it impossible to use desire as a substitute for goodness.

Suppose, for the sake of argument, an account built on a Desire Satisfaction Principle despite these flaws. The resulting position would render value relations obsolete. Utility functions from objects to real numbers can represent an amount of desire that allows for "greater than" comparisons. Desiring *a* more than *b* means that the amount of desire for *a* is greater than the amount of desire for *b*, i.e., $u(a) > u(b)$ holds. According to the theory of scale types introduced by psychologist Stevens (1946) and formally worked out in measurement theory (see, e.g., Roberts 1979; Krantz et al. 1971, 1989, 1990), talking about amounts in this way means that the utility function $u(.)$ rests at least on an interval scale and more likely on a ratio scale.[17] A corresponding value relation can be extracted from such a utility representation in a mechanical way by defining $x \succ y \Leftrightarrow_{Def.} u(x) > u(y)$ and $x \sim y \Leftrightarrow_{Def.} u(x) = u(y)$. This construction makes the value relation dispensable and requires assumptions much stronger than merely talking about "better than" comparisons within a value. Utility functions guarantee that all value comparisons are complete and transitive, provided that additional constraints are met in case there are uncountably many comparison objects. Utility functions also make all value comparisons compensatory, which is a dubious assumption. To cut a long story short, desire understood in this way is a stronger value representation than a mere value relation. It makes the latter redundant.[18]

[17] On an interval scale, any linear transformation $u'(x) = a \cdot u(x) + b$ for positive non-zero constant *a* and positive constant *b* represents the same information as $u(x)$. On a ratio scale, only transformations of the type $u'(x) = a \cdot u(x)$ are allowed for positive non-zero constant *a*, meaning that the 0-point is meaningful and shared. In contrast, an ordinal utility function only represents an underlying preference relation, but talking about amounts of desire would be meaningless on such a scale.

[18] I have argued in Rast (2022a, 2022b) that these utility representations are inadequate for values in general. These arguments are independent of the current point and go beyond the scope of this article.

Even if one is willing to defend such an account, the property of comparison object *a* of being desired to amount $u(a)$ by person *X* cannot serve as a better-making property. The comparison $u(a) > u(b)$, not the amount of desire for *a* itself, makes *a* better than *b*, and this comparison violates the circularity prohibition of Section 2.1. Finally, even under a desire-as-utility view, when we ask why a particular object *a* is better than *b* in a given evaluation situation, the reason cannot just be that it is more desirable. Rather, *a* is more desirable *because* it has some property that *b* lacks. Desire is not blind, something in the desired object needs to spark it.

3.3 The objectivity of value disagreement

To recapitulate, objective better-making properties are sufficient conditions for the truth of value comparisons. These properties are also necessary for justifying value comparisons, so every justification of a value comparison has an objective component. However, one point of the previous sections was that these justifications are typically more exhaustive. Part of a justification may also concern what constitutes a better-making property for a particular value and how different values enter an overall value assessment. Finally, a disagreement may also arise over the relevance of specific values. For example, someone might deny that comparisons of a product's packaging design ought to enter its evaluation. In contrast, someone else might insist that it is an essential aspect of the purchasing experience. Because of these additional possibilities, one might doubt that broader aspects of a justification need to rest on narrow facts.

Moral intuitionists and particularists like Dancy (2004) have expressed one such doubt. According to Dancy, there cannot be an overarching systematic theory that justifies moral judgments and morally relevant value comparisons. Our moral practices are too context-dependent and have too many exceptions to allow for general theories. Instead, we must rely on moral intuitions in each evaluative context. These enter broader justifications of value statements.

It is worth noting, however, that moral intuitionism and particularism are compatible with the approach presented thus far. Sometimes a justification may appeal to intuitions, and it is also possible to have different justifications in different contexts. Nevertheless, it seems doubtful that intuitions alone can be decisive for particular value disagreements.

The problem is that intuitions are not generally a source of evidence. I follow Hintikka (1999) in this regard, though my own take is a bit less radical. In my point of view, intuitions may provide evidence in moral philosophy due to certain anthropological constants, but I agree with Hintikka that they are methodologically useless for resolving disagreements. Suppose most people share roughly the same intuitions about a value statement. That means the value statement is uncontroversial, and most people agree about it. In that case, there is no demand for a justification, and there will be widespread agreement over the better-making properties. Such cases may occur, but they are of little interest in the light of error-theoretic arguments like those of Mackie (1977). Many interesting value statements trigger persistent disagreements. So suppose there are conflicting intuitions instead. Then intuitions themselves cannot resolve a disagreement, although they might help to address it. There are essentially three ways to deal with such cases:

1. One might deal with them like in the ice cream example. The result is moral relativism.
2. One might claim that some people have mistaken intuitions or misidentify them. This leads to moral skepticism and an error theory.
3. Justifications may involve something else besides intuitions, such as moral and narrow facts.

According to the thesis defended in Section 3.1, the first response means that the alleged value statement does not concern value but only subjective preference. There is no fundamental disagreement between people who seemingly disagree about such statements, or the disagreement is about something else. The second response is likewise possible. However, it is a long stretch to claim these are the only possibilities. At least *some* value comparisons can reasonably be expected to fall into the third category. So what about the third case?

Factual disagreements can be persistent, and their resolution may require detailed domain knowledge. Nobody would expect non-specialists to be able to determine whether a statement in physics is reasonably well-confirmed or false; physicists do that, and they need to study physics for years to acquire the skills to judge and advance physical theories. Similarly, problems of what constitutes better-making properties and how to combine different values into an overall assessment might hinge on moral facts. Scanlon (2014) and Parfit (2011) defend moral facts based on “domain pluralism”, the thesis that the truth of statements and the existence of corresponding facts are established differently by different domains of inquiry. Science is concerned with narrow facts, mathematical reasoning is concerned with mathematical facts and the existence of mathematical objects, moral reasoning is concerned with moral facts, and so forth. If this view is correct, moral and axiological facts might make the theories that support value statements true or at least more adequate than other theories. Some of these facts might not be narrow in the sense introduced in Section 1.

Domain pluralism is controversial. What would these non-narrow facts be, and how do we access them? Are they like mathematical truths? This article does not need to answer these questions and decide whether domain pluralism is acceptable. Whether moral facts exist is independent of justifying and ranking the overall merits of theories that support value comparisons. Error theory, moral relativism, naturalism, non-cognitivism, and moral realism have one thing in common: Theories are *not* compared according to their moral value. We compare them according to how close we believe they are to being true, and, in a more practical sense, according to theory virtues and merits exemplifying (broadly conceived) epistemic value. There is no reason to believe that axiological theories work substantially differently than theories in other fields from an epistemological point of view. A justification of a value statement rests on a supporting theory and corresponding beliefs, which may include metaethical and normative stances, and any theory is ultimately assessed on the basis of its overall merits. Epistemic values decide the outcome of such an evaluation. Which justifications and supporting theories are most likely true? Which justification has best explanatory adequacy? Which one integrates best with other value-related issues and metaethical theories? Which one is internally most coherent? Justifying a value comparison requires answering these questions, which cannot be answered by intuitions alone.

So, the answer to the question of how to deal with questions of the third kind is that, ultimately, the epistemic merits of supporting theories decide between competing justifications of value statements. It is a separate question whether those merits reliably track moral facts and in which way, and it seems likely that viable answers to these questions vary from value to value. Different types of values have different supporting theories with different overall merits, and we need to address each of them separately.

Is this the trivial position mentioned at the beginning of this article? Although it remains close to it, the new position is no longer trivial. First, better-making properties are not trivial, and whether a comparison object has better-making property or lacks it depends on narrow facts. This aspect of value comparisons is objective. Value comparisons between hypothetical comparison objects are equally objective. In this case, law-like statements from well-confirmed theories allow us to derive the relevant facts about the comparison objects. For instance, to have any validity, causal consequences of hypothetical courses of action that give rise to better-making properties are based on law-like statements about the world, and the theories supporting these statements are empirical. Second, ranking theories according to their overall merits is far from being trivial, as the vast body of literature on abduction and inference to the best explanation illustrates.[19] The epistemic evaluation involved in inference to the best explanation does not involve moral value. Even when non-narrow facts are involved in this evaluation, epistemic values trump other types of value and ultimately guide our judgments about value statements. All aspects of value disagreement are objective in this sense.

This position remains compatible with the view that there is sometimes no acceptable justification for a particular type of value statement. Judging that there is no acceptable justification is itself an evaluative position, though one that might remain agnostic about the original value statement. In that case, the proper response acknowledges that there is no corresponding value. This response is similar to how we (epistemically should) deal with existence claims in other domains of inquiry. For example, as Russell (1952) famously pointed out in his rejection of theism, the claim that there is a teapot flying in orbit between Earth and Mars has no good enough justification, so the default assumption ought to be that there is no such teapot. Likewise, if there is no good enough justification for a value statement, the default assumption is that there is no underlying value.

4. Conclusion

The above arguments support the thesis that value disagreements are disagreements about facts but do not say anything about the existence of such facts in a particular case. That is the right kind of theory because it matches how we deal with alleged facts in other domains. We rank theories and justifications according to their overall merits, and this evaluative process rests on epistemic values and theory virtues. So, the

[19] See, among many others, Peirce (1955), G. H. Harman (1965), Hintikka (1999), Magnani (2001), Lipton (2004), Gabbay and Woods (2005), Minnameier (2004), Schurz (2008), Mohammadian (2021), McCain and Poston (2017), and Niiniluoto (2018).

conclusion of this article is that value disagreements are objective and rest on epistemic values, provided there is a value behind them. In contrast, seemingly subjective value disagreements are no value disagreements because they are no disagreements.

Acknowledgments

This work was supported by Fundação para a Ciência e Tecnologia, project DL 57/2016/CP1453/CT0002, DOI 10.54499/DL57/2016/CP1453/CT0002, as well as project UIDB/00183/2020, DOI 10.54499/UIDB/00183/2020. Many thanks go to several anonymous reviewers and to the participants of the Open Seminar at the Nova Institute of Philosophy for helpful comments on an earlier version of this article.

REFERENCES

Beardsley, Monroe C. 1965. "Intrinsic Value." *Philosophy and Phenomenological Research* 26 (1): 0031-8205. https://doi.org/10.2307/2105465.

Brogaard, Berit. 2008. "Moral Contextualism and Moral Relativism." *The Philosophical Quarterly* 58 (232): 385–409. https://doi.org/10.1111/j.1467-9213.2007.543.x.

———. 2012. "Moral Relativism and Moral Expressivism." *The Southern Journal of Philosophy* 50 (4): 538–556. https://doi.org/10.1111/j.2041-6962.2012.00141.x.

Broome, John. 1999. *Ethics out of Economics.* Cambridge: Cambridge University Press. https://doi.org/10.1017/cbo9780511605888.

Carlson, Erik. 1997. "A Note on Moore's Organic Unities." *The Journal of Value Inquiry* 31 (1): 55–59. https://doi.org/10.1023/A:1004282127888.

———. 2014. "'Good' in Terms of 'Better'." *Noûs* 50 (1): 213–223. https://doi.org/10.1111/nous.12061.

———. 2018. "Value Theory (Axiology)." In *Introduction to Formal Philosophy,* edited by Sven Ove Hansson and Vincent F. Hendricks, 523–534. Springer. https://doi.org/10.1007/978-3-319-77434-3_28.

———. 2020. "Organic Unities and Conditionalism About Final Value." *The Journal of Value Inquiry* 54 (2): 175–181. https://doi.org/10.1007/s10790-019-09688-3.

Chang, Ruth. 2002. "The Possibility of Parity." *Ethics* 112: 669–688. https://doi.org/10.1086/339673.

Chisholm, Roderick, and Ernest Sosa. 1966. "On the Logic of 'Intrinsically Better'." *American Philosophical Quarterly* 3 (3): 244–249.

Cohen, Stanley. 2002. *Folk Devils and Moral Panics.* Routledge. https://doi.org/10.4324/9780203828250.

van Dalen, Dirk. 1974. "Variants of Rescher's Semantics for Preference Logic and Some Completeness Theorems." *Studia Logica* 33 (2): 163–181. https://doi.org/10.1007/bf02120492.

Dancy, Jonathan. 2004. *Ethics without Principles.* Oxford: Clarendon Press. https://doi.org/10.1093/0199270023.001.0001.

Dasgupta, Shamik. 2018. "Realism and the Absence of Value." *The Philosophical Review* 127 (3): 279–322. https://doi.org/10.1215/00318108-6718771.

Dworkin, Ronald. 1986. *Law's Empire.* Cambridge, Mass.: Harvard University Press.

Gabbay, Dov, and John Woods. 2005. *The Reach of Abduction.* Amsterdam: Elsevier.

Goodman, Nelson. 1955. *Fact, Fiction, and Forecast.* Cambridge, MA: Harvard University Press.

Gustafsson, Johan E. 2013. "Value-Preference Symmetry and Fitting Attitude Accounts of Value Relations." *The Philosophical Quarterly* 63 (252): 476–491. https://doi.org/10.1111/1467-9213.12025.

———. 2015. "Still Not 'Good' in Terms of 'Better'." *Noûs* 50 (4): 854–864. https://doi.org/10.1111/nous.12122.

Hansson, Sven Ove. 1990. "Defining 'Good' and 'Bad' in Terms of 'Better'." *Notre Dame Journal of Formal Logic* 31 (1): 136–149. https://doi.org/10.1305/ndjfl/1093635338.

———. 2001. *The Structure of Values and Norms.* Cambridge: Cambridge University Press. https://doi.org/10.1017/cbo9780511498466.

———. 2018. "Formal Investigations of Value." In *Introduction to Formal Philosophy*, edited by Sven Ove Hansson and Vincent F. Hendricks, 499–534. Cham, Switzerland: Springer. https://doi.org/10.1007/978-3-319-77434-3_27.

Harman, Gilbert H. 1965. "The Inference to the Best Explanation." *The Philosophical Review*, 74 (1): 88–95. https://doi.org/10.2307/2183532.

———. 1975. "Moral Relativism Defended." *Philosophical Review* 84:3–23. https://doi.org/10.1093/0198238045.003.0001.

———. 1996. "Moral Relativism." In *Moral Relativism and Moral Objectivity,* edited by Gilbert Harman and Judith Jarvis Thomson, 3–64. Cambridge: Blackwell Publishers.

Hintikka, Jaakko. 1999. "The Emperor's New Intuitions." *Journal of Philosophy*, 96 (3): 127–147. https://doi.org/10.5840/jphil199996331.

Kagan, Shelly. 1998. "Rethinking Intrinsic Value." *The Journal of Ethics* 2 (4): 277–297. https://doi.org/10.1023/a:1009782403793.

Korsgaard, Christine M. 1983. "Two Distinctions in Goodness." *The Philosophical Review* 92 (2): 169. https://doi.org/10.2307/2184924.

Krantz, David H., R. Duncan Luce, Patrick Suppes, and Amos Tversky. 1971/1989/1990. *Foundations of Measurement, Volumes I-III.* New York: Academic Press.

Lasersohn, Peter. 2005. "Context Dependence, Disagreement, and Predicates of Personal Taste." *Linguistics and Philosophy* 28 (6): 643–686. https://doi.org/10.1007/s10988-005-0596-x.

———. 2008. "Quantification and Perspective in Relativist Semantics." *Philosophical Perspectives* 22 (1): 305–337. https://doi.org/10.1111/j.1520-8583.2008.00150.x.

Lipton, P. 2004. *Inference to the Best Explanation.* In *A Companion to the Philosophy of Science*, edited by W. H. Newton-Smith, 184–193. Wiley. https://doi.org/10.1002/9781405164481.ch29.

Luce, Duncan R. 1956. "Semiorders and a Theory of Utility Discrimination." *Econometrica* 24 (2): 178–191. https://doi.org/10.2307/1905751.

Mackie, John L. 1977. *Ethics: Inventing Right and Wrong.* London: Penguin.

Magnani, Lorenzo. 2001. *Abduction, Reason, and Science: Processes of Discovery and Explanation.* Kluwer Academic Publishers.

McCain, Kevin, and Ted Poston, eds. 2017. *Best Explanations: New Essays on Inference to the Best Explanation.* Oxford: Oxford University Press.

McDowell, John. 1985. "Values and Secondary Qualities." In *Morality and Objectivity*, edited by Ted Honderich, 110–129. London: Routledge.

McNally, Louise, and Isidora Stojanovic. 2017. "Aesthetic Adjectives." In *The Semantics of Aesthetic Judgment*, edited by James Young, 17–37. Oxford: Oxford University Press.

Minnameier, Gerhard. 2004. "Peirce-Suit of Truth – Why Inference to the Best Explanation and Abduction Ought Not to be Confused." *Erkenntnis* 60 (1): 75–105. https://doi.org/10.1023/b:erke.0000005162.52052.7f.

Mohammadian, Mousa. 2021. "Abduction - the Context of Discovery + Underdetermination = Inference to the Best Explanation."

Synthese 198: 4205–4228. https://doi.org/10.1007/s11229-019-02337-z.

Moore, George Edward. 1903. *Principia Ethica.* Cambridge: Cambridge University Press.

———. 1922. *Principia Ethica.* 2 (reprinted). Cambridge: Cambridge University Press. https://doi.org/10.2307/j.ctv1jk0jrs.22.

Niiniluoto, Ilkka. 2018. *Truth-Seeking by Abduction.* Cham, Switzerland: Springer. https://doi.org/10.1007/978-3-319-99157-3.

O'Neill, John. 1992. "The Varieties of Intrinsic Value." *Monist* 75 (2): 119–137. https://doi.org/10.5840/monist19927527.

Oddie, Graham. 2013. *Value Realism*. In *The International Encyclopedia of Ethics*, edited by Hugh LaFollette. Wiley-Blackwell. https://doi.org/10.1002/9781444367072.wbiee588.

Parfit, Derek. 2011. *On What Matters.* Vol. 1. New York: Oxford University Press. https://doi.org/10.1093/acprof:osobl/9780199572809.001.0001.

Peirce, Charles Sanders. 1955. "Abduction and Induction." In *Philosophical Writings of Peirce,* edited by Justus Buchler, 150–156. Dover.

Perrine, Timothy. 2018. "Basic Final Value and Zimmerman's The Nature of Intrinsic Value." *Ethical Theory and Moral Practice* 21 (4): 979–996. https://doi.org/10.1007/s10677-018-9938-y.

Putnam, Hilary. 1980. "Models and Reality." *Journal of Symbolic Logic* 45 (3): 464–482. https://doi.org/10.2307/2273415.

Rabinowicz, Wlodek, and Toni Rønnow-Rasmussen. 2000. "A Distinction in Value: Intrinsic and for its Own Sake." *Proceedings of the Aristotelian Society* 100 (1): 33–51. https://doi.org/10.1111/j.0066-7372.2003.00002.x.

———. 2005. "Tropic of Value." In *Recent Work on Intrinsic Value,* edited by Toni Rønnow-Rasmussen and Michael J. Zimmerman, 213–226. Springer.

Rachels, Stuart. 1998. "Counterexamples to the Transitivity of Better Than." *Australasian Journal of Philosophy* 76 (1): 71–83. https://doi.org/10.1007/1-4020-3846-1_20.

———. 2001. "A Set of Solutions to Parfit's Problems." *Nous* 35 (2): 214–238. https://doi.org/10.1111/0029-4624.00294.

Rast, Erich. 2022a. *Theory of Value Structure.* London: Lexington Books.

———. 2022b. "The Multidimensional Structure of 'Better Than'." *Axiomathes,* 32 (2): 291–319.

Roberts, Fred. S. 1979. *Measurement Theory.* Reading, MA: Adison Wesley. https://doi.org/10.1017/cbo9780511759871.

Russell, Bertrand. 1952. "Is There a God?" Commissioned but never published. *Illustrated Magazine.*

Scanlon, Thomas M. 2014. *Being Realistic about Reasons.* Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199678488.001.0001.

Schurz, Gerhard. 2008. “Patterns of Abduction.” *Synthese* 164 (2): 201–234. https://doi.org/10.1007/s11229-007-9223-4.

Sider, Theodore. 2022. “Dasgupta’s Detonation.” *Philosophical Perspectives* 36 (1): 292–304. https://doi.org/10.1111/phpe.12169.

Smith, Barry C. 2007. “The Objectivity of Taste and Tasting.” In *Questions of Taste: The Philosophy of Wine,* edited by Barry C. Smith, 41–77. Oxford: Oxford University Press.

Sorokowski, Piotr, Maciej Karwowski, Michał Misiak, Michalina Konstancja Marczak, Martyna Dziekan, Thomas Hummel, and Agnieszka Sorokowska. 2019. “Sex Differences in Human Olfaction: A Meta-Analysis.” *Frontiers in Psychology* 10 (February). https://doi.org/10.3389/fpsyg.2019.00242.

Stevens, Stanley Smith. 1946. “On the Theory of Scales of Measurement.” *Science* 103 (2684): 677–680. https://doi.org/10.1126/science.103.2684.677.

Stojanovic, Isidora. 2007. “Talking about Taste: Disagreement, Implicit Arguments, and Relative Truth.” *Linguistics and Philosophy* 30 (6): 691–706. https://doi.org/10.1007/s10988-008-9030-5.

———. 2015. “Evaluative Adjectives and Evaluative Uses of Ordinary Adjectives.” In *Proceedings of LENLS12: Language Engineering and Natural Language Semantics,* edited by Daisuke Bekki and Eric McCready. The Japan Society for Artificial Intelligence.

Temkin, Larry S. 1987. “Intransitivity and the Mere Addition Paradox.” *Philosophy & Public Affairs* 16 (2): 138–187.

———. 2012. *Rethinking the Good: Moral Ideals and the Nature of Practical Reasoning.* New York: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199759446.001.0001.

Vincke, P., and Marc Pirlot. 1997. *Semiorders: Properties, Representations, Applications.* Dordrecht: Springer.

Wong, David B. 1984. *Moral Relativity.* Berkeley: University of California Press. https://doi.org/10.1525/9780520335028.

Zimmerman, Michael J. 2001. *The Nature of Intrinsic Value.* Rowman & Littlefield.

European Journal of Analytic Philosophy EuJAP | 2024 | Vol. 20 | No. 1 | 183-207
https://doi.org/10.31820/ejap.20.1.8

INTEGRATIVE BIOETHICS: A BLIND ALLEY OF EUROPEAN BIOETHICS

Tomislav Bracanović[1]

[1] Institute of Philosophy, Croatia

Review paper – Received: 17/03/2024 Accepted: 22/07/2024

ABSTRACT

Integrative bioethics is a predominantly Croatian school of thought whose proponents claim to have initiated an innovative and recognizably European concept of bioethics capable of dealing with the most pressing issues of our time. In this paper, a critical overview of the integrative bioethics project is undertaken to show that it is, in fact, a poorly articulated and arguably pseudoscientific enterprise fundamentally incapable of dealing with practical challenges. The first section provides the basic outline of integrative bioethics: its historical development, major proponents, geographical context and philosophical foundations. The second section considers its main theoretical shortcomings: the absence of normativity, collapse into ethical relativism and frequent intratheoretical inconsistencies. The third section addresses the issue of typically pseudoscientific features of integrative bioethics: verbose language, constant self-glorification and isolation from mainstream science. The fourth and concluding section of the paper argues that integrative bioethics—regarding its quality, reception and identity—does not merit the “European bioethics” label and is better described as a blind alley of European bioethics.

Keywords: integrative bioethics; pluriperspectivism; inconsistency; ethical relativism; pseudoscience; European bioethics.

Correspondence: tbracanovic@ifzg.hr

1. Introduction

Integrative bioethics is a predominantly Croatian school of thought founded at the end of the 20th century whose proponents claim to have initiated an innovative and recognizably European concept of bioethics capable of dealing with the most pressing issues of our time. By relying on already existing criticisms of integrative bioethics (Bracanović 2012; Ivanković and Savić 2016; Savić and Ivanković 2017) and by taking into account its proponents' more recent publications, this paper aims to show that actually the opposite is true: that integrative bioethics is a poorly articulated and arguably pseudoscientific enterprise that is fundamentally incapable of dealing with bioethical challenges and as such does not merit the "European bioethics" label.

The paper has four sections. Its second section is purely descriptive and provides the basic outline of integrative bioethics: its historical development, major proponents, geographical context and philosophical foundations. The third section is a criticism focused on three shortcomings of integrative bioethics: the absence of normativity, inevitable collapse into ethical relativism and frequent inconsistencies. The fourth section addresses the issue of a large number of typically pseudoscientific features of integrative bioethics: verbose language, constant self-glorification and isolation from mainstream science. Based on preceding considerations, the fifth and concluding section of the paper argues that integrative bioethics cannot be considered European bioethics when it comes to its quality, reception or identity.

2. Integrative bioethics: History, geography and philosophy

Integrative bioethics is a predominantly Croatian brand of bioethics established at the end of the 20th century.[1] Its development is usually associated with the period when Ante Čović—the founding father of integrative bioethics, formerly ethics professor at the Department of Philosophy of the Faculty of Humanities and Social Sciences in Zagreb—initiated and led three research projects with the financial support of the Croatian Ministry of Science: *Bioethics and Philosophy* (1996-2002),

[1] The following outline of integrative bioethics is partly based on the document *Koncept i projekt integrativne bioetike* published on the Centre of Excellence for Integrative Bioethics website, https://www.bioetika.hr/wp-content/uploads/2016/02/ZCI-IB-koncept-i-projekt.pdf. It also draws on the booklet *Zehn Jahre Integrative Bioethik an der Fern Universität in Hagen 2009-2019*, available at https://www.fernuni-hagen.de/bioethik/docs/10_jahre_integrative_bioethik.pdf (both websites accessed August 4, 2024).

Bioethics and Philosophy (2002-2006) and the *Foundations of Integrative Bioethics* (2007-2011). The international expansion of integrative bioethics began in 2004 when the circle of scholars gathered around Čović's projects connected with the circle of scholars associated with the project *Nutzenkultur versus Normenkultur: Zu den intrakulturellen Differenzen in der westlichen Bioethik*, led by Walter Schweidler at the Ruhr University in Bochum (Germany). In the ensuing years, the two groups organized seven conferences on bioethics in Southeast Europe (Croatia, Bosnia and Herzegovina, and Serbia) and several international summer schools on integrative bioethics (Croatia, Germany, Greece, and Bulgaria). The most important event of integrative bioethics is the annual conference *Lošinj Days of Bioethics*, held in Mali Lošinj in Croatia for over twenty years.

The development of integrative bioethics is also reflected in the growth of the number of its centers in Croatia: Referral Centre for Bioethics in Southeast Europe (founded in Zagreb in 2006), Documentation and Research Centre for European Bioethics "Fritz Jahr" of the University of Rijeka, Centre for Integrative Bioethics of the Faculty of Philosophy in Zagreb, Centre for Integrative Bioethics of the Faculty of Philosophy in Split, Centre for Integrative Bioethics of the J. J. Strossmayer University in Osijek (all founded in 2013) and the Centre of Excellence for Integrative Bioethics (founded in Zagreb in 2014).

The establishment of the Centre of Excellence for Integrative Bioethics resulted from Čović's 2012-2013 project (funded by the University of Zagreb) *Integrative Bioethics: Developing the Centre of Excellence and the Doctoral Program at the University of Zagreb*. Although the doctoral program in Zagreb was not established, similar programs were launched in other European cities: in Sofia (Bulgaria), there is an MA program called "Integrative bioethics"; at the distance-learning university in Hagen (Germany) exists a module (encompassing a number of courses, lectures and summer schools) in integrative bioethics, and the University of Crete (Greece) runs an MA and Ph.D. program that "operates according to the integrative-bioethical foundations".[2]

The history of integrative bioethics is also the history of its publishing projects. Its vital publication hubs are the two journals of the Croatian Philosophical Society, *Filozofska istraživanja*, published since 1980 in Croatian language, and *Synthesis Philosophica*, published since 1986 in several foreign languages. Although neither of these journals initially

[2] See *Koncept i projekt integrativne bioetike*, 9.

specialized in “bioethical” or “integrative” issues, they progressively opened their pages to such topics since the mid-1990s, especially as the proponents of integrative bioethics assumed editorial positions. Moreover, after 2006, when Čović was appointed editor-in-chief of both journals for the second time, both journals were officially proclaimed the “journals for integrative thought”.[3] As for other publishing projects, a key role is played by the publishing house Pergamena from Zagreb. Since 1997, when Čović established and became the editor of its “Bioethics” series, it has published almost 50 books and collections of papers on various bioethical topics (dominated, of course, by authors of an integrative-bioethical orientation). Abroad, the publishing house Academia Verlag from Sankt Augustin in Germany published, from 2005 to 2014, six collections of papers devoted primarily to integrative bioethics topics and issues. The Bioethics Society of Bosnia and Herzegovina also (between 2007 and 2012) published three collections of papers dedicated to various questions of integrative bioethics.

As its proponents tell us, the place of integrative bioethics in the global development of bioethics is unique and essential. Bioethics in the 20th and 21st centuries, according to Čović (2011), had three developmental stages. The first stage was “new medical ethics”, focused on moral reflection about issues arising within healthcare and biomedical research. The central work of this developmental stage of bioethics was the *Principles of Biomedical Ethics* by Beauchamp and Childress (^{1}1979, 2013). The second stage was “global bioethics”, making a turn towards “ethical pluralism” and “scientific interdisciplinarity”, as well as towards a much broader scope of problems related to life and its social, political, and ecological context. The central work of this stage was *Global Bioethics: Building on the Leopold Legacy* by Potter (1988). The third stage, according to Čović, is his own “integrative bioethics”, in which methodological turn was made not only to “ethical pluralism” but also to “pluriperspectivism”. The scope of integrative bioethics encompasses not

[3] Čović was editor-in-chief of both journals in two terms: first time from 1984 to 1993 and second time from 2006 until today. No papers on bioethics were systematically published in *Filozofska istraživanja* and *Synthesis Philosophica* before the mid-1990s. Čović himself published many papers in both journals during the 1980s, but none were about bioethics (to be precise, all his papers then were about Marx and Marxism). Papers from Čović’s Marxist period (1974-1988) are reprinted in his book *Marxism as the Philosophy of the World* (1988). Although bioethics was in full swing already during the 1970s (when significant works by Aldo Leopold, Hans Jonas, Van Rensselaer Potter, Tom Beauchamp and James Childress were published), the “Marxist” Čović seems to have been entirely disinterested for (or unaware of) it. The “bioethical” Čović, born around the mid-1990s, (re)discovered bioethics and all these authors. This transformation from “Marxism” to “bioethics” is consistent with Ana Borovečki’s (2014, 1049-50) assessment that in 1990s Croatia, “the impetus for the developments in the field of bioethics were the changes in the political system”, prompting a large number of former professors of subjects like Marxism to “reinvent” themselves as bioethicists.

only issues related to healthcare and biomedical research (as was the case with “new medical ethics”) or to issues related to life and its social, political, and ecological context (as was the case with “global bioethics”), but also

> (…) the philosophical-historical dimension, in which the character of the scientific-technical epoch and the role of modern science are illuminated, the changes in the fundamental relations of man to what is historically given are considered, and the processes of refraction of world-historical epochs are detected. (Čović 2011, 20-21)

In addition to implicit reference to his publications, Čović singles out Jurić’s (2007) paper on Potterian “roots” or “footholds” of integrative bioethics as one of the most important works of this “integrative” stage of bioethics.

The question is, of course, what makes integrative bioethics so unique compared to its alternatives? Its proponents’ answer is the following: Integrative bioethics is a response to the “misuse of scientific results that can cause irreversible and catastrophic consequences for man and life as a whole” (Čović 2004, 164) but also to “bioethical reductionism” or the one-sidedness of other bioethical traditions.[4] The scope of problems they plan to address is very broad and they define their bioethics as

> (…) an open area for the encounter and the dialogue between different sciences and activities, as well as for different approaches and worldviews, which is meant to articulate, discuss and resolve ethical questions related to life, to life as a whole and to all parts of that whole, to life in all its forms, stages, phases and appearances. (Jurić 2007, 83)

The methodological principles of integrative bioethics are best presented through their “official” definitions: (1) *multidisciplinarity* (the gathering of “all human sciences and activities that are relevant for bioethical questions”), (2) *interdisciplinarity* (to “encourage dialogue and to find a mode of cooperation between all these disciplines”), (3) *transdisciplinarity* (to “overcome mutual differences” and unify them “into a unique bioethical view focused on questions that cannot be unraveled from the perspective of *one* science or *one* area”), (4) *pluriperspectivity* (meaning “unification and dialogical mediation of not

[4] A relatively recent paper in English about their main tenets is Čović and Jurić (2018).

only scientific, but also of non-scientific, i.e. a-scientific contributions", such as "diverse ways of reflection, diverse traditions of thought and cultural traditions, that is, diverse views that rest on cultural, religious, political and other particularities"), and (5) *integrativity* (gathering "all the abovementioned differences into a unique *bioethical view*, rather than into a disciplinary and disciplined scientific framework") (all quotations are from Jurić 2007, 84-5). This set of principles should create "footholds and standards for orientation when it comes to questions about life or about conditions and circumstances of its preservation" (Čović 2004, 11).[5]

Integrative bioethicists very often describe their position in laudatory "European" terms as the project that "Europeanizes bioethics" by "regenerating the spiritual potential of the European philosophical heritage" (Čović 2005, 12), as the "developmental shift" that "transferred bioethics from the United States to Europe" (Čović 2011, 21), as "original and foundational concept of the European bioethics" (Čović 2023, 14), as bioethics that "transcended the imagined framework of South and Southeast Europe" and "encompassed the entire European context" (Pavić 2014, 583), as "innovative and recognizably European concept of bioethics" (Tomašević 2013, 494) and as the "striking development of the bioethical discipline in Central and Southeastern Europe in the last thirty years" (Perušić 2019, 323). As will be shown here, none of these descriptions is justified.

3. Integrative bioethics: Problems with normativity, relativism and consistency

What qualities should a new and unique bioethical theory have if it hopes to deal with pressing issues caused by the development of science and technology? A minimal set of such qualities would undoubtedly include a specific set of normative principles for resolving moral conflicts, the internal consistency between its essential parts and a clearly defined scope of problems it wants to address. By relying on objections to integrative bioethics developed in Bracanović (2012), Ivanković and

[5] The word "orientation" is carefully chosen here. Relying on philosophers like Jürgen Mitelstraß, Friedrich Kaulbach, and Werner Stegmeier, integrative bioethicists present their enterprise as "orientational science" in pursuit of "orientational knowledge" (see Čović 2006, 2009; Cifrić 2006; Jurić 2007; Pavić 2014; Perušić 2019). Orientational knowledge is the "knowledge about the goals for which scientific knowledge will be applied, and for which it will never be aplied", or the knowledge that "guides a person as to the way and the limits of the application of scientific knowledge" (Cifrić 2006, 298).

Savić (2016), and Savić and Ivanković (2017), as well as by analyzing some integrative bioethicists' more recent publications, I will try to corroborate the view that all these qualities are conspicuously absent from integrative bioethics.

A severe problem of integrative bioethics is its lack of normativity or action-guiding capacity. This can be summarized as follows: (1) integrative bioethicists are right to highlight the moral threats posed by scientific and technological advancements, (2) they are right to emphasize that dealing with these threats requires considering all relevant perspectives, but (3) they do not deliver when it comes to providing their unique account as to how one should decide between mutually exclusive perspectives when facing particular bioethical dilemmas. Since finding solutions to such dilemmas is the *raison d'être* of the entire field, integrative bioethics fails in the most critical mission: telling us how to choose among diverse perspectives and arrive at the morally correct answers. Collecting the opinions of all affected parties in various bioethical dilemmas is praiseworthy, but this is typically done by descriptive sciences like sociology or psychology researching, for example, the public opinion on issues such as healthcare, preservation of the environment, animal rights, etc. The normative or action-guiding principles that would distinguish integrative bioethics *qua* bioethics are simply absent from its agenda. This absence of normativity is a severe problem, especially as integrative bioethicists emphasize that they do not wish merely to "articulate" but also to "*resolve ethical questions* related to life" (Jurić 2007, 83, emphasis added). A convenient illustration of this problem can be provided via Katinić's "round table" account of integrative bioethics:

> Figuratively speaking, integrative bioethics is conceived as a huge round table where experts of different profiles and representatives of different domains of social life sit and in a lively and fruitful discussion find the best solutions to complex and difficult problems such as the treatment of newly conceived human beings, transhumanist theories and practices, genetically modified organisms, energy crisis, etc. (Katinić 2012, 599)

Assume that participants in this integrative round table represent conservative and liberal worldviews, respectively, discussing the permissibility of abortion. Both sides will probably be prepared to listen to (maybe even agree with) the scientific theories about the development of the fetus and its characteristics. Will this, however, eliminate the fundamental moral disagreement between them? Hardly. For conservatives, even the early embryo will have that one additional (normative) property that science, by definition, cannot address: the absolute right not to be destroyed, which is comparable to the right of an adult person. For liberals, even a relatively mature fetus will lack that one additional (normative) property that science, by definition, cannot address: a right to life that could outweigh the mother's right to control her body. Anyone familiar with this long-lasting debate should know these views are fundamental (practically defining) for the parties in this dispute. Conservatives can say, of course, that the liberal views about the value of fetal life are wrong, whereas liberals can say the same for the conservative views.

How can an integrative bioethicist settle this dispute with their insistence on pluriperspectivism? While other bioethical theories (e.g. deontological or utilitarian) have specified criteria for making decisions in such cases, integrative bioethics inevitably ends up in a normative *cul-de-sac*: it lacks the basic normative principles needed to determine which participants of their round table hold morally acceptable views and which hold morally unacceptable ones. It, therefore, fails as a bioethical theory and collapses into a relativistic mosaic of diverse but normatively equivalent moral perspectives.

Lovro Savić and Viktor Ivanković (2016, 2017) criticized integrative bioethics along the same lines by introducing the notion of "semantic incommensurability". According to them, integrative bioethicists' enthusiasm for treating all perspectives as equally respectable participants in a bioethical dialogue implies that these perspectives are "non-hierarchical and cannot claim superiority in reaching truths over other acknowledged perspectives" (Ivanković and Savić 2016, 328). As we have seen, this nips in the bud the integrative bioethics' potential to resolve conflicts between perspectives. However, even if integrative bioethicists somehow agree that some perspectives have to be excluded, the "semantic incommensurability" problem will remain: the vocabularies of participants in the dialogue may look the same ("commensurable"), although they radically differ when it comes to the meaning ("semantics") of their central terms. A term that Savić and Ivanković (2017, 274; drawing on Fan 1997, 309) use to illustrate this is "autonomy". In the Western context, "autonomy" means self-determination, a subjective conception of the good and individual independence. In the East Asian context, it means family determination, an objective conception of the good and the value of harmonious dependence. "Semantic incommensurability" may affect various bioethical terms (such as "life," "death" or "dignity") and integrative bioethicists need to address it (which they do not) if they want to avoid pointless dialogues between perspectives that, despite their superficial similarities, talk past each other.

Let us now consider how integrative bioethicists respond to objections like these. Not long after the appearance of the "absence of normativity" objection (Bracanović 2012), Amir Muzur (2014) offered his response in a letter to the editor published in the journal *Developing World Bioethics*. Although Muzur's response is brief, it seems to be considered in the circle of integrative bioethicists as "the best defense of integrative bioethics from the narrowing of the imposed normativity" (Smiljanić 2022, 571). Muzur sketched several strategies of possible defense, but three of them deserve to be singled out and briefly commented on:

(a) "Why", asks Muzur, "should ethics and bioethics be (only) normative at all?" (2014, 109) Except for the fact that every relevant dictionary defines bioethics as a normative discipline, an obvious answer to this question is that setting norms or guiding action is the main reason why bioethics came into being in the first place. Integrative bioethicists themselves, as we have seen, present their school of bioethics as a discipline that is supposed to "resolve ethical questions related to life" (Jurić 2007, 83). Since "resolving ethical questions" is undoubtedly a normative activity, Muzur's idea of removing the normative component from bioethics is inconsistent with the primary motivation behind establishing integrative bioethics.

(b) "Normativness", Muzur is protesting, "imposes instant, one-sided solutions and thus often leads to mistakes" (2014, 109). It is unclear why he sees "one-sidedness" as a necessarily bad by-product of "normativity". Consider the analogy: A judge in the court of law reaches the verdict (normative judgment) based on impartial consideration of facts and arguments presented by both parties. That the judge ultimately decides in favor of one party does not mean that they are one-sided. The same applies to bioethical judgments: After impartially considering all arguments about a specific issue, we make a judgment that we believe is objective and correct. Muzur, however, seems to think that *any* normative judgment, as soon as it is made and irrespective of *how* it is made, is necessarily a one-sided imposition of one's norms or values on others. Such a typically relativistic approach paralyzes any bioethical decision-making process.

(c) For Muzur, bioethics, instead of being normative, "might be closer to a kind of buying time for humaneness until technology

> and science (if ever) provide us with crucial answers about life" (2014, 109). This might be the pinnacle of inconsistency within the integrative bioethics school. The task of integrative bioethics, as we are often told, is to deal with the "misuse of scientific results that can cause irreversible and catastrophic consequences for man and life as a whole" (Čović 2004, 164). However, if integrative bioethics (as Muzur maintains) is only about "buying time" until "technology and science" find answers to the burning ethical questions, then its historical role may not be as crucial as its founders typically claim. They claim, namely, that integrative bioethics is a spark of a "new ethical culture" that will provide us with "epochal orientation" (Čović and Jurić 2018)—not that it is some lowly placeholder for some future science and technology. In other words, Muzur's "buying time for humaneness" thesis may be a case not only of intra-theoretical inconsistency but also of intra-theoretical heresy.

The inconsistency of integrative bioethics becomes especially visible when one takes a closer look at the positions of its various proponents on ethical relativism (which, as we have seen, is a serious problem for its normative aspirations). Not all integrative bioethicists seem to view relativism as necessarily problematic. Sonja Kalauz (2011, 256-57), for example, defines integrative bioethics in a highly relativistic way, as a "polyvalent discipline" that has a "logically structured form" and "with the help of which every active participant, in accordance with his own theoretical and methodological templates, can come to the final normative judgment" (the talk about "one's own theoretical-methodological template" seems to imply not only relativist but also subjectivist reading of integrative bioethics). Although not willing to explicitly acknowledge the relativist status of integrative bioethics, Jos Schaefer-Rolffs (2012) interprets its "pluriperspectivism" in a way that is difficult to differentiate from a dictionary definition of ethical relativism: for him, pluriperspectivism means "(a) the non-hierarchic discourse of (b) multiple different points of view on one topic that are (c) rooted in different ideals and worldviews" (2012, 114). Some defenders of integrative bioethics also define its "orientational knowledge" in typically relativistic terms: as "a social norm" or "a set of patterns of mutual relations in the community" or as the "criterion of how it should be, as the community requires, and not as it actually is" (Smiljanić 2022, 570).

And yet, when it comes to the inner circle of the discipline's founders, they vigorously dissociate themselves from ethical relativism. Consider Jurić's dismissal of the relativistic interpretation of pluriperspectivism:

> Terrible "accusation" directed towards pluri-perspectivism ("Pluri-perspectivism is nothing but pure relativism") has no ground. Certain "relative relativism" inside the pluriperspectivistic way of discovering, viewing and constructing is unavoidable, just like in any approach which tends to be comprehensive, but it is something different from the "absolute relativism" of monoperspectivistic approach, because it can in no way embrace the whole: it always sacrifices some (massive) segments of the life and the world in order to achieve theoretical rigidity, self-sufficient coherence and consistency, in other words—"mythical" ideals of "exactness" and "objectivity". (Jurić 2012, 89)

In addition to being inconsistent with interpretations of "pluriperspectivism" offered by other defenders of integrative bioethics, an evident problem with this defense against the charge of relativism is its vagueness. What exactly is "relative relativism" as distinct from "absolute relativism"? What is this mysterious "whole" that the "monoperspectivistic" approach cannot embrace? What is it that the pluriperspectivistic approach "discovers", "views", and "constructs"? None of this is explained, despite the promise that pluriperspectivism is the superior tool of integrative bioethics that outcompetes all other schools of bioethics.

That the pluriperspectivist bioethical approach is nothing more than relativism in disguise should also become apparent from Jurić's claim that "exactness" and "objectivity" are "mythical ideals". Integrative bioethicists seem to have two mutually exclusive aims. On the one hand, they want to gain as many theoretical allies as possible (such as lawyers, physicians, or theologians), which explains the aggressive advertising of their "pluriperspectivism". On the other hand, they desperately want to avoid all associations with ethical relativism—if for no other reason than because most of their theoretical allies (especially theologians) do not subscribe to relativism. Unfortunately for them, sitting on this bioethical fence cannot go undetected forever, despite all the intentional and unintentional vagueness surrounding their normative agenda.[6]

[6] Probably aware of the danger of inevitable collapse into ethical relativism, Čović (2009) published a paper on integrative bioethics and the problem of truth. It mentions many things, from the fact that truth is a "Pilate's question" to the fact that already Aristotle was preoccupied with it. It says nothing, however, about how precisely integrative bioethics avoids the danger of the relativity of moral truth. It is interesting that no integrative bioethicist ever attempted to neutralize the moral relativity objection by relying on metaethical theories such as prescriptivism, quasi-realism or particularism.

Besides maintaining its internal consistency and providing a specific set of normative or action-guiding principles, a contender for a new and unique school of bioethics should have a clearly delineated scope of the issues it attempts to deal with. Integrative bioethics fares terribly in this respect too—not because it tries to cover too little ground, but because it tries to cover too much of it. Luka Perušić provides a vivid illustration of this overreach of integrative bioethics:

> If we produce vehicles whose exhaust pipes pollute the environment, it is a bioethical issue, just as the use of mobile devices containing ores mined by minors is a bioethical issue; the Panama Papers is a bioethical issue, nootropics are a bioethical issue and political and trade agreements and alliances, excessive production of toilet paper, regulation of the legal capacity of mentally challenged people and extraplanetary expansion, terrorism and surveillance, the concept of prisons and penitentiaries, GMO and the application of artificial intelligence, gender issues and the status of plants and animals, inter-religious conflicts, education and training systems, huge oxygenation and warfare and philosophical questions about the phenomena that arise in all problems, entail the area of the moral dimension of life and thus necessarily enter (integrative) bioethics as possible subjects of investigation. (Perušić 2019, 346-47)

If all the mentioned issues, from the production of toilet paper to extraplanetary expansion, are typically integrative-bioethical issues, a common-sense question arises: Which issues then remain to be dealt with by, for example, biomedical ethics, applied ethics, AI ethics, ethics of war, ethics of sexuality, ethics of information, political philosophy, social philosophy, environmental ethics or, simply, ethics? If integrative bioethics is *the* approach for dealing with all these issues, questions and problems, do we even need any other approach? If we are to believe its proponents, integrative bioethics will ultimately put all other practical or applied philosophical disciplines out of work. Given its weaknesses discussed so far (but also those to be addressed in the next section), this could not be further from the truth.[7]

[7] It may be difficult to say whether integrative bioethicists aim at establishing a specific bioethical theory for dealing with concrete problems or a more general approach to bioethics (a kind of Lakatosian research program). Both options are equally problematic. The first one (a bioethical *theory*), as we could see, is plagued by relativism, inconsistency and the lack of normativity. The second one (a bioethical *approach* or *program*) is burdened by the absence of a distinctive core consisting of its unique governing principles. Pluriperspectivism is a poor candidate for such a core

4. Integrative bioethics: The pseudoscientific features problem

A severe objection to integrative bioethics is that it has too many pseudoscientific features, especially the verbose and obscure language, a constant and unjustified self-glorification, a penchant for conspiracy theories and isolation from mainstream science.[8]

Verbose and obscure language is frequently used by various types of pseudoscience. Although new disciplines tend to generate new terminology and writing styles, integrative bioethics took this tendency too far. Some of its hard-to-understand phrases were already registered in Bracanović (2012), such as “phylonic responsibility”, “theoretical absurdism”, “epochal orientation” or “inductio ad absurdum”. In the meantime, integrative bioethicists—ironically, in an attempt to explain their discipline—generated a novel series of claims of the same level of unintelligibility. For example, Perušić explains:

> As a paradigmatic system that possesses a kind of method algorithm, integrative bioethics determines its own horizon of problem reception based on the fundamental determinants of its cognitive and practical activity. (Perušić 2019, 385)

The multidisciplinarity, pluriperspectivity and integrativity, explains Hrvoje Jurić, were necessitated, among others things, by the fact that “we are living in the world” in which “the science of nature lost its right to philosophy”, the fact that “we are living in the world where the philosophy lost its right to poetry”, and the fact that “we are living in the world where the poetry became so marginalized that it lost any right” (2012, 86). The necessity of the integrative bioethics itself, explains Željko Pavić, follows from the fact “that life—even in its ‘non-living’ form—happens as a constant mutual overflow, fusion, separation,

because the idea that all relevant perspectives must be considered when investigating specific issues is almost trivially true (maybe even a matter of basic academic integrity). The integrative bioethicists’ alarmist plea to include as many perspectives as possible in the bioethical debate creates the impression that bioethics has tragically failed in this respect. This is false. Quick and convenient evidence of inherent pluralism of contemporary bioethics can be found, for example, in the variety of thematic specializations of a large number of contemporary bioethics journals, such as the *Journal of Medical Ethics, Journal of Agricultural and Environmental Ethics, Developing World Bioethics, Asian Bioethics Review, Christian Bioethics, International Journal of Feminist Approaches to Bioethics, Literature and Medicine* (a very diverse list could go on).

[8] The original objection was put forward by Bracanović (2012), based on a classic study in pseudoscience by Gardner (1957). What follows is a further elaboration of how integrative bioethics fares concerning three groups of pseudoscientific features (verbose and obscure language, a constant self-glorification and isolation from mainstream science). The penchant for conspiracy theories, although a distinctive feature of many pseudoscientific enterprises (and most likely of integrative bioethics as well), is a topic that is too complex to deal with in such a limited space.

differentiation, rise and fall" and that "no single scientific 'subject area' nor any idea of life can replace or explain life itself" (2014, 585). Luka Janeš explains that

> (…) integrative bioethics with its consideration of the general values of Earth's plurality, come as a certain 'post-technological Prometheus' who ought to banish enclosed darkness of technicized science with the burning flame of morality governed by the principle of All-Oneness. (Janeš 2017, 47)

It is difficult not only to make sense of these "explanations", but also to see how they all constitute explanations of the same thing (integrative bioethics).

Self-glorifying claims, as another typically pseudoscientific feature, are very common among integrative bioethicists. As indicated in the first section of this paper, integrative bioethicists have strong convictions about the historical and global importance of their enterprise, and they do not hesitate to describe it in terms like the "original and foundational concept of the European bioethics" or the "innovative and recognizable European concept of bioethics".

An intriguing method of self-glorification is to pick great names from the history of philosophy and science and interpret them as their predecessors. For example, in their paper on German priest Fritz Jahr (credited for coining the term "bioethics" in 1927), Amir Muzur and Iva Rinčić claim that Jahr's work "might be interpreted as an anticipation by several decades of the integrative bioethics perspectivism of Croatian bioethicist Ante Čović" (2011, 136). In her paper on Russian existentialist Nikolai Berdyaev, Marija Selak claims that his "notion of 'new medievalism' can be understood as the predecessor of the concept of integrative bioethics" (2009, 612). German philosopher Karl Löwith, as Selak claims in a different article, is also "a precursor and incentive to the idea of integrative bioethics" (2011, 525). Slavko Amulić claims that the work of the famous physicists Fritjof Capra "perfectly fits into the orientational framework of bioethics as the pluriperspectival area" (2007, 422). Dževad Hodžić claims that American mathematician Alfred N. Whitehead is "interesting and significant for the integrative horizon of bioethics" (2011, 296). The founding father of integrative bioethics himself, Čović, claims, for example, that "Plato's dialogues can be read as elementary exercises in pluriperspectivism", as well as that "for the historical-philosophical reconstruction of the pluriperspectivist understanding of truth especially important are explicit forms of perspectivism endorsed by Leibniz, Nietzsche and Ortega y Gasset" (Čović 2009, 191). The champion of self-glorification is probably Janeš (2018, 313), who talks about "the explosive power of optimism and of the scientific, life-augmenting cognitive light that integrative bioethics exudes in relation to the potential treatment of psychological suffering" (notice the hint about no less than potentially healing powers of integrative bioethics).

Integrative bioethicists desire to be seen in good company and keep up appearances. To what extent, however, is that desire justified? It is complicated to provide evidence that something is not as important as someone claims it to be (since the only evidence of the non-importance of something is the absence of evidence of its importance). Still, we can mention two pieces of indirect evidence that this self-glorification is an unjustified peculiarity of the Croatian branch of integrative bioethics.

The first evidence is the MA and Ph.D. program in bioethics at the University of Crete. As we have seen, the Croatian integrative bioethicists proudly claim that it "operates on the basis of integrative-bioethical principles".[9] However, if we examine the publicly available data about this program,[10] it does not seem to have any kinship to integrative bioethics. For example, the program does not have a single course on anything "integrative" or "pluriperspectivist", but it does have many well-conceived and bioethically relevant courses like "Conceptual foundations of bioethics", "Introduction to modern biology", "Philosophy of science" or "Theories of distributive justice". As for the required literature, no publications of integrative bioethicists are mentioned and almost all courses are based on English-language (some would say "analytic philosophy") classics like John Rawls, Tom Beauchamp, Ronald Dworkin, Helga Kuhse, Peter Singer, Bernard Williams, etc. Also telling is the following detail: Whereas Croatian integrative bioethicists claim that bioethics is about *everything* related to "life in all its forms, stages, phases and appearances" (Jurić 2007, 83), avoiding the language of "normativity" because it "imposes instant, one-sided solutions and thus often leads to mistakes" (Muzur 2014, 109), their Greek colleagues describe their MA and Ph.D. bioethics program in reasonable terms as "primarily the *normative* investigation of moral challenges resulting from *developments in the life sciences and biotechnology*" (emphasis added).[11] Apparently, the only conceptual connection between the Crete MA and

[9] See in footnote 2 the cited document *Koncept i projekt integrativne bioetike*, 9.

[10] Available at http://bioethics.fks.uoc.gr/en/MainFrameSet.htm (accessed August 4, 2024)

[11] See the "Director's note" at http://bioethics.fks.uoc.gr/en/MainFrameSet.htm (accessed August 4, 2024)

Ph.D. program with the Croatian brand of integrative bioethics seems to be the word “bioethics”.

The second piece of evidence that should make one skeptical about the self-glorifying claims of integrative bioethicists can be obtained by browsing recent German literature on bioethics and applied ethics. Why German? Remember that the internationalization of integrative bioethics occurred due to the cooperation between two groups of philosophers (Croatian and German), who, among other things, published six volumes of papers with the German Academia Verlag. Did this publishing project have any *Wirkungsgeschichte* in the German bioethical community? Apparently not, and if it did, it surely was not as revolutionary as integrative bioethicists would like us to believe it is. For example, in 2015, Sturma and Heinrichs (2015), in cooperation with the Deutsche Referenzzentrum für Ethik in den Biowissenschaften (DRZE), published the *Handbuch Bioethik*, containing entries on 28 concepts of bioethics, 46 bioethical topics and eight interfaces between bioethics and other disciplines or social areas. Integrative bioethics is not mentioned.[12] This is surprising if integrative bioethics really is “the widest concept of European bioethics” (Perušić 2018, 316) and “the original and foundational concept of European bioethics” (Čović 2023, 14). Even more surprising—and somewhat ironical—is that even the closest German partners of Croatian integrative bioethicists are also not too eager to mention “integrative bioethics” in their other publications. For example, in 2018, Walter Schweidler published his *Kleine Einführung in die Angewandte Ethik*. In this book, Schweidler discusses many bioethically important topics (from science, technology and medicine to economy, society and environment), but he does not mention “integrative bioethics” or any of its proponents. In a nutshell, the entire integrative bioethics agenda seems to be assigned a much greater value by its Croatian proponents than by their German colleagues.

Isolation from mainstream science, according to Michael Gardner (1957), means that pseudoscientist stands “outside the closely integrated channels through which new ideas are introduced and evaluated”, does not “send his findings to the recognized journals”, in most cases “is not well enough informed to write a paper with even a surface resemblance to a significant study”, speaks “before organizations he himself has founded, contributes to journals he himself may edit, and (…) publishes books

[12] Similarly, in *Handbuch Angewandte Ethik*, edited by Stoecker, Neuhäuser and Raters (2011), there is not even a trace of mention of integrative bioethics. The only thing “integrative” mentioned is P. Urlich’s “integrative ethics of economy” (*integrative Wirtschaftsethik*).

only when he or his followers can raise sufficient funds to have them printed privately" (1957, 11). Anyone familiar with integrative bioethics must be aware of the following: (1) Integrative bioethicists rarely, if ever, talk at conferences not organized by themselves or their partners. (2) Integrative bioethicists rarely, if ever, publish in journals not edited by themselves or their partners (many of their papers are published in journals *Filozofska istraživanja* and *Synthesis Philosophica*, whose editor-in-chief is Čović). (3) Integrative bioethicists rarely, if ever, publish books with publishers they do not control (almost all books on integrative bioethics in Croatia are published with Pergamena, a publishing house whose editor of the "Bioethics" series is Čović). If one searches the Web of Science database for the phrase "integrative bioethics", it is almost impossible to find a paper published in an independent journal or by an author not a member of their circle.[13] Works of integrative bioethicists are also rarely cited in papers published by non-members of their circle.[14]

A detail supporting the "isolation from mainstream science" thesis about integrative bioethics is that, in 2013, the Croatian Minister of Science publicly criticized prominent integrative bioethicists for abusing their positions in two journals of the Croatian Philosophical Society (*Filozofska istraživanja* and *Synthesis Philosophica*). It was discovered that the members of their editorial boards (which is the same in both journals) published a large number of papers in these journals, facilitating thus their academic promotions (some of them even got promotions to

[13] The Web of Science search for the phrase "integrative bioethics" (in "all fields" and for all "document types") yields 37 papers. Of those 37 papers, 24 were published in integrative bioethicists' "home" journals *Filozofska istraživanja* and *Synthesis Philosophica*, 5 in other journals (3 in Croatian, 2 in foreign journals) with integrative bioethicists as their (co)authors, 3 are critiques by Bracanović, Ivanković and Savić, 1 is a review of a book that has "integrative bioethics" as its subtitle. The search also yields 4 papers mentioning the phrase "integrative bioethics" published in foreign journals but, interestingly, with no reference to its Croatian papers (the search was performed on January 20, 2024).

[14] Is integrative bioethics a unique "citation cartel"? Additional data would be needed to answer this question. However, if one takes as the test case the publications of the founder of integrative bioethics (Čović), certain preliminary positive evidence exists. According to the Web of Science, his best-cited work is the book *Etika i bioetika* (Čović 2005), which has only 12 citations, all in the journals *Filozofska istraživanja* and *Synthesis Philosophica*. His second best-cited work is the 2018 paper (co-authored with Jurić), which has three citations, all in *Filozofska istraživanja* and *Synthesis Philosophica* (moreover, two of those are self-citations). His 2006 paper on pluralism and pluriperspectivism also has three citations, all in *Filozofska istraživanja*. Since almost all these citations stem from papers published *after* Čović became editor-in-chief of *Filozofska istraživanja* and *Synthesis Philosophica* in 2006, the "citation cartel" hypothesis could be worthy of further investigation.

senior positions based *exclusively* on papers published in “their” journals).[15]

A further detail supporting the same thesis is the following: In 2019, the Faculty of Humanities and Social Sciences of the University of Zagreb launched an investigation into whether Čović should have his full professor title revoked. A committee appointed for this purpose, consisting of three philosophers working in different traditions, reported, among other things, that Čović

> (…) did not fulfill the prescribed conditions for promotion to the position of full professor, largely because he violated the basic and generally accepted norms of academic ethics for years, which led to the fact that, for the vast majority of his works, there is a sound suspicion that they did not undergo the necessary and impartial professional evaluation and verification before publication.[16]

The committee concluded that

> (…) in almost 30 years of his university career (from 1976 to 2005), Professor Čović failed to publish a single original scientific article anywhere else except in those two journals in which he was the editor (in the vast majority of cases the editor-in-chief) during that period, or there was a suspicion of bias in the evaluation of his articles.[17]

The Faculty Council of the Faculty of Humanities and Social Sciences in Zagreb accepted this report and decided (with 47 votes in favor, 22 against and 13 invalid) to initiate the process of revoking Čović’s title.[18]

[15] See Tanja Rudež: Čović i Jurić karijeru su gradili tako da su sami sebi objavljivali radove u časopisima, *Jutarnji list*, April 10, 2013, available at https://www.jutarnji.hr/life/znanost/covic-i-juric-karijeru-su-gradili-tako-da-su-sami-sebi-objavljivali-radove-u-casopisima-1136426 (accessed August 4, 2024).

[16] The quote is translated from the report of the committee, which is publicly available at https://www.srednja.hr/app/uploads/2019/01/Izvjestaj-zvanje-%C4%8Covi%C4%87.pdf (accessed August 4, 2024).

[17] The data and quote are translated from the report of the committee, which is publicly available at https://www.srednja.hr/app/uploads/2019/01/Izvjestaj-zvanje-%C4%8Covi%C4%87.pdf (accessed August 4, 2024).

[18] In the end, Čović’s full professor title was not revoked as the Faculty decision was not confirmed by the Scientific Field Committee for Philosophy and Theology, which oversees scientific promotions in Croatia. The committee was chaired by Čović’s close colleague, who played the crucial role in all of his academic promotions. See https://www.srednja.hr/faks/covjek-kojemu-nitko-nije-mogao-nista-prorektoru-covicu-nece-se-oduzeti-zvanje-redovitog-profesora (accessed August 4, 2024).

To make the long story short, integrative bioethicists seem pretty isolated from mainstream science and rarely exposed to independent evaluation of their ideas.[19]

5. Concluding remarks: Is integrative bioethics "European"?

Practical disciplines like ethics and applied ethics are relatively diverse in Croatia, encompassing analytic, continental and neo-scholastic approaches. Still, integrative bioethics is undoubtedly the most widespread and visible. Only integrative bioethicists run a number of regional centers (and the independent Centre of Excellence), at least two philosophical journals, a regular annual conference, a bioethics book series with an independent publisher—they even managed to introduce their bioethical teachings into the ethics curriculum for high schools.[20] Such an expansion of integrative bioethics in Croatia can be explained either by the fact that bioethics is an attractive field in itself or, alternatively, by the programmatic promise of integrative bioethicists that everyone (philosophers, theologians, physicians, artists, even laypersons) has a guaranteed place at their pluriperspective "round table". A complementary explanation (for which there is not enough space here) would be a kind of "sociology of integrative bioethics" examining possible connections between the expansion of integrative bioethics and the academic and political positions held by its proponents during the past 30 years in Croatia (ranging from ministers and deputy ministers of science, over university vice-rectors and heads of philosophy

[19] A piece of evidence confirming the same story is Vlatko Smiljanić's paper "The history of defamation of integrative bioethics" (2022) published in *Filozofska istraživanja*. The conjunction of its following three features is noteworthy: (1) It enthusiastically glorifies integrative bioethics (e.g. describing its "meteoric rise in the Croatian and European scientific and professional community") and aggressively denigrates its critics (e.g. accusing them of "defamation", "diabolization", and "denunciation", even calling some of them "half-crazy"). (2) It is arguably one of the worst papers ever published in this journal (riddled with obscure concepts, logical flaws, even unintelligible sentences), which is to some extent explainable by the author's lack of formal training in philosophy (he is a historian) and this being his first philosophical publication. (3) It was published in a journal whose editor-in-chief (Čović), deputy editor (Jurić), managing editor (Perušić), and many editorial board members are prominent advocates of integrative bioethics. In summary, despite undergoing strict doctrinal and scholarly quality control by the highest authorities of integrative bioethics, this paper impeccably exemplifies Gardner's (1957, 11) depiction of pseudoscientific practices. *Sapienti sat*.

[20] In Igor Lukić's (2021) high-school ethics textbook, the presentation of integrative bioethics spans several pages, presenting in a positive light its concepts like "pluriperspectivism" and "integrativity" and extensively quoting its main proponents such as Čović and Jurić. Of course, the question is whether such a novel, local and controversial bioethical theory is a fitting material for fourth-graders. The reviewers who evaluated and recommended the textbook for use in schools obviously found it perfectly fitting (bear in mind, however, that one of its reviewers, as we find out from its opening pages, was Jurić himself).

departments, to members of various committees in charge of things like government subsidies for scientific books and journals or academic promotions).[21]

All in all, there is no doubt that this curious spread of integrative bioethics in Croatia is *not* a testimony that integrative bioethics is a "developmental shift" that "transferred bioethics from the United States to Europe" or the "original and foundational concept of European bioethics" or the project that "Europeanizes bioethics" by "regenerating the spiritual potential of the European philosophical heritage". In Croatia, the adjective "European" is frequently (mis)used as a "virtue signal", intended to indicate that some enterprise has transcended the local context and become globally known and appreciated in terms of accomplishing either European quality, European reception or European identity. We can complete our discussion by summarizing how integrative bioethics fares concerning these three levels of its hoped-for "Europeanization".

It should have become evident by now that, judging by its accomplishments, integrative bioethics does not deserve the "European" label. Here is just a selection of keywords that should remind one of its theoretical shortcomings: absence of normativity, inconsistency, poorly defined scope of problems, not addressing concrete bioethical issues, verbose and obscure language, constant and unjustified self-glorification, low scholarly standards, isolation from the mainstream science. Why have they failed, in nearly three decades, to publish anything of bioethical significance? An educated guess could be that their research program was designed and controlled by latecomers to bioethics who spent the formative years of their academic careers working in a typically Marxist-socialist paradigm, who were strangers to English-language (bio)ethical literature, and who continued to apply their old patterns of murky reasoning to newly discovered bioethical issues. Of course, their younger colleagues and students *could* perform much better by engaging with more recent bioethical debates and literature. Yet, it seems they may have succumbed to self-censorship and decided not to go beyond the standards

[21] The need for a critical discussion on integrative bioethics in both the Croatian and European philosophical contexts arises not only from its dangerous potential to blur the boundary between reputable and substandard scholarly work in ethics and applied ethics. A financial cost also needs to be considered. The nearly three-decade-long expansion of the integrative bioethics agenda has been accompanied by substantial financial support from public sources (for research projects, conferences, books, journals, etc.). However, if integrative bioethics is a murky enterprise with pseudo-scientific undertones, one cannot but wonder whether this support could have been utilized better. For all these reasons, a periodic philosophical check-up of what is happening in and around this peculiar school of bioethics seems welcome.

set by the founders of the movement. This is probably one of the reasons why integrative bioethicists never apply their allegedly unique methods to concrete bioethical problems but remain focused on the eternal "laying of the foundations" (*Grundlegung*) of their discipline. This is a strange destiny for an allegedly revolutionary school of bioethics. Instead of becoming inherently practical, focusing on specific problems created by science and technology, integrative bioethics remains highly theoretical, focusing predominantly on itself.

Integrative bioethicists, as we have seen in the first section, unabashedly claim that their brand of bioethics is the "original and foundational concept of the European bioethics" and a "striking development of the bioethical discipline in Central and Southeastern Europe in the last thirty years" that "encompassed the entire European context". In the third section, however, we could see that such claims are way too exaggerated, not only because most activities of integrative bioethicists (especially of their Croatian branch) are always limited to the same circle of scholars organizing conferences, summer schools, round tables and lectures with more or less the same circle of participants, not only because they typically publish their papers and books only in venues they control, but also because the standard reception of integrative bioethics, in terms of its advocates being cited in publications by independent scholars, is practically non-existent. (Moreover, as we could also see, even their closest foreign partners omit to mention integrative bioethics in their other publications and projects.) Integrative bioethicists have a stable collaboration with scholars and institutions from several European countries, but that is "business as usual" that many scholars from Croatia and neighboring countries are engaged in without claiming any European glory. Integrative bioethics, therefore, has no recognizable European reception because the belief in its "epochal" role rarely travels beyond the narrow circle of their main proponents.[22]

Finally, although integrative bioethicists claim that they are "Europeanizing bioethics" by "regenerating the spiritual potential of the European philosophical heritage", nothing in their writings justifies such a claim. Ironically, if one takes a closer look at the philosophical heritage they most frequently invoke, it does not seem particularly European or, for that matter, particularly philosophical. Integrative bioethicists frequently point out, for example, that their position has "footholds" in the work of

[22] Probably the highest-ranking journals in which integrative bioethics was discussed are *Developing World Bioethics, Bioethics,* and *Medicine, Health Care and Philosophy,* in which criticisms of integrative bioethics by Bracanović (2012), Ivanković and Savić (2016), and Savić and Ivanković (2017) were published.

V. R. Potter. However, Potter was neither European (he was American) nor a philosopher (he was a biochemist and oncologist). They also frequently invoke the ideas of Aldo Leopold, although he was also an American and non-philosopher (his education was in forestry). They are particularly keen, as we could see in the second section, to find their "predecessors" or theoretical allies amongst a heterogeneous group of thinkers, like American mathematicians and physicists, Russian existentialists or German pastors. Although this group also includes several German philosophers, this is too meager and unsystematic to justify any talk of a unique European identity of integrative bioethics.[23] In a nutshell, integrative bioethics turns out to be a blind alley of European bioethics.

Acknowledgments

This paper was produced within the project "Ethics and Social Challenges" at the Institute of Philosophy in Zagreb, monitored by the Ministry of Science, Education and Youth of the Republic of Croatia and financed through the National Recovery and Resilience Plan 2021–2026 of the European Union—NextGenerationEU. I am grateful to two anonymous reviewers for the *European Journal of Analytic Philosophy* for their valuable comments.

REFERENCES

Amulić, Slavko. 2007. "Poredbenost perspektiva." *Filozofska istraživanja* 27(2): 407-25.

Beauchamp, Tom L., and James L. Childress. ^{1}1979, 2013. *Principles of Biomedical Ethics*. New York, Oxford: Oxford University Press.

Borovečki, Ana. 2014. "Croatia." In *Handbook of Global Bioethics,* edited by Henk A. M. J. ten Have and Bert Gordijn, 1049-65. Dordrecht, Heidelberg, New York, London: Springer.

[23] An ingredient that could make integrative bioethics a truly European project, possibly even boost its overall clarity, is a stronger reliance on the tradition of analytic philosophy. Integrative bioethicists, unfortunately, have a strong aversion to analytic philosophy, despite its firm European identity and roots in, for example, Frege, Wittgenstein, Vienna circle or Polish logic. One could argue that even Beauchamp and Childress' "principlism" has a more recognizable European identity than integrative bioethics. Their four principles (autonomy, beneficence, nonmaleficence and justice) involve an obvious basis in and systematic elaboration of European intellectual traditions like deontology (Kant and Ross), utilitarianism (Bentham and Mill), specific theories of justice and rights (Locke and Hegel), Hippocratic tradition, etc. More recent editions of their *Principles of Biomedical Ethics* (2013) also include the extensive elaboration of Aristotelian virtue ethics and its importance for biomedical ethics.

Bracanović, Tomislav. 2012. “From Integrative Bioethics to Pseudoscience.” *Developing World Bioethics* 12(3): 148-56. https://doi.org/10.1111/j.1471-8847.2012.00330.x

Cifrić, Ivan. 2006. “Bioetička ekumena: Potreba za orijentacijskim znanjem.” *Socijalna ekologija* 15(4): 283-310.

Čović, Ante. 1988. *Marksizam kao filozofija svijeta*. Zagreb: Hrvatsko filozofsko društvo.

———. 2004. *Etika i bioetika: Razmišljanja na pragu bioetičke epohe*. Zagreb: Pergamena.

———. 2005. “Bioethik unter den Bedingungen des Postkommunismus––Fallbeispiel Kroatien.” In *Bioethik und kulturelle Pluralität / Bioethics and Cultural Plurality,* edited by Ante Čović and Thomas Sören Hoffmann, 148-172. Sankt Augustin: Academia Verlag.

———. 2006. “Pluralizam i pluriperspektivizam.” *Filozofska istraživanja* 26(1): 7-12.

———. 2009. “Integrativna bioetika i problem istine.” *Arhe* 6(12): 185-94.

———. 2011. “Pojmovna razgraničenja: moral, etika, medicinska etika, bioetika, integrativna bioetika.” In *Bioetika i dijete,* edited by Ante Čović and Marija Radonić, 11-24. Zagreb: Pergamena.

———. 2023. “Uvod / Introduction.” In *21st Lošinj Days of Bioethics*: *Program and Abstracts,* 11-15. Zagreb: Croatian Philosophical Society.

Čović, Ante, and Hrvoje Jurić. 2018. “Epochal Orientation, New Ethical Culture, and Integrative Bioethics.” *Formosan Journal of Medical Humanities* 19(1-2): 20-30.

Fan, Ruiping. 1997. “Self-Determination vs. Family-Determination: Two Incommensurable Principles of Autonomy.” *Bioethics* 11(3-4): 309-322. https://doi.org/10.1111/1467-8519.00070

Gardner, Martin. 1957. *Fads and Fallacies in the Name of Science*. New York: Dover Publications.

Hodžić, Dževad. 2011. “Whiteheadova filozofija prirode i bioetika.” *Filozofska istraživanja* 31(2): 291-297.

Ivanković, Viktor, and Lovro Savić. 2016. “Integrative Bioethics: A Conceptually Inconsistent Project.” *Bioethics* 30(5): 325-35. https://doi.org/10.1111/bioe.12235

Janeš, Luka. 2017. “Paradogma of the Psychic Entropy of Evil and the Palingenesis of All-Oneness.” *Synthesis Philosophica* 32(1): 31-50. https://doi.org/10.21464/sp32103

———. 2018. “Budućnost filozofije psihe u Hrvatskoj.” *Filozofska istraživanja* 38(2): 293-314. https://doi.org/10.21464/fi38205

Jurić, Hrvoje. 2007. “Uporišta za integrativnu bioetiku u djelu Van Rensselaera Pottera.” In *Integrativna bioetika i izazovi* *suvremene civilizacije,* edited by V. Valjan, 77-99. Sarajevo: Bioetičko društvo BiH.

Jurić, Hrvoje. 2012. "Multi-Disciplinarity, Pluri-Perspectivity and Integrativity in the Science and Education." *The Holistic Approach to Environment* 2(2): 85-90.

Kalauz, Sonja. 2011. *Sestrinska profesija u svjetlu bioetičkog pluriperspektivizma*. Zagreb: Pergamena.

Katinić, Marina. 2012. "Filozofija za djecu i mlade i integrativna bioetika." *Filozofska istraživanja* 32(3-4): 587-603.

Lukić, Igor. 2021. *Etika 4: Koracima budućnosti. Udžbenik etike u četvrtom razredu srednjih škola*. Zagreb: Školska knjiga.

Muzur, Amir. 2014. "The Nature of Bioethics Revisited: A Comment on Tomislav Bracanović." *Developing World Bioethics* 14(2): 109-10. https://doi.org/10.1111/dewb.12008

Muzur, Amir, and Iva Rinčić. 2011. "Fritz Jahr (1895-1953) – The Man who Invented Bioethics." *Synthesis Philosophica* 51(1): 133-39.

Pavić, Željko. 2014. "'Pluriperspektivizam' – slučaj jedne natuknice u *Filozofskome leksikonu*." *Filozofska istraživanja* 34(4): 577-600.

Perušić, Luka. 2018. "Kozmobioetika: uvodna rasprava o bioetičkim aspektima kozmičkog društva." *Obnovljeni život* 73(3): 311-328. https://doi.org/ 10.31337/oz.73.3.2

———. 2019. "Narav i metoda integrativne bioetike: rasprava." In *Integrativno mišljenje i nova paradigma znanja,* edited by Ante Čović and Hrvoje Jurić, 323-412. Zagreb: Pergamena.

Potter, Van Rensselaer. 1988. *Global Bioethics: Building on the Leopold Legacy*. East Lansing: Michigan State University.

Savić, Lovro, and Viktor Ivanković. 2017. "Against the Integrative Turn in Bioethics: Burdens of Understanding." *Medicine, Health Care and Philosophy* 21(2): 265-76. https://doi.org/10.1007/s11019-017-9799-5

Schaefer-Rolffs, Jos. 2012. "Integrative Bioethics as a Chance: An Ideal Example for Ethical Discussions." *Synthesis Philosophica* 27(1): 107-122.

Schweidler, Walter. 2018. *Kleine Einführung in die Angewandte Bioethik*. Wiesbaden: Springer.

Selak, Marija. 2009. "Bioetički osvrt na filozofiju Nikolaja A. Berdjaveva." *Filozofska istraživanja* 29(3): 603-14.

———. 2011. "Philosophy of the World and Philosophy of Karl Löwith as a Precursor and Incentive to the Idea of Integrative Bioethics." *Jahr* 2(4): 525-32.

Smiljanić, Vlatko. 2022. "Povijest difamiranja integrativne bioetike." *Filozofska istraživanja* 42(3): 561-78. https://doi.org/10.21464/fi42309

Stoecker, Ralf, Christian Neuhäuser, and Marie-Luise Raters, eds. 2011. *Handbuch Angewandte Ethik*. Stuttgart, Weimar: Verlag J. B. Metzler.

Sturma, Dieter, and Bert Heinrichs, eds. 2015. *Handbuch Bioethik*. Stuttgart, Weimar: Verlag J. B. Metzler.

Tomašević, Luka. 2013. "Razvoj bioetike u Hrvatskoj." *Crkva u svijetu* 48(4): 488-503.

European Journal of Analytic Philosophy EuJAP | 2024 | Vol. 20 | No. 1 | 209-234
https://doi.org/10.31820/ejap.20.1.9

ARE COMPOSITE SUBJECTS POSSIBLE? A CLARIFICATION OF THE SUBJECT COMBINATION PROBLEM FACING PANPSYCHISM

Siddharth S[1]

[1] Sai University, India

Original scientific paper – Received: 10/02/2024 Accepted: 17/08/2024

ABSTRACT

Panpsychism, the view that phenomenal consciousness is present at the fundamental physical level, faces the subject combination problem—the question of whether (and how) subjects of experience can combine. While various solutions to the problem have been proposed, these often seem to be based on a misunderstanding of the threat posed by the subject combination problem. An example is the exchange in this journal between Siddharth (2021) and Miller (2022). Siddharth argued that the phenomenal bonding solution failed to address the subject combination problem, while Miller responded that Siddharth had (among other things) misunderstood the problem that the phenomenal bonding solution was trying to solve. In this paper, I seek to clarify the real subject combination problem facing panpsychism, and on this basis, evaluate the various attempts at defending the possibility of subject composition.

Keywords: panpsychism; combination problem; subject composition; consciousness.

Correspondence: siddharth.nias@gmail.com

Introduction

> A spectre is haunting panpsychism—the spectre of the subject combination problem.

Panpsychism, the view that phenomenal consciousness—ontologically subjective and qualitative phenomena that have a 'what-it-is-like' feel associated with them[1]—is present at the fundamental physical level, has regained prominence in the past two decades as a viable middle-path between physicalism and dualism.[2] However, critics argue that panpsychism faces a 'hard' problem of its own, of explaining whether (and how) microphysical entities that are themselves bestowed with subjectivity—are subjects of experience—can combine to form subjects of macrophysical entities such as human beings.[3] Such composition, it is claimed, is unintelligible and impossible. The *subject combination problem*, as it has come to be known, thus threatens to derail panpsychism's claim as a viable middle-path between physicalism and dualism.

In response, some panpsychists have argued that subjects can indeed compose, and proposed solutions to the subject combination problem (Goff 2016; Miller 2017; Roelofs 2019; Goff and Roelofs forthcoming). Siddharth (2021), in an article published in this journal, offered a critique of the *phenomenal bonding* (PB) solution proposed by Goff (2016) and Miller (2017), and argued that it failed to adequately address the problem. The PB solution was first proposed by Goff (2016), who contended that it was possible for subjects to enter into a relation that necessitated—brought into existence—a composite subject. The relation that fulfilled this role was the phenomenal bonding relation. While Goff conceded that we have no positive conception of the PB relation, Miller (2017) thought otherwise. He proposed that the *co-consciousness* relation—the relation "in virtue of which conscious experiences have a conjoint

[1] See Nagel (1974) for more on 'what-it-is-like' talk.

[2] For arguments in favour of panpsychism, see Chalmers (2016a), Goff (2017), Maxwell (1979), Mørch (2014), Rosenberg (2004), and Strawson (2006a, 2006b, 2016). See Freeman (2006), Brüntrup and Jaskolla (2016), Seager (2019), and Skrbina (2009) for discussions of various issues related to panpsychism. Panpsychists commonly distinguish between two versions of the view: *micropsychism*, wherein the microphysical entities (such as quarks, electron, etc.) are taken to be fundamental; and *cosmopsychism,* wherein the cosmos-as-a-whole is taken to be the fundamental entity. While it is only micropsychism that faces the subject combination problem strictly speaking, cosmopsychism faces an analogous problem—the de-combination problem (see Miller 2018a). In this paper, I focus only on the subject combination problem facing micropsychism, and hence use the term panpsychism to refer only to this version of the view.

[3] For instance, see James (1890), Coleman (2014), and Goff (2009). See Chalmers (2016b) for a comprehensive discussion of the combination problem.

phenomenology or a conjoint what-it-is-like-ness" (Miller 2017, 548)—could play the role of the phenomenal bonding relation, and that we could form a positive conception of inter-subject co-consciousness based on our knowledge of intra-subject co-consciousness. Siddharth (2021) argued that the proponents of the PB solution were guilty of begging the question, and that Miller's proposal to form a positive conception of inter-subject co-consciousness did not work.

In response, Miller (2022) claimed that Siddharth's critique was off the mark for the following reasons:

1. Siddharth's critique was based on the intuition that subjects were ontologically unified and private; however, he gives no justification for these theses.
2. In arguing that the PB solution does not show how subject composition is possible, Siddharth commits the strawman fallacy; the proponents of PB were not addressing the *mereological problem* (the question of the possibility of composite subjects), but the *subject-summing-problem* (the question of the mechanism of composition). The mereological problem, nevertheless, has been addressed by others (Miller 2018b; Roelofs 2019; Goff and Roelofs forthcoming), claimed Miller.
3. Contra Siddharth, analogical extension can be used to form a positive conception of the PB solution.

The exchange between Siddharth and Miller highlights the need for a clarification of the threat posed by the subject combination problem to panpsychism, and on this basis, a comprehensive evaluation of the theories of subject composition that have been offered in response to the problem. These are what I seek to do in this paper.

I begin by showing that the real subject combination problem is the question of whether such composite subjects are possible in the first place (§1), followed by an examination of Miller's response to Siddharth (§2) where I argue that Siddharth's critique of the PB solution is correct. I then evaluate other attempts at addressing the mereological problem, and show that contra Miller's claim, they do not show that composite subjects are possible; given this, the phenomenal bonding solution (and other similar proposals) are either trivial, or guilty of begging the questions. I conclude by rearticulating the subject combination problem facing panpsychism in light of these discussions.

1. Clarifying the subject combination problem

Let us begin with William James' influential articulation of the problem. It is worth repeating his oft-quoted passage here:

> Take a hundred of them [feelings], shuffle them and pack them as close together as you can (whatever that may mean); still each remains the same feeling it always was, shut in its own skin, windowless, ignorant of what the other feelings are and mean. There would be a hundred-and-first feeling there, if, when a group or series of such feelings were set up, a consciousness belonging to the group as such should emerge. And this 101st feeling would be a totally new fact; the 100 original feelings might, by a curious physical law, be a signal for its creation, when they came together; but they would have no substantial identity with it, nor it with them, and one could never deduce the one from the others, or (in any intelligible sense) say that they evolved it. (James 1890, 160)

Here, James describes a "feeling" as "shut in its own skin, windowless". A little later, he refers to feelings as the "most absolute breaches in nature" (James 1890, 226). This aspect of subjects has been cashed out and understood in various ways. As Siddharth (2021) notes, it intuitively seems that subjects are ontological unities, entities that are "fundamentally unified, utterly indivisible" (Strawson 2009, 378). Further, subjects seem to be such that a token experiential quality experienced by one subject cannot be experienced by another subject. Let us call these two aspects of a subject its ontological *unity* and *privacy*:

> *Ontological Unity:* Subjects of experience are ontological unities, such that the unity/singleness is not just a matter of convention or abstraction.
>
> *Ontological Privacy:* Subjects of experience are such that the token phenomenal quality experienced by one subject is not available to be experienced by another subject.

What justifies the conception of subjects as ontologically unified and private? I do not think there is a rigorous defence to be offered; nevertheless, I still think that these are intuitions that a panpsychist cannot reject.

To explain human consciousness—subjectivity and experientiality—panpsychists contend that consciousness ought to be present at the fundamental, micro level. Given that the only subjects that humans have access to are their own, panpsychists (and others) take human subjectivity to be the paradigm of subjectivity *simplicter*. However, as James pointed out and we just saw, human subjects *seem* to be characterised by ontological unity and privacy. That there is such a *seeming* is agreed upon by everybody.[4] Even proponents of subject composition do not deny that subjects seem to be ontologically unified and private, but only that the *seeming* is an accurate indicator of how subjects really are.

Given that human subjectivity is the only kind of subjectivity we have direct epistemic access to, intuitions that we form based on what we know about human subjects ought to hold primacy over other metaphysical intuitions and possibilities we may entertain, insofar as these other intuitions and possibilities apply to subjects (human or otherwise). This is not to say that unity-privacy intuitions cannot be rejected, but that the burden of proof is on those who want to reject them and conceive of subjects in ways that violate these intuitions. Let us call this the *epistemic primacy of human subjectivity* principle:

> *Epistemic Primacy of Human Subjectivity (EPHS)*: In discussions about the metaphysics of subjects-of-any-sort, intuitions that are based on how human subjects seem to be hold primacy over other metaphysical intuitions and possibilities that are incompatible with these intuitions, unless we have a positive conception of subjects that violate these intuitions.

In summary, the ontological unity-privacy intuitions form the bedrock of our conception of human subjects. This is the background against which the possibility of composition of subjects is rejected by James and others.

1.1 The general and the special questions

One can ask two kinds of questions about composition/combination of any entities, including subjects. First is the modal question, of whether

[4] For example, Barnett (2010, 161) takes the intuition, "Pairs of people themselves are incapable of experience" to be obvious and accepted by all almost everyone, including functionalists such as Putnam (1967); and further argues that the best explanation of this datum is that persons are simples. Barnett (2008) further demonstrates how this intuition elucidates various other intuitions in the philosophy of mind, offering a cohesive explanatory framework.

composite subjects are possible at all. Following van Inwagen (1990),[5] let us refer to this as the *general* question.

> *General subject composition question (GSCQ)*: What *is* a composite subject?

Only if one is able to answer the GSCQ in a non-circular manner—i.e. without assuming that something such as parts and wholes exist[6]—can it be claimed that composite subjects are possible.

If one assumes that composite subjects are possible, one can ask a further question: how should the parts be related such that they compose a subject? What are the mechanisms through which subject composition occurs? Let us refer to this as the special question.[7]

> *Special subject composition question (SSCQ): If* composite subjects are possible, how should micro-subjects be related so that they form a composite subject?

While one can give various answers to the SSCQ, these answers would be relevant only if subject composition is possible in the first place i.e. if we are able to define a composite subject in a non-circular manner. In this regard, the general question is more foundational than the special question.

With this distinction in place, one can ask what needs to be done to show that subject composition is possible. One needs to provide a satisfactory answer to the GSCQ, of course. How would such a response look, though? To reject the unity-privacy intuitions, it is not enough that we identify these intuitions as the basis of the problem and simply reject them. We need to take into account EPHS and show how it is possible for subjects to compose despite what we know of human subjectivity. Neither would it suffice to describe subject composition in structural terms, as the question is not one of how subjects ought to be structured for composition to occur *if* composition is possible; but what composition *is,* and whether it is possible in the first place. A structural response

[5] Van Inwagen (1990) articulated the *general* and *special* questions of composition *simpliciter,* and not specifically of subjects.

[6] Per van Inwagen, a response to the General Composition Question is “to find a sentence containing no mereological terms that was necessarily extensionally equivalent to ‘the *x*s compose y’” (1990, 39)

[7] Vaidya (2022) refers to the general and special questions pertaining to subject de-combination—the question of how a ‘big’ subject can contain within it ‘smaller’ subjects—as the *modal* and *mechanical* aspects of the problem respectively.

would be acceptable only if the proposed structural arrangement makes it transparent to us how a subject can be understood as a composite, despite the unity-privacy intuition.

I cannot think of an adequate response to the GSCQ, and this underpins my belief that subject composition is not possible. James too can be understood as claiming that there is no non-circular, coherent answer to the GSCQ, and on this basis claiming that subject combination is not possible.

It is worth noting that van Inwagen (1990) himself opined that there is no satisfactory definition of composition *simpliciter* that does not refer to mereological terms such as 'whole', 'part' or 'compose', and hence that there is no non-trivial answer to the general composition question.[8] The idea of composition itself, thus, is problematic. I briefly take this up again in § 3.1.2.

1.2 Goff

Consider Goff's (2016) version of the subject combination problem, the *no-summing-of-subjects-argument (NSS)*:

> 1. Conceptual Isolation of Subjects: For any group of subjects, instantiating certain conscious states, it is conceivable that just those subjects with those conscious states exist in the absence of any further subject.
>
> 2. Transparency Conceivability Principle: For any proposition P, if (A) P involves only quantifiers, connectives, and predicates expressing transparent concepts, and (B) P is conceivably true upon ideal reflection, then P is metaphysically possibly true.
>
> 3. Phenomenal Transparency: Phenomenal concepts are transparent.
>
> 4. Metaphysical Isolation of Subjects: For any group of subjects, instantiating certain conscious states, it is possible that just

[8] Why does a response to the special composition question not suffice as a response to the general question? Van Inwagen argues for this by showing how from two sentence of the following sort: a). "(There is a y such that the *x*s bear F to y) if and only if the *x*s are G" and b. "There is at most one y such that the *x*s bear F to y", one cannot deduce a sentence of the form "The *x*s bear F to y if and only if Φ" unless Φ contains both 'F' and the free variable 'y'" (van Inwagen 1990, 39–40).

> those subjects with those states exist in the absence of any further subject (from 1, 2, and 3).
>
> 5. For any group of subjects, those subjects with those conscious states cannot account for the existence of a further subject (from 4).
>
> 6. Therefore, panpsychism is false (from 5). (Goff 2016, 291-2)

Here, Goff partially echoes James when he says (in premise 4) that subjects are metaphysically isolated, that it is possible that *n* subjects exist without further n+1th subject existing. In contrast, Chalmers, in his *subject-summing-argument*, assumes the stronger premise that "It is *never* the case that the existence of a number of subjects with certain experiences necessitates the existence of a distinct subject" (2016, 86, emphasis added). Goff's premise 4 is weaker as it does not rule out the possibility that the n+1th could exist as a further, contingent fact of reality. Chalmers' premise, on the other hand, excludes such a possibility. Goff himself, in his (2009) seems to adopt the stronger position, saying,

> The existence of a group of subjects of experience, S1…SN, instantiating certain phenomenal characters, *never necessitates* the existence of a subject of experience T, such that what it is like to be T is different from what it is like to be any of S1…SN. (Goff 2009, 130, emphasis added)

However, he adopts the weaker premise in his (2016), and on this basis, concludes *only* that a group of subjects *cannot* account for the existence of a further subject (in premise 5).

What is the basis of Goff's move from premise 5 to 6, though? Why does *Metaphysical Isolation of Subjects* (MIS) entail the falsity of constitutive panpsychism? Such an entailment will work only if MIS entails that further subjects are not possible. If this weren't the case, MIS holds no demons–if a composite subject were possible, and if the problem were merely that we do not know the relations between subjects that lead to composition, there would be no reason to think that MIS entails the falsity of constitutive panpsychism. In other words, NSS works only if it is interpreted as arguing that we have no satisfactory response to the GSCQ (and not merely that we do not have a response to the SSCQ).

Goff seems to acknowledge that NSS is about the coherence, and hence the possibility, of composite subjects when he says, "When metaphysical possibility is so radically divorced from conceptual coherence (…) I start to lose my grip on what metaphysical possibility is supposed to be" (2016, 290). Here, he clearly recognises that it is the coherence of subject-summing (and hence the possibility) of subject-summing that is in question. This recognition is further evidenced in the line of response he adopts—he precisely questions the move from premise 4 to 5, asking why one should assume that the conceivability of n subjects existing by themselves without an n+1th subject entails that the n+1th is impossible. In other words, he claims that:

> WeakP: It is conceivable that n subjects exist without necessitating an n+1th subject

Does not entail:

> StrongP: An n+1th subject is impossible.

However, he also recognises that only StrongP entails the falsity of panpsychism, not WeakP. Given this, he simply assumes that composite subjects are possible, thus side-stepping the GSCQ and answering only the SSCQ. Goff seems to justify this move by noting that since panpsychism is otherwise theoretically desirable and hence likely true, composition of subjects *has* to be possible. This now leaves him in a position where he is "pre-theoretically committed to composite objects of some sort" (Goff 2016, 299).

Given that Goff motivates the NSS through James' articulation of the combination problem, simply claiming that WeakP does not entail StrongP and responding only to the SSCQ is too easy a move. It does not address the intuitions underlying James' articulation—unity and privacy of subjects—but simply dismisses them. If such a move were acceptable, the question of subject combination would not be a 'hard' problem in the first place.

1.3 Summary

From this discussion, we can take away two key insights regarding the context of the subject combination problem (including the NSS):

a. Given its Jamesian origin, the NSS is a 'hard' problem for panpsychists only if it is interpreted as arguing that composite subjects are impossible, and not just that we do not know the mechanisms of such composition. In other words, the relevant question posed by the subject combination problem (and NSS) is the GSCQ, not the SSCQ.

b. Any response to the GSCQ will have to take into account the unity and privacy intuitions; given EPHS, rejection of these intuitions requires us to show how subjects, as known through human subjectivity, can be non-unified-private.

2. Miller's response to Siddharth

With this background, I now consider Miller's (2022) responses to Siddharth's (2021) rejection of the phenomenal bonding solution.

2.1 Unsubstantiated intuitions

Miller (2022) points out that Siddharth's (2021) critique is based on the unity-privacy intuitions, for which Siddharth gives no justification. Miller is correct in claiming this, for Siddharth indeed does not justify these intuitions, but only notes that they underlie the subject combination problem (including the NSS). However, as noted in § 1.1, the unity-privacy intuitions are based on how human subjects—the only subjects we have direct epistemic access to—seem to be for us. James too recognises this. In light of the Jamesian origins of the subject combination problem(s) including NSS, and EPHS, it is the rejection of the unity and privacy intuitions that requires justification. The burden of proof, thus, is on the proponents of the PB solution. In the absence of such justification, Siddharth's contention that the proponents of the PB solution beg the question is correct.

2.2 The strawman fallacy

Miller (2022) contends that the proponents of the PB solution were not addressing the mereological question of the possibility of composite subjects, but only the question of whether the existence of *n* subjects can necessitate a further subject (which he identifies with the NSS).

Miller is partly correct in that the NSS, as articulated by Goff (2016), reduces the question of possibility to the question of mechanism of composition (i.e. GSCQ to SSCQ). However, as shown in section § 1.2, this is not an acceptable move; Goff (2016) himself seems to recognise that James' articulation of the subject combination problem is the question of the possibility of composite subjects (i.e. the GSCQ), and that only this question was a problem for panpsychists. It was this acknowledgment that grounded Goff's response that WeakP does not entail StrongP. However, given the Jamesian origin of the subject combination problem, Goff's response is inadequate. It is either a trivial response as it does not address the elephant in the room (the alleged impossibility of subject composition) or begs the question against the real question of subject composition.

Given this, Miller's claim that "[t]he subject summing argument is not about the incoherence of composite subjects" but "the lack of a transparent, *a priori* explanatory relationship between the fundamental level conscious facts, and the non-fundamental conscious facts" (2022, 10) does not hold water. Rather, in construing the NSS as an objection that is concerned merely with the relationship between the micro and macro conscious facts (i.e. the SSCQ) and not the question of the incoherence of composite subjects, it is Miller (and Goff 2016) who are guilty of the strawman fallacy.

My response here would be irrelevant if, as Miller (2022) claims, others (Roelofs 2019; Goff and Roelofs forthcoming; Miller 2018b) have already addressed the question of whether composite subjects are possible. I examine these views in §3, and show that contrary to what Miller claims, they do not establish the possibility of composite subjects.

2.3 Analogical extension

Siddharth (2021) had objected to Miller's (2017) proposal that the co-consciousness relation could fulfil the role of the PB relation by pointing out that co-consciousness holds between qualities and not subjects; whereas, Miller (2017) had prescribed that for a relation to be the PB relation "it must hold between subjects qua subjects of experience" (Miller 2017, 542, 546). In response, Miller (2022) clarified that this prescription requires only that subjects should be related, directly or indirectly, by the PB relation, and not that subjects qua subjects must be the relata of the PB relation. Since the qualities related by the co-consciousness relation are the qualities of the respective subjects, Miller contends that co-consciousness relation indirectly relates the subjects.

Miller's clarification entails that it is enough if the PB relation holds between subjects qua experiential qualities, and not subjects qua subjects of experience as he had originally stipulated. This change aside, would a relation that holds between subjects indirectly, by relating their qualities, suffice to form a positive conception of the PB relation? The co-consciousness relation, by relating qualities, serves to phenomenally unify them for the subject experiencing these qualities. That is, it results in the phenomenology of unity—of experiencing the two qualities together—in the subject. All examples of co-consciousness that we are aware of are between qualities experienced by the same subject. In claiming that this relation (which we know of as only holding between intra-subject qualities), in addition to unifying the phenomenal qualities can also unify subjects *qua* subjects, Miller is conflating quality combination for subject combination—unless he also believes that qualitative unity metaphysically necessitates, i.e. brings about, the ontological unity of subjects. Such a necessitation, though, must be argued for. After all, if qualitative unity alone could bring about (and suffices as an explanation for) the unity of subjects, subject-summing would not have been a problem in the first place and would follow merely from the fact that human subjects experience unified phenomenology.

Further, Miller offers introspection as a means of accessing inter-subject co-consciousness, saying:

> Non-fundamental subjects, like humans and non-human animals, are composites with large proper parts that are also subjects. These proper parts undergo a subset of the experiences of the whole. Because of this, when a human subject introspects, it is thereby introspecting inter-subjective relations, viz. the relations that hold between the subjects that compose it. (Miller 2022, 14)

His claim that we can access inter-subject co-consciousness relation through introspection would be true only if a). subject composition is possible in the first place, and b). the token qualities related by co-consciousness in a human subject's experiences are also token qualities experienced by different micro-subjects. *If* these two are assumed to be true, co-consciousness could be considered for the PB role, as part of a response to the SSCQ (even then, the entailment identified in the previous paragraph will have to be justified). However, as noted in §1, the real question about subject composition is not the SSCQ but the GSCQ. The hypothesis that co-consciousness can serve the PB role does nothing to address the GSCQ, but simply assumes that composite subjects are possible (as Miller himself admits).

In summary, we see that all three of Miller's (2022) responses to Siddharth's (2021) critique of the PB solution fail to address the real issue. I now turn to see if the mereological question has been addressed elsewhere.

3. Responses to the mereological problem: Miller, Goff, and Roelofs

Miller (2022) claims that the mereological problem—the question of whether subject composition is possible—has been addressed by Miller (2018b), Roelofs (2019), and Goff and Roelofs (forthcoming), thus paving the way for his positive conception of the PB relation. I take up each of these studies to examine if Miller's claim is correct.

3.1 Goff and Roelofs

Goff and Roelofs seek to defend the following thesis:

> **Weak Sharing (WS)**: A single experience may belong to multiple distinct subjects. (Goff and Roelofs forthcoming, 2)

They state explicitly what they mean by 'defend', saying:

> *[w]hile we cannot positively establish the possibility or actuality of mental sharing,* we hope to show that philosophers who have independent reasons to postulate it in particular cases need not hold back from doing so". (Goff and Roelofs forthcoming, 1, emphasis added)

Here, by definition, 'distinct' subjects are non-identical subjects that *overlap* (the ones that do not overlap are referred to as 'discrete'). On such an understanding, a defence of WS is a defence of the view that *if* subjects *can* overlap, a single token experience could belong to overlapping subjects.

We thus see that in defending WS, Goff and Roelofs do not want to show that overlap of subjects (and hence their composition) is possible, but only that if overlap/composition of subjects were possible then sharing of phenomenal content is also possible.[9] That is, they do not seek to answer the GSCQ, but only the SSCQ.

That the question they seek to answer is the SSCQ is reiterated, explicitly or implicitly, multiple times in the course of their argument. They identify five arguments against phenomenal sharing and respond to these.

[9] Consider another statement where Goff and Roelofs explicitly state this: "So our aim is to defend the principle of Weak Sharing*: to the extent that two subjects overlap*—one containing the other as a proper part, or both sharing a single proper part—they may share particular experiences" (forthcoming, 3, emphasis added).

In their responses, it is clear that they assume the possibility of overlapping subjects, and then go on to show that these arguments do not work. For example, their response to the *Privacy argument*—wherein critics note that experiences are accessible only by, and hence 'private' to their subjects—is to claim that privacy holds true only for discrete subjects, and not distinct subjects. This follows from the definition that distinct subjects are those that do not overlap; however, the very possibility of discrete subjects is not established. In effect, they have simply kept aside the intuition that subjects are always ontologically private, and instead adopted a weaker intuition.

As another example, consider their application of WS to panpsychism. One of the panpsychist ontologies that they think WS allows for is what they call *hybrid panpsychism*, which consists of the following two theses:

- Step 1 – It is a basic law of nature that when micro-level subjects, M1, M2…Mn, stand in certain physical relations to another, the resulting state of affairs causes a fundamental subject S to emerge, such that: (i) S is composed of all and only M1, M2…Mn, and (ii) S shares all and only the phenomenal properties of M1, M2…Mn. Call such a subject a 'basic macro-level subject'.

- Step 2 – It is a basic law of nature that when a basic macro-level subject emerges, it causes numerous other co-located subjects to emerge, such that the phenomenal properties of those subjects are grounded by subsumption in the phenomenal properties of the basic macro-level subject. (Obviously both principles leave out a lot of detail that would need to be filled in on the basis of empirical investigation). (Goff and Roelofs forthcoming, 24)

Here, they do not show that the composition of M1, M2…Mn into S is possible—they simply posit that such composition is enabled by a basic law of nature. Similarly, they do not show how it is possible for a macro-level subject to be co-located with numerous other subjects, but simply state that this is enabled by a basic law of nature.

This view is similar to the one that James considers at the end of his oft-quoted passage (see §1), and is hence open to the question he poses: what makes it the case that S is composed of M1, M2…Mn, and not a wholly distinct subject? Further, what makes it the case that the macro-level subject and other subjects that emerge (in step 2) are co-located? Even more pertinently, can a fundamental law of nature bring about something that is incoherent and unintelligible in the first place? As noted by

Siddharth (2021), such a move can be used to justify any incoherent and unintelligible relation in a brute manner.

The point, as shown in §1, is the following: given EPHS, the ontological unity and privacy theses will have to be refuted, and not merely set-aside as a matter of definition as Goff and Roelofs have done. Failing this, the very notion of composite/overlapping/co-located subjects remains incoherent and unintelligible; and any attempt to show that subjects can compose, or experiences can be shared would end up begging the question.

3.1.1. Mereological nihilism

It is noteworthy that Goff and Roelofs (forthcoming) often allude to composition of physical entities to make the case for phenomenal sharing. The assumption here is that the composition of physical entities is not a problematic notion. James himself would have disagreed—he denied the possibility of physical composition. He contended that physical entities that we take to be composites—chairs, rocks, molecules, etc.—are composites only in relation to other entities, saying:

> All the 'combinations' which we actually know are EFFECTS, wrought by the units said to be 'combined', UPON SOME ENTITY OTHER THAN THEMSELVES. Without this feature of a medium or vehicle, the notion of combination has no sense. (James 1890, 97, original emphasis)

Per James, a chair is a composite entity only to the extent that they appear as unified entities to humans. One can extend this and say that for our purposes, including our scientific practise, it makes sense to think of the chair as a single, unified, composite entity. However, to a creature that has a much, much stronger visual resolution than ours, paradigmatic solid entities (rocks, chairs, etc.) that appear to us as unified entities might appear merely as collections of multiple entities arranged in a relatively stable structure. The alleged composition of these simples, thus, is only in relation to our cognitive setup and interests.

Van Inwagen (1990), whose distinction between the general and special composition questions I earlier outlined, refers to such a view as *mereological nihilism*. He also claims that there is no non-circular response to the general composition question, which partly motivates his favourable evaluation of mereological nihilism.[10] He suggests ways in which everyday facts can be rearticulated in a nihilist-friendly language. For example, the statement 'There is a chair five metres away' ought to be understood as, 'There are simples arranged *chairwise* five metres away'. Such articulation helps us reframe truth conditions for the veracity of everyday facts, and partly alleviate the counter-intuitiveness of the view. Further, it has been shown by Brenner (2018) that composites posited in scientific theories are not indispensable and can be replaced by nihilist-friendly variations of these theories. Often, scientists do not even consider alternate articulations of their theories that do not posit composites, mostly as a matter of habit and convenience.

I note all these not to make a case for mereological nihilism,[11] but to show that the very possibility and coherence of physical composition are questionable, and not something that can be taken for granted. Given this, alluding to physical composition in support of subject composition, as Goff and Roelofs do, does little to make subject composition more acceptable.

3.2 Roelofs

Roelofs' (2019) response to critics of subject composition is similar to Goff and Roelofs' (forthcoming). He identifies the intuitions of unity (which he calls *independence*) and privacy as the basis of the subject combination problem; in response, he argues that we could adopt the weaker versions of these intuitions, that subjects are independent and private *except in the case of overlapping subjects*. Roelofs characterises experiential overlap in terms of the following two theses:

> Experience Inheritance (EI): Whenever a part of aggregate x undergoes an experience (instantiates an experiential property), x undergoes that same experience. (Roelofs 2019, 79)
>
> Micro-Unity Hypothesis (MUH): The inner nature of one, some, or all of the fundamental physical relations is phenomenal

[10] Van Inwagen, however, does not embrace full mereological nihilism; he makes an exception, claiming composition occurs only when simples are arranged to constitute a living organism. Similarly, Merricks (2001) argues that simples compose only when they form a subject of experience, and never otherwise.

[11] It has often been noted that the biggest challenge facing mereological nihilism is in accommodating human consciousness (van Inwagen 1990; Merricks 2001; Markosian 2008). Panpsychism, by positing consciousness at the fundamental level, makes it easier to accept mereological nihilism. See Kadić (2024), Siddharth and Bhojraj (forthcoming) for a defence of mereological-nihilist-panpsychism.

> unity; when two microsubjects are related in the relevant way, their experiences become unified, establishing a composite experience that subsumes them. (Roelofs 2019, 80)

These two theses tell us that *if* composite subjects exist, they derive their experiential content from the experiential content of its parts, and that microsubjects are related by some fundamental physical relation. However, by themselves, EI and MUH do not answer the question of what a composite subject *is* or whether composite/overlapping subjects are possible in the first place. In short, EI and MUH address the SSCQ, not the GSCQ.

It is noteworthy that Roelofs' account of composite subjectivity is based on a deflationary view of composition. He articulates this in terms of the following thesis:

> *Substantive Indiscernibility of Parts and Aggregate* (SI): For every property had by some part of an aggregate, that aggregate has a corresponding property, and for every property had by an aggregate, one or more of its parts have (individually or collectively) a corresponding property. (Roelofs 2019, 84)

Such a view has also been called *Composition as Identity* (CAI) by others. David Lewis (1991) characterises it as follows:

> I say that composition—the relation of part to whole, or, better, the many-one relation of many parts to their fusion—is like identity. The 'are' of composition is, so to speak, the plural form of the 'is' of identity. Call this the Thesis of Composition as Identity. (Lewis 1991, 82)

According to Lewis, the composite is not a substantial ontological addition to the world; and descriptions of the same region of the world in terms of parts or wholes are just different ways of describing the same reality. In other words, composition is *ontologically innocent*:

> Mereology is ontologically innocent. To be sure, if we accept mereology, we are committed to the existence of all manner of mereological fusions. But given a prior commitment to cats, say, a commitment to cat-fusions is not a further commitment. The fusion is nothing over and above the cats that compose it. It just is them. They just are it. Take them together or take them separately, the cats are the same portion

> of Reality either way. Commit yourself to their existence all together or one at a time, it's the same commitment either way. If you draw up an inventory of Reality according to your scheme of things, it would be double counting to list the cats and then also list their fusion. In general, if you are already committed to some things, you incur no further commitment when you affirm the existence of their fusion. The new commitment is redundant, given the old one. (Lewis 1991, 81-2)

On the basis of such a deflationary view of composition, Roelofs (2019) contends that composite subject are just structured arrangements of microsubjects. On this view, the existence of a human subject follows *a priori* from the existence of microsubjects that are arranged *humanwise*. EI thus becomes an *a priori* truth about composite experiences (see Roelofs 2019, 107–8).

The problem with CAI is that it is not clear whether it is any different from mereological nihilism. If CAI is ontologically innocent, in what way is the composite a 'real' entity, and not a mere epistemic posit? No doubt the epistemic posit holds special significance for humans; nevertheless, as James contended, such a posit is relational—in relation to us, humans. Rather than entailing mereological universalism (as Lewis contends), ontologically innocent CAI seems to entail mereological nihilism.[12] Importantly, if CAI entails mereological nihilism, and composites are ontologically innocent epistemological posits, EI cannot follow as an *a priori* truth about composite experiences, for there exist no composites in the first place.

We thus see that Roelofs fails to show that composite subjects are possible. He does not refute the unity-privacy intuitions, but simply adopts weaker versions of these intuitions. As noted earlier, the point of the subject combination problem (and the NSS) is that, given EPHS, the unity-privacy theses must be refuted and not merely set aside; any theory of subject composition that fails to address this issue is either trivial or guilty of begging the question. Further, the notion of composition that underlies Roelofs' proposal—the composition as identity view—threatens to reduce composition to an epistemological notion, and hence fails to illuminate whether composite subjects are metaphysically possible.

[12] See Calosi (2016) for a more rigorous argument for the claim that CAI entails mereological nihilism.

3.3 Miller

In his comprehensive study defending constitutive panpsychism, Miller (2018b) considers and responds to various versions of the combination problem, of which two are relevant to our purpose here: a). response to Coleman's (2014) claim that subjects are perspectival, which by definition excludes other perspectives, and b). response to Barnett's (2010) contention that persons (and subjects) are simples.

Miller identifies Coleman as claiming that a "subject's perspective is defined inclusively by what it experiences, but also exclusively by what it does not experience i.e. with an additional "to-the-exclusion-of-all-else' clause" (2018b, 120). What might ground Coleman's claim that a perspective is exclusory this way? Miller again considers various options, one of which is that Coleman assumes a "two-level account" of consciousness, according to which perspectives have "pure awareness" which "is an exclusory structural feature" (2018b, 126). The details of the two-level account of consciousness are not relevant for our purpose here; what is relevant is Miller's response to the possibility that exclusion is a structural feature of pure awareness. Consider Miller's characterisation of such a structural feature, and his reasons why it does not rule out subject composition:

> On (2) [the view that exclusion is a structural feature] the awareness itself accounts for the exclusion. It does not somehow impart a phenomenal character of exclusion to the content, but it is instead itself an exclusionary structural feature. This means that the awareness itself rules out the possibility of the scope of another awareness being wholly overlapped by it, thus ruling out proper parthood of subjects.
>
> The problem with (2) is that it is not clear as to why precisely the pure awareness is in fact exclusory. We can stipulate that a subject's perspective is defined in such a manner as to be exclusory and we can stipulate that the awareness of the two-level model accounts for this, but the explanatory relation between the two is quite opaque. What is it about the awareness itself that grounds and explains why a subject's perspective is exclusory? I have grappled with this issue and I cannot see what it is. A pure awareness must exclude other pure awarenesses as proper parts, but why?

> If (2) does not offer an (partially) illuminating explanation of exclusion, then I will take it that (2) is not responsible for exclusion. (Miller 2018b, 127-8)

This passage is illuminating, for it gets to the heart of my disagreement with Miller (and perhaps the other constitutive panpsychists). Miller is correct in his characterisation of the structural feature—subjects *qua* subjects (or pure awareness), as we intuitively understand them, are such that they exclude other subjects, thus ruling out any overlap between them. This is what I understand James as saying when he calls them the "most absolute breaches in nature". Everyone, including the constitutive panpsychists, seems to agree that subjects at least seem to be this way for us. Whether there is a further justification for this or not is a further question. Miller here then goes on to say that in the absence of such further justification—as response to the question of "What is it about the awareness itself that grounds and explains why a subject's perspective is exclusory?"—he thinks the intuition can be ignored. I, on the other hand and as noted in § 1, think that this intuition cannot be ignored. Rather, it has primacy over any other metaphysical intuition or possibility we may entertain, unless we have an independent, positive conception of a composite subject that violates these intuitions. Given the near-universal acknowledgement (even if not acceptance) of the unity-privacy intuition, and with no basis to overturn it, I do not believe that simply setting it aside is acceptable.

Miller (2018b) makes a similar move in his response to Barnett (2010), who argues that the simplicity of subjects of experience is the best explanation of the datum that it is impossible for any pair of people to be conscious. Miller responds by claiming that the simplicity of subjects is not the best explanation of the datum, and that the datum could be better explained if we accepted that: a). a pair of people do not bear the right sort of relations, such as the phenomenal bonding relation or some other physical relation *qua* their 'deep', intrinsic nature; and a pair of subjects that does bear such relations could be conscious, and b). human beings are conscious composites.

In § 2, and the preceding sub-sections of this section, we have seen that phenomenal bonding, and other relations fail to provide a positive conception of composite subjects. For this reason, option a). of Miller's response is a non-starter. More interesting is Miller's response to Barnett ruling out the possibility that humans are conscious composites. Miller accuses Barnett of assuming the following without justification:

> **Composite presentation conditional:** if something is presented to our mind as composite, then we find it absurd that it could be identical to a subject of experience. (Miller 2018b, 162)

And sets it aside, saying:

> How then can we respond? (…) [W]e can simply note that Barnett does nothing to support the absurdity claim. Granted, he gives a helpful and illustrative intuition pump, but unless one already concedes the absurdity, then it is not persuasive. In short: the absurdity in Barnett's argument is neither demonstrated or (sic) justified. (Miller 2018b, 162)

Similar to his response to Coleman (2014), Miller claims that the intuition that subjects cannot compose need not be accepted without further justification. I disagree. Given the failure of the PB and other proposed solutions, the unity-privacy intuition, and EPHS, the incoherence and impossibility of composite subjects ought to be accepted. The burden of proof is thus on those who want to claim that composite subjects are possible. In the absence of such proof, they end up begging the question.

3.4 Summary

In his response to Siddharth's (2021) critique of the PB solution, Miller (2022) claimed that the PB solution was not intended to show that composite subjects are possible, and that this possibility had been established by others (Miller 2018b; Roelofs 2019; Goff and Roelofs forthcoming). In this section, I have shown that these studies fail to show that composite subjects are possible. In such a scenario, the PB solution is either trivial, or guilty of begging the question against the real subject combination problem.

4. Rearticulating the subject combination problem

Based on the discussions in the previous sections, I propose that the subject combination problem facing panpsychism ought to be understood as the following argument:

I. *Ontological Unity-Privacy Intuition:* It seems to us that human subjects of experience, in their very being, are ontological unities such that their experiential content cannot be shared with another subject.

II. *Epistemic Primacy of Human Subjectivity (EPHS)*: In discussions about the metaphysics of subjects-of-any-sort, intuitions that are based on how human subjects *seem* to be hold primacy over other metaphysical intuitions and possibilities that are incompatible with these intuitions, unless we have a positive conception of subjects that violate these intuitions.

III. We have no transparent conception of a subject that is not ontologically unified and private.

From I, II and III,

IV. Subjects-of-any-sort (or just 'subjects') are ontologically unified and private.

Further,

V. *Composite Subject:* A composite subject is such that its subjectivity and experiential qualities are constituted by the microsubjects that compose them.

From IV and V

VI. Composite subjects are impossible, or it can never be the case that subjects compose.

Any response to the question of subject combination will have to address this argument i.e. reject at least one of I, II, III or V. Premise I seems to be acceptable to constitutive panpsychists. Premise V too is straightforward and follows from our intuitive (and circular) definition of composition. Some constitutive panpsychists (Goff 2016; Roelofs 2019; Goff and Roelofs forthcoming) can be understood as rejecting EPHS (premise II), and hence IV and VI. My response to their views in the earlier sections has been that they do not offer any justification for their rejection of EPHS; hence, their responses are either trivial (for it does not address the real issue), or guilty of begging the question. Miller (2017, 2022) can be understood as rejecting III and claiming that the co-consciousness relation fulfils the PB role, and on this basis forming a transparent conception of a subject that violates the unity-privacy intuition. My response here has been that contrary to what Miller claims, we cannot form a positive conception of an inter-subject co-consciousness relation without begging the question.

It seems to me that ultimately, one's position in this debate depends on what one thinks of the unity-privacy intuitions. Miller (2018b) is correct in noting that those who appeal to these intuitions do not say much to illuminate them. To the extent that my denial of subject composition is based on unsubstantiated intuitions, my response is also open to accusations of begging the question.

Nevertheless, I think the deniers of subject composition are on firmer ground. The unity-privacy intuitions are based on what we seem to know of the only kind of subjects we have direct access to—human subjects—and hence have priority over mere abstract possibilities. Any attempts to deny these intuitions and EPHS is faced with the question: on what basis? To me, it is not clear if entities that violate unity-privacy can even be called 'subjects'—such entities would be no different from the mysterious 'proto-phenomenal', 'neutral' quiddities[13] and non-subjective experiential qualities[14] that some Russellian monists posit. One could of course take the Kantian route and contend that our unity-privacy intuitions do not tell us anything about how subjects really are (Kant 1781/1998, A351-54). This would be acceptable if one were to claim that knowledge of the real nature of subjects are beyond our reach, not when one wants to defend the possibility of real composite subjects.

For these reasons, if one is a realist about our knowledge of subjects (and experiences), one ought to accept that composite subjects are not possible, however attractive panpsychism is independent of the subject combination problem.

Acknowledgments

I am grateful to the editor of EuJAP and the two anonymous reviewers for their useful comments on earlier versions of this paper. Parts of this paper are borrowed from my doctoral dissertation, the research for which was conducted at the National Institute of Advanced Studies, Bengaluru.

REFERENCES

Barnett, David. 2008. "The Simplicity Intuition and Its Hidden Influence on Philosophy of Mind." *Nous* 42 (2): 308–35. https://doi.org/10.1111/j.1468-0068.2008.00682.x.

[13] See Chalmers (2016a) for more on protophenomenal properties.
[14] See Coleman (2012).

———. 2010. “You Are a Simple.” In *The Waning of Materialism*, edited by Robert C Koons and George Bealer, 161–74. New York: Oxford University Press.

Brenner, Andrew. 2018. “Science and the Special Composition Question.” *Synthese* 195: 657–78.

Brüntrup, Godehard, and Ludwig Jaskolla, eds. 2016. *Panpsychism: Contemporary Perspectives*. Oxford: Oxford University Press.

Calosi, Claudio. 2016. “Composition is Identity and Mereological Nihilism.” *The Philosophical Quarterly* 66 (263): 219–35.

Chalmers, David J. 2016a. “Panpsychism and Panprotopsychism.” In *Panpsychism: Contemporary Perspectives*, edited by Godehard Brüntrup and Ludwig Jaskolla, 19–47. Oxford: Oxford University Press.

———. 2016b. “The Combination Problem for Panpsychism.” In *Panpsychism: Contemporary Perspectives*, edited by Godehard Brüntrup and Ludwig Jaskolla. Oxford: Oxford University Press.

Coleman, Sam. 2012. “Mental Chemistry: Combination for Panpsychists.” *Dialectica* 66 (1): 137–66. https://doi.org/10.1111/j.1746-8361.2012.01293.x.

———. 2014. “The Real Combination Problem: Panpsychism, Micro-Subjects, and Emergence.” *Erkenntnis* 79:19–44. https://doi.org/10.1007/s10670-013-9431-x.

Freeman, Anthony. 2006. *Consciousness and Its Place in Nature: Does Physicalism Entail Panpsychism?* Exeter: Imprint Academic.

Goff, Philip. 2009. “Can the Panpsychist Get around the Combination Problem?” In *Mind That Abides: Panpsychism in The New Millennium*, edited by David Skrbina, 129–35. Philadelphia, USA: John Benjamins Publishings.

———. 2016. “The Phenomenal Bonding Solution to the Combination Problem.” In *Panpsychism: Contemporary Perspectives*, edited by Godehard Brüntrup and Ludwig Jaskolla, 283–302. Oxford: Oxford University Press.

———. 2017. *Consciousness and Fundamental Reality*. USA: Oxford University Press.

Goff, Philip, and Luke Roelofs. forthcoming. “In Defence of Phenomenal Sharing.” In *The Phenomenology and Self-Awareness of Conscious Subjects*, edited by Julien Bugnon and Martine Nida-Rümelin. Routledge.

Inwagen, Peter van. 1990. *Material Beings*. Ithaca, N.Y: Cornell University Press.

James, William. 1890. *The Principles of Psychology*. New York: Henry Holt & Company.

Kadić, Nino. 2024. “Monadic Panpsychism.” *Synthese* 203 (38). https://doi.org/10.1007/s11229-023-04464-0.

Kant, Immanuel. 1781/1998. *Critique of Pure Reason (Translated and Edited by Paul Guyer & Allen W. Wood)*. Cambridge: Cambridge University Press.

Lewis, David. 1991. *Parts of Classes*. Oxford: Basil Blackwell.

Markosian, Ned. 2008. “Restricted Composition.” In *Contemporary Debates in Metaphysics*, edited by Theodore Sider, John Hawthorne, and Dean W. Zimmerman, 341-363. Oxford: Blackwell.

Maxwell, Grover. 1979. “Rigid Designators and Mind-Brain Identity.” *Minnesota Studies in the Philosophy of Science* 9: 9.

Merricks, Trenton. 2001. *Objects and Persons*. New York: Oxford University Press.

Miller, Gregory. 2017. “Forming a Positive Concept of the Phenomenal Bonding Relation for Constitutive Panpsychism.” *Dialectica* 71 (4): 541–62. https://doi.org/10.1111/1746-8361.12207

———. 2018a. “Can Subjects Be Proper Parts of Subjects? The De-Combination Problem.” *Ratio* 31 (2): 137–54. https://doi.org/10.1111/rati.12166

———. 2018b. “The Combination Problem for Panpsychism: A Constitutive Russellian Solution.” University of Liverpool. https://livrepository.liverpool.ac.uk/id/eprint/3030931

———. 2022. “A Reply to Siddharth’s ‘Against Phenomenal Bonding.’” *European Journal of Analytic Philosophy* 18 (1): (D1)5-18. https://doi.org/10.31820/ejap.18.1.4

Mørch, Hedda Hassel. 2014. “Panpsychism and Causation: A New Argument and a Solution to the Combination Problem.” University of Oslo.

Nagel, Thomas. 1974. “What Is It like to Be a Bat?” *Philosophical Review* 83: 435–50.

Putnam, Hilary. 1967. “The Nature of Mental States.” In *Art, Mind and Religion*, edited by W. H. Capitan and D. D. Merrill, 37–48. Pittsburgh: University of Pittsburgh Press.

Roelofs, Luke. 2019. *Combining Minds: How to Think About Composite Subjectivity*. New York: Oxford University Press.

Rosenberg, Gregg. 2004. *A Place for Consciousness*. New York: Oxford University Press.

Seager, William, ed. 2019. *The Routledge Handbook of Panpsychism*. Routledge.

Siddharth, and Tejas Bhojraj. forthcoming. “Leibnizian Panpsychism or: How I Learned to Stop Worrying and Love the Combination Problem.” *Journal of Consciousness Studies*.

Siddharth, S. 2021. “Against Phenomenal Bonding.” *European Journal of Analytic Philosophy* 17 (1): (D1)5-16. https://doi.org/10.31820/ejap.17.1.3

Skrbina, David, ed. 2009. *Mind that Abides: Panpsychism in the New Millennium*. Amsterdam: John Benjamins.

Strawson, Galen. 2006a. “Realistic Monism - Why Physicalism Entails Panpsychism.” *Journal of Consciousness Studies* 13 (10–11): 3–31.

———. 2006b. “Reply to Commentators with a Celebration of Descartes.” In *Consciousness and Its Place in Nature: Does Physicalism Entail Panpsychism?*, edited by Anthony Freeman, 184–280. Exeter: Imprint Academic.

———. 2009. *Selves: An Essay in Revisionary Metaphysics*. Oxford: Oxford University Press.

———. 2016. “Mind and Being: The Primacy of Panpsychism.” In *Panpsychism: Contemporary Perspectives*, edited by Godehard Brüntrup and Ludwig Jaskolla, 75–112. New York: Oxford University Press.

Vaidya, Anand Jayprakash. 2022. “Analytic Panpsychism and the Metaphysics of Rāmānuja’s Viśiṣṭādvaita Vedānta.” *The Monist* 105: 110–30. https://doi.org/10.1093/monist/onab026

European Journal of Analytic Philosophy EuJAP | 2024 | Vol. 20 | No. 1 | 235-254
https://doi.org/10.31820/ejap.20.1.10

IS KINDNESS A VIRTUE?

Kristján Kristjánsson[1]

[1] University of Birmingham, UK

Original scientific paper – Received: 27/02/2024 Accepted: 18/08/2024

ABSTRACT

This article swims against the stream of academic discourse by answer the title question in the negative. This contrarian answer is not meant to undermine the view that kindness is a good thing; neither is it, however, an example of a mere philosophical predilection for word play. I argue that understanding kindness as a virtue obscures rather than enlightens, for the reason that it glosses over various distinctions helping us make sense of moral language and achieving "virtue literacy". I survey some of the relevant psychological literature before moving on to philosophical sources. I subsequently delineate the alternative ways in which coherent virtue ethicists can say everything that they want to say about kindness by using much better entrenched and less bland terms. I offer a view of kindness as a cluster concept in the same sense as the Wittgensteinian concept of a game. Finally, I elicit some implications of this view for practical efforts at character education.

Keywords: virtue ethics; Aristotle; kindness; moral virtue; umbrella concept; cluster concept.

Correspondence: k.kristjansson@bham.ac.uk

1. Introduction: Umbrella concept or cluster concept?

The question of whether kindness is a virtue may seem odd. In Google, the search string “kindness a virtue” elicits 514,000 hits. A quick look at the first dozen of those indicates that most answer the question in the affirmative—albeit typically indirectly, by assuming (without argument) that kindness is indeed a virtue; subsequently employing it as an example of a paradigmatic virtue when introducing virtue ethics of a religious or secular kind. Recently, BBC Radio 4 broadcast a series of radio programmes on the virtue of kindness, accompanying a large UK national research project. Moreover, in the VIA-model, the most widely used psychological system of virtues—self-described as the “social science equivalent of virtue ethics” (Peterson and Seligman, 2004, 89)—kindness features among 24 strengths of character: more specifically as one of the three strengths (along with love and social intelligence) instantiating the overarching virtue of “humanity”.

My aim in this article is to swim against the stream and answer the question in the negative. This contrarian answer is not meant to undermine the view that kindness is a good thing; neither is it, however, an example of mere philosophical pedantry: an ill-famed professional predilection for playing with words. I will be making the substantive claim that understanding kindness as a virtue obscures rather than enlightens, for the reason that it glosses over various distinctions helping us make sense of moral language and achieving what virtue ethicists call “virtue literacy” (Jubilee Centre 2022; cf. Vasalou 2012).

To elaborate upon what I mean by the title question, it is helpful to nuance it as follows: does the term “kindness” refer to (i.e., identify, pick out) a discrete disposition that can helpfully be called a “moral virtue”? It is this specific question that I propose to address and answer in the negative. My study will be conducted to a large extent within the parameters of what is commonly referred to as Aristotelian (or neo-Aristotelian, if updated by contemporary research findings) virtue ethics. The reason for this choice is simply that Aristotle offered the most rigorous specification of moral virtue available to us, and that the majority of current Western theorising about virtues has an Aristotelian provenance. However, most of what I have to say has a wider application and will, hopefully, carry traction also for anyone interested in the contemporary discourse about (moral) virtues that takes place outside of the charmed circle of Aristotelians.

The standard historical view of moral virtues is that they constitute settled *dispositions* (acquired states of character, or *hexeis*, in Aristotle's language), concerned with excellent choices and functioning in a number of significant and distinguishable socio-moral spheres of human life that are conducive to human flourishing (Nussbaum 1988). For each virtue, the term "dispositional set" is perhaps more apt than "disposition", for each virtue is typically seen to comprise a unique set of perception/recognition, emotion, desire, motivation, behaviour, and comportment or style, applicable in the relevant sphere, where none of the factors can be evaluated in isolation. The person possessing the virtue of compassion, for example, *notices* easily situations in which a lot of others has been undeservedly compromised, *feels* for the needs of those who have suffered such undeserved misfortune, *desires* that their misfortune be reversed, *acts* (if humanly possible) for the relevant (ethical) reasons in ways conducive to that goal, and *exudes* an aura of empathy and care.

In addition to the above general conditions, Aristotle places a higher bar on a trait to constitute a moral virtue. It must 1) be driven by the right intrinsically motivating emotions;[1] 2) hit the golden mean between the extremes of excess and deficiency; 3) be performed knowingly, autonomously, and for the right reasons, overseen by the intellectual virtue of *phronesis*; 4) result from a 'firm and unchanging' state of character (see esp. Aristotle 1985, 40 [1105a30-33]); 5) include a clear behavioural component, not just a proclivity to behaviour as is the case for virtuous emotional traits. Thus, the various commendable emotional traits that Aristotle analyses in his *Rhetoric* (2007), such as compassion (*eleos*) and righteous indignation (*nemesis*), fail his strict test as full-blown virtues (see Kristjánsson 2018, ch. 1).

It could be argued that if I invoke Aristotle's strict conditions for a trait to constitute a virtue, my argument that kindness is not a moral virtue will only target straw men, as a) Aristotle himself did not designate kindness as a virtue, and b) most contemporary writings about kindness as a virtue do not apply Aristotle's criteria. Although neither a) nor b) are quite true––as a) Aristotle did discuss kindness as, at least, a virtuous emotion (see Section 3 below), and b) many philosophers who refer to kindness as a moral virtue do so from a standpoint that can only be described as Aristotelian or quasi-Aristotelian (see, e.g., McDowell 1979; Crisp 2008)

[1] The only exception to this rule are the social-glue virtues of friendliness, truthfulness about oneself, and wit in casual social encounters that people nowadays associate with manners rather than morals (Aristotle 1985, 107–114 [1126b11–1128b9]). Aristotle obviously had no specific concept of the "moral" (as distinct from the "characterological") to work with.

—I will relax some of those strict conditions. In order for my definition of virtue to fit, for instance, with Peterson and Seligman's (2004) positive psychological definition of virtues and character strengths,[2] I will leave conditions 2), 3), and 5) out of the equation. Thus, I omit the famous reference to the "golden mean", simply because it is surplus to my current requirements here, and I do not confine "moral virtue" to *phronetic* virtue, as Aristotle does in his official definition.[3] In any case, Aristotle is elsewhere happy to designate merely habituated developmental dispositions as virtues, although they have yet to become *phronesis*-infused complete virtues. Moreover, I am ready to specify the various positive emotional traits that Aristotle analyses in his *Rhetoric* (2007) as full-blown virtues, although he refrains from it there for the rather obscure reason, it seems, that they do not necessarily include an enacted behavioural element, as distinct from a behavioural proclivity (Kristjánsson 2018, ch. 1). Let me simply stipulate that my term "moral virtue" here also includes "virtuous emotions" as it does in the positive psychological system, in which various emotional traits, such as gratitude, make the grade as overarching or specific virtues. That makes my task in this article more demanding, however, because it does not allow me to reject kindness as a moral virtue for being merely a virtuous emotional disposition.

To resume the earlier thread, we saw that within virtue theories such as Aristotle's each virtue term (like "compassion") typically refers to a specific inter-connected dispositional set unique to a discrete experiential "sphere of human life" (Nussbaum 1988): say, in compassion, the sphere of undeserved misfortunes.[4] In Plato's system, but not Aristotle's, there is

[2] Their taxonomy is unusual from a philosophical perspective. They posit six overarching "virtues" and twenty-four subordinate empirically measurable "character strengths" through which the virtues are represented (Peterson and Seligman 2004, chs. 2–3). It must be admitted that the distinction between virtues and character strengths is not entirely clear; "character strengths" could just as well have been called "specific virtues", with the "virtues" understood as umbrella constructs on the understanding elaborated later in this section. In any case, all these traits would fall under Aristotle's definition of virtue as a stable character state, although Peterson and Seligman do not apply all of his strict criteria.

[3] Despite being heavily criticised for omitting the golden-mean architectonic, which creates various conceptual and moral problems (see, e.g., Ng and Tay 2020; cf. Morgan et al. 2015), I have not seen any responses from positive psychologists on why they insist that "the more is always the better" for every virtue. However, McGrath (2019) explains why positive psychology makes do without an intellectual virtue of *phronesis*. Interestingly, Peterson and Seligman (2004) retain the strict condition from Aristotle that virtues and character strengths must be intrinsically valuable: an unexpected concession given that instrumentalism about value is the dominant paradigm in psychology (Fowers 2010).

[4] Nussbaum (1996, 31) alters this to the sphere of outcomes not caused primarily by the sufferer's own culpable actions. Although that sphere does not coincide fully with the sphere of undeserved misfortunes, both presumed spheres are well circumscribed with respect to discrete experiential

one moral *master virtue* trumping the others in cases of conflict (namely, justice); in Aristotle's there is, however, an intellectual *meta-virtue* that oversees all the moral (and civic) virtues and adjudicates upon potential virtue conflicts: the above-mentioned *phronesis*. To complicate matters, Aristotle also makes space for what I call "umbrella virtues" that incorporate more than one moral virtue. Those assume two main forms. The first is that of a virtue which, while possessing some unique content of its own, also incorporates all the other moral virtues, and "does not arise without them", but "magnifies" them; this is the virtue of great-heartedness or magnanimity (*megalopsychia*) (1985, 99 [1123a1–3]).[5] Second, there are virtues that simply combine the content and moral salience of other underlying virtues without adding anything substantive to them; the prime example in Aristotle (2007) is justice (*nemesis*) as a virtuous emotion bringing together under one umbrella the four underlying virtues having to do with pleasure at deserved, and pain at undeserved, fortune or misfortune (see Kristjánsson 2006, ch. 3).[6]

In light of these complications, it is in order to extend slightly our guiding question: does the term "kindness" refer to a discrete disposition that can helpfully be called a "moral virtue", either in the specific sense of a single virtue or as an "umbrella term" referring to a unified combination of specific related virtues? Without getting ahead of my argument in Section 2, where I pinpoint the fuzziness of ordinary-language uses of "kindness" that have made their way into social scientific studies, I gather that my answer to the first part of the question will not sound unduly radical. It requires no deep scrutiny to notice that "kindness" does not designate a sphere of human experience with anywhere near the same type of specificity as, say, compassion (on either Aristotle's or Nussbaum's understanding, recall Footnote 4 above), or—to take another moral virtue, generosity: the sphere of appropriate giving. The view that kindness is an umbrella virtue, like the emotional virtue of justice (*nemesis*), sounds initially more plausible. However, my intention is to reject that part of the question also; hence, refusing kindness the label of a 'moral virtue' on either understanding.[7]

contexts. Notice that the condition about virtues referring to distinct sphere of human experience applies to all moral virtues, be those complete (*phronetic*) or still only habituated.

[5] *Megalopsychia* is only available to people with considerable material riches and certain larger-than-life personalities. However, as an enabler of great deeds, it also places psycho-moral burdens on them—to be constantly at others' beck and call—and can thus be characterised as a burdened virtue.

[6] Somewhat confusingly, Aristotle (2007, 1386b–1387a) also uses *nemesis* as a term for one of the four underlying virtues, namely pain at someone's undeserved good fortune, or what I called "righteous indignation" above.

[7] A reviewer referred me to a recent paper by David Carr (2022) on love as a non-virtue. Although love is in some ways an easier target to hit at in this sense than kindness, because of its increasingly eclectic and fuzzy uses, I think that much of what Carr argues about love applies, *mutatis mutandis*,

To anticipate, my overall view on kindness is that it is a *cluster concept* in the same sense as the Wittgensteinian concept of a game (Wittgenstein 1973). Cluster concepts distinguish themselves from umbrella concepts by not simply collating a number of related characteristics under one umbrella. The same cluster concept can refer to a number of fairly distinguishable phenomena that do share some vague similarity but cannot be easily defined or categorised as tokens of the same type. A cluster concept is specified by a weighted list of criteria, such that no one of these criteria is either necessary or sufficient for membership. Without a shared common cognitive core, what connects the criteria are family resemblances: for example, tennis as a game is connected to chess in the sense of having two players; it is, however connected to football because both are played with a ball. It is difficult to come up with a comprehensive definition of a "game", although Suits (2005) makes a healthy stab at it. Yet the possibility of a reasonable-sounding comprehensive definition of a concept *C* does not mean that *C* is not a cluster concept. Google tells us that a game is "an activity that one engages in for amusement or fun", but that definition is clearly liable to counter-examples.[8] What about professional football qua *game*; and what about the mind-games people play to manipulate one another? Contrast this with decathlon, which is a specific game/sport that shares conceptually many of the same logical/structural characteristics as Aristotle's *megalopsychia*. The term "decathlon" thus functions as a conceptual umbrella rather than a cluster concept: it incorporates other sports but synergises them in a certain way (see Kristjánsson and Fowers 2024a).

I argue that "kindness" is more akin to "game" in this respect than to "decathlon". To be sure, the word "kindness" conjures up a broad image of positive personal characteristics, but these characteristics are eclectic; they have very little in common structurally except being "morally good" in a sense that is too thin to carry weight within standard forms of virtue ethics. My core methodological assumption here is that the success criteria for an account of kindness as a virtue (in addition to tallying with some basic linguistic intuitions about the meaning of the word "kindness") are that it *either* specifies a disposition with the required specificity to constitute a single virtue—inter alia, by identifying a

to kindness: namely, that the more kindness appears to resemble a virtue, the less it looks like kindness in the ordinary sense, and vice versa.

[8] Notably, the idea of a family resemblance can of course be conveyed without the example of a game, i.e. just by pointing to the resemblance of family members sitting around the table at a typical family dinner!

distinct sphere that it is "about"—or a broader disposition that collates, and possibly synergises, a number of specific dispositions aiming cognitively at the same broad sphere but coming at it from different directions. My claim is that existing accounts of kindness fail to satisfy either of these criteria. To evidence this claim, I survey some of the relevant psychological literature in Section 2 before moving on to philosophical sources in Section 3. For those who hoped that the latter would help bring kindness talk back from its social scientific "language on holiday" (Wittgenstein 1973, §38, 232) and infuse it with conceptual rigour, Section 3 may be a disappointment. In Section 4, I address the "so-what" question, constantly hanging over conceptual studies like the sword of Damocles. I try to give a clear answer on why this analysis matters.

2. Recent psychological sources on kindness

There are abundant sources to choose from here, and I need to be selective. The most obvious place to start is with the Values-in-Action (VIA) model, as that has proved to be hugely popular with psychologists and educators since its inception (Peterson and Seligman 2004). I will leave the more general critical philosophical and psychological discourses about it[9] out of the current purview and focus solely on its inclusion and analysis of kindness.

In the VIA-model, kindness is one of three lower-order virtues appearing under the high-order virtue of *humanity*; the other two are love and social intelligence. In the chapter on kindness in the original *Handbook* (Peterson and Seligman 2004, 325–335), the title word "kindness" has "generosity, nurturance, care, compassion, altruistic love, and niceness" in brackets, apparently meaning that kindness serves as a general designator for all of them. I explained the general relationship between virtues and character strengths in positive psychology in Footnote 2 above. It is clear from Peterson and Seligman's taxonomy (2004, ch. 1) that "humanity" is an umbrella concept that is meant to cover the extensions of its three underlying strengths, including kindness. It is not

[9] For the former criticism, see Kristjánsson (2013). Regarding the latter, positive psychology is often criticised for not having published the primary data from around the world that presumably went into the creation of the original virtue taxonomy. In any case, subsequent factor analyses of millions of self-reported survey data from the VIA measure consistently fail to reproduce the original six-factor structure, but normally yield just three factors, coinciding broadly with the moral/civic, intellectual, and performative (cf. McGrath 2015). For a recent passionate defence of the VIA-model, see McGrath (2022).

as easy to decipher the relationship between kindness and all the underlying terms that are seen as instantiations of kindness. Any philosophically inclined reader will find reason to pause on the first page of this account when various statements are listed that a kind person "would strongly endorse" (Peterson and Seligman 2004, 326). Those include some uncontroversially kindness-sounding ones, such as "People in need require care", but also a more loaded statement such as "All human beings are of equal worth". That latter statement has, as far as I can see, nothing to do with kindness but all to do with respect. An elitist or a radical nationalist, who believes some people are of less worth than he is (and perhaps his fellow nationals), can still consider kindness the right attitude to treat the "inferior" people. The plot thickens further when Christian *agape* (love, charity) and Buddhist *karuna* (compassion) are introduced as being in the same "network", for those hark back to quite different world-views. When, on top of that, David Hume's sympathy is invoked by Peterson and Seligman as one more member of the kindness set, Wittgenstein's language-on-holiday complaint about social science really begins to hit home. The way all of this is formulated is that kindness constitutes a "network of closely related terms indicating a common orientation of the self toward the other" (Peterson and Seligman 2004, 326). That seems to indicate that kindness is, indeed, understood here as a cluster concept rather than an umbrella concept, although this is not made explicit.

As they come to listing possible measures of kindness, for practical purposes, the authors correctly point out that there are not many of those around. Hence, they rely on validated tests of altruism instead. However, given that standard psychological accounts of altruism in psychology typically consider moral reasoning and social responsibility among its main components (e.g., Batson et al. 1986), the awkwardness of testing kindness via altruism soon becomes apparent. Google defines altruism as "*disinterested and selfless* concern for the well-being of others" (my italics). For once, a simple dictionary definition seems to do a good philosophical job. The striking difference between an altruistic and a kind motive is that the former is disinterested but the latter is interested (i.e. emotion-imbued). The textbook (pantomime?) altruist is a Kantian who, while not motivated by other-regarding emotions, relies on a universalist principle to steer herself into helping others.[10] This is where the "social responsibility" and detached "moral reasoning" components enter in.

[10] Philips and Taylor (2009, 41) even ascribe the elision of kindness in the 19th century to the rise of Kantianism and Protestantism. For a while, they argue, kindness became the prerogative of "clergymen, romantic poets and women".

Moreover, this is precisely why kindness cannot be an umbrella concept containing altruism; the two are largely incompatible as moral characteristics. For however vague the word "kindness" is in everyday discourse, it is at least clear enough to exclude Kantian-styled altruism.[11]

Fortunately, positive psychology has moved on since 2004, and a current website (Miller 2019) presents a much more nuanced and thoughtful account of the supposed virtue of kindness. This website freely acknowledges the complications and controversies regarding a definition of "kindness". Nonetheless, it offers the following specification:

> (…) a benevolent and helpful action intentionally directed towards another person, it is motivated by the desire to help another and not to gain explicit reward or to avoid explicit punishment. (Miller 2019)

This specification serviceably seems to rule out Kantian altruism; however, it comes perilously close to equating kindness with prosociality, which is a social scientific term that virtue ethicists tend to avoid. Although I have already indicated that kindness does not lend itself to an explicit definition with necessary and sufficient conditions, any more than the concept of a game or other cluster concepts, the specification on offer here seems to clash with at least some fairly common understandings of kindness. For example: (a) Why define it as a state rather than a trait? (b) Why only "directed towards another person" but not towards animals/pets? (c) Why must it be manifested as an action? Surely, sometimes people are barred from acting on their kind motivations for various reasons (e.g., disabilities, a lack of resources); and in some cases kindness is best displayed by intentional inaction: withdrawing from a charged scene and allowing others to sort out their affairs. (d) Why must there be no expectation of a reward? What about the famous "double benefit" that is meant to be derived from young people's volunteering; does it detract from the merit of their kind acts if they are *also* motivated by the hope those will enhance their CVs for the future?

I admit that these are quick-fire responses, and that Miller's specification could possibly be amended to take account of them. However, even after such tweaks, the specification does not come anywhere close to satisfying

[11] There are other philosophical and lay uses of "altruism" that do not exclude an emotional motivation. However, the tests mentioned by Peterson and Seligman (2004) seem to have a Kantian/Kohlbergian provenance.

either success criteria for a definition of a virtue, set out in Section 1. Going from the frying pan of trying to define an ill-definable construct, the author later jumps straight into the conceptual fire by claiming that kindness and compassion are, in the end, one and the same thing. Whether one understands compassion along Aristotle's restricted lines as referring only to pain at another's *undeserved* bad fortune,[12] or makes it more inclusive by understanding it as pain at another's bad fortune *tout court* (namely, as sympathy), compassion is clearly a much narrower concept than kindness (cf. Crisp 2008, 244).

Let us now turn to a recent attempt to design a psychological instrument to measure kindness—ameliorating the previously mentioned lacuna in the psychometric field. Canter, Youngs, and Yaneva (2017) administered a 40-item self-report questionnaire to 165 people and came up with three main factors of kindness: as benign tolerance, empathic responsivity, and principled proaction. While I acknowledge the relevance of this work and all the effort that has gone into it, it is no secret that the quality of the factors elicited depends on the credibility of the original items with which participants are presented (in an exploratory factor analysis). The authors concede that the items had a varied provenance: pilot discussions, items used in previous studies, and theoretical issues identified in the literature. Some of items are bound to raise philosophical eyebrows. For example, one wonders why the item "I admit when I don't know something" (falling under "benign tolerance") should have been included in the first place. That seems to be about intellectual humility, not kindness—however broadly one understands the latter term. Similarly, "I open doors to let people through" conveys a sense of agreeableness or politeness (a distinct Aristotelian-style virtue of civility or considerateness, see Kristjánsson 2023), rather than kindness. Furthermore, most of the items falling under "principled proaction" seem to be more easily relatable to generosity (again, a clearly demarcated Aristotelian virtue) more so than kindness: for instance, "I give to charity". Without wanting to detract from the merits of this exercise, I consider the most important finding to be the authors' concession that kindness is not readily construed as a single, structured concept.

Any credible psycho-moral concept has to pass a test of developmental adequacy; we must be able to say something about how it develops and, consequently, how it can be educated. Therefore, Tina Malti's recent (2021) article on the development of kindness is potentially of great

[12] Aristotle thinks that pain at deserved bad fortune (that we would normally refer to nowadays as pity) is a vice: namely, the excess of compassion (see further in Kristjánsson 2018, ch. 4).

interest for present purposes. I learned a lot from this article, but not so much about the development of kindness specifically as about the development of a person's moral capacities in general. Malti begins with such a broad definition of kindness (as relating to "the precariousness of every human life and the beauty of imperfection" as well as entailing "feelings of respect for all others and their dignity": 2021, 630) that it is almost impossible to think of any moral developmental construct that does not fall under this specification. She divides her discussion up into explorations of kind emotions, kind cognitions, and kind behaviours. That is a helpful conceptualisation, but given the extreme permissiveness of the original definition of "kindness" (which arguably goes even further than the vagueness of ordinary language allows), we end up with a veritable smorgasbord of constructs and their developmental trajectories. Does kindness lie somehow at their intersection? I am not sure, and Malti does not persuade us that this is the case. Particularly worrying from an Aristotelian perspective is her insistence that each of the three components can be self- as well as other-oriented; thus, making much of constructs such as "self-kindness". Those will sound fairly alien to most virtue ethicists, however, be those Aristotelian or not.[13]

All in all, then, psychological studies of kindness have not succeeded in identifying a concept of kindness with a clear common core, nor have they made a strong case for kindness as a helpful umbrella concept, approaching a common core from different directions. Indeed, psychologists have not made much progress in tidying up the vagaries of ordinary language. Yet it is clear that their intention is to conceptualise, operationalise, and measure a lay concept of kindness as a virtue, and they frequently use the "virtue" word. The image of kindness that emerges from contemporary psychology is, however, far from that of either a specific virtue or a discrete umbrella-like virtue trait.

3. Some philosophical sources on kindness

Given that some of Wittgenstein's haughtiness towards social science seems to ring true in the case of kindness, can philosophers do any better? Obviously, for virtue ethicists, at least of Aristotelian or quasi-Aristotelian persuasion, the natural entry point will be in Aristotle's own texts. Kindness does not emerge in the (non-exhaustive) list of virtues in

[13] Cf., e.g., Peter Geach's well-known remarks about self-love as a potential virtue: "A man's self-concern is unworthy of the name of love: and if it were love, the man who thinks he is trying to extend that sort of personal interest even to all the other persons he knows will pretty certainly be kidding himself" (1977, 74).

the *Nicomachean Ethics*. It does, however, appear in the analyses of various virtuous emotions in the *Rhetoric*. I have already dismissed, albeit cursorily, Aristotle's own misgivings about considering those as full-blown virtues, so we may appear initially to have hit the jackpot here. Indeed, many of the things Aristotle says about kindness in his uncharacteristically quick treatment (2007, 137–139 [1385a16–1385b11]) seem to give succour to the idea that kindness does indeed fit into the architectonic of a virtue. It is defined as helpfulness towards someone in need, not in return for anything, nor for the advantage of the helper himself, but for that of the person helped. Although the analysis is elliptical in an Aristotelian sense in that the excess and deficiency forms are not enlisted, it does not seem to be a tall order to add the missing bits and pieces.

Unfortunately, this impression is illusory. Although the standard translation of the emotional virtue explored here, *kharis* in Greek, is "kindness" or "kindliness" (see, e.g., the 2007 translation, while it rightly notes that the word has various meanings, p. 137), David Konstan (2006, ch. 7) has argued persuasively that the specific meaning of *kharis* in the *Rhetoric* is the inclination to return favours received, namely gratitude. Indeed, Aristotle is not analysing the emotion of *kharis* here at all, but rather *ekhô kharin*: the kindly feeling one experiences when receiving a gift. It is no wonder, then, that Aristotle is quick to deliver his account of this virtuous emotion as a discrete one, for gratitude constitutes a fairly specific state and trait with clear cognitive and motivational components (see Kristjánsson 2018, ch. 3). Aristotle, however, offers no help to us in specifying kindness—on contemporary understandings—as a virtue.

Despite kindness appearing in almost uncountable (in Google Scholar) philosophical writings about virtues and virtue ethics,[14] I tried hard but failed to identify a single philosophical article that sets out to define kindness explicitly as an Aristotelian moral virtue, with its standard components and parameters, although many seem to assume an Aristotelian architectonic of virtue implicitly (see, e.g., Crisp 2008). The closest I came to an explicit understanding of kindness as an Aristotelian virtue was an article by John McDowell (1979), a classic and much-quoted one. Although the article is not cited mainly for its focus on kindness, but rather for its account of the uncodifiability of virtues in general, McDowell takes kindness throughout as the paradigmatic example of a moral virtue to which his general account will then apply.

[14] The search term "kindness AND virtue AND philosophy" elicit 289,000 hits.

One will look in vain for a clear specification of kindness in this article. Yet McDowell says about kindness that the

> (…) kind person has a reliable sensitivity to a certain sort of requirement which situations impose on behaviour. The deliverances of reliable sensitivity are cases of knowledge and (…) a kind person knows what it is like to be confronted with a requirement of kindness (McDowell 1979, 331–332).

The problem is that McDowell does not specify what exactly is "specialised" in those specialised sensitivities towards kindness, although he later says those have to do with "proper attentiveness to others' feelings" (1979, 333). But then, again, what counts as "proper" here? McDowell seems simply to have chosen kindness in this article as an illustration, because of its prevalence in ordinary language, without taking account of the fact that kindness is not a good example of the kinds of virtues whose incarnations flower in Aristotle's virtue ethics.

Quite a different take on kindness can be found in Alan Wilson's (2017) article on how to avoid the conflation of moral and intellectual virtues. As with McDowell, kindness is not the main topic of Wilson's article. However, it enters his argument in a way that is highly pertinent for present purposes. Wilson tries to contrive a way out of the conundrum of how to distinguish systematically between moral and intellectual virtues when the dividing line between them seems to be thin. When exactly, for example, is honesty an intellectual and when a moral virtue? Wilson's solution is motivation-based: intellectual virtues can be identified by their shared motivation for cognitive contact with reality whereas moral virtues are identified by the characteristic motivations of justice and kindness.

Wilson's solution, while ingenious, is outside of the present purview. What matters is his definition of kindness as a broad motivation to protect and promote (others') well-being. I think he hits the nail on the head to understand kindness as a broad motivation of this kind.[15] Far from being antithetical to my view of kindness as a cluster concept, Wilson's characterisation actually supports it. Understood as a broad motivation,

[15] It could be argued that not all kindness is even virtue ethically relevant at all; consider "light-weight" forms of kindness such as smiling kindly at the shopkeeper, which seem to have to do with good manners rather than morals. However, interestingly enough, Aristotle failed to make a distinction between manners and morals, and considered friendliness in casual social interactions to be morally (i.e., characterologically) virtuous, even if not accompanied by any underlying virtuous emotions (Aristotle 1985, 107–108 [1126b11-29]).

kindness attached itself to various attitudes, virtues, beliefs, and gestures—just as the broad motivation to have fun attaches itself to various rituals and practices that we call “games”. We refer to the plethora of these kindness-as-a-motivation-attached phenomena as being “kind”. There is nothing wrong with that usage. However, these items are too varied to fall under a single umbrella of an overarching virtue that we could call “kindness”.[16] Rather, they are part of a cluster concept whose items are connected by their vague family resemblance of being similarly motivated, while otherwise having very little in common:[17] compare, say, giving a large chunk of your income to charity versus holding the elevator door open for an arriving person in a department store. Wilson carefully explains how a virtue such as compassion counts as a moral virtue in the first place because of its containing the overarching motivation of kindness, but he avoids positing a distinct sphere of human activity (in Nussbaum’s 1988 sense) to which a moral virtue of kindness uniquely refers—which is just as good, because it would be impossible to identify such a sphere.

4. Why this matters: Educational ramifications

Virtue ethics is perhaps most influential these days in its practical incarnation as character education, both within schools and professional ethics education (Jubilee Centre 2022). This extension of virtue ethics is very much in line with Aristotle’s contention that the purpose of moral inquiry “is not to know what virtue is, but to become good, since otherwise the inquiry would be of no benefit to us” (1985, 35 [1103b27-29]). However, as inconvenient as a fuzzy definition of a virtue term is for philosophical and psychological studies, it is virtually devastating for

[16] Notice that sharing the motivation of kindness does not come anywhere close to the criterion of an umbrella concept of having a “shared common cognitive core”. If it did, all the moral virtues would simply be instances of one umbrella virtue. However, that is not what Wilson (2017) is arguing. The relationship of generosity and compassion—although having a shared motivation of kindness driving them—is much closer to that of tennis and chess (which share the motivation of wanting to play) than that of, say, righteous indignation and satisfied indignation (which share the cognitive content of aiming towards deservingness). Wilson’s account, as I understand it, is therefore not to be best interpreted as an argument for kindness as an umbrella concept. That said, Wilson does refer casually in his article to kindness as “a virtue”, without any argument, perhaps simply relying on the received wisdom from ordinary language. He did the same in an earlier article (Wilson 2016).

[17] It could be argued that, on this understanding, kindness is not a true cluster concept because acts of kindness have one necessary feature in common, unifying everything in the set: namely motivation. However, the fact that traits *x*, *y*, and *z* share the same motivation does not establish either that they are about the same sphere or that they all fall under the same umbrella concept. Analogously, in a way, all moral traits are motivated by a concern for what psychologists would call “prosociality”, but it would not be helpful to claim that they are all therefore instantiations of a single umbrella concept; consider, e.g., moral traits as distinct as (proper) pride and (proper) compassion.

the process of carving out character educational interventions. For those to work, we need to be pedantically clear about what sort of concept we are working with, how that refers to a specific psycho-moral quality, and what strategies are most effective in cultivating this quality in classroom contexts. Kindness is, I argue, particularly badly fit for that purpose.

Recall some of the specific moral virtues and virtuous emotions that Aristotle demarcates in the *Nicomachean Ethics* and the *Rhetoric*: compassion, generosity, agreeableness, friendship. These are specified with meticulous precision, the main focus being on their cognitive content: what they are *about*. What is more, Aristotle provides systematic advice about how to educate them as virtues. Although he is much more detailed on the early stages of that process, where the cultivation takes place through emotional contagion/sensitisation, social osmosis, emulation of moral exemplars, and habituation (learning by doing), he also gives clues about how to infuse those virtues with *phronesis* at a later stage, once the soul of the student is prepared for metacognitive intellectual pursuits. It is no wonder that the most advanced theories of character education in modernity have done little more than systematise Aristotle's account as that of "caught", "taught", and "sought" method of character cultivation and bring it up to date with empirical evidence (Jubilee Centre 2022).

To be sure, the four virtues that I mentioned above could all be called "kind", but it does not add anything to an educational account of those well-entrenched dispositions to try to educate them together under one label of "kindness"—let alone add kindness to them as a discrete additional virtue. Quite the opposite, it simply waters down the educational content. Admittedly, virtues "hunt in packs", as it is often put, and they form various conceptual and substantive alliances (see, e.g., Gulliford and Roberts 2018). However, there is no convenient conceptual umbrella bringing all "kind" virtues together; they serve very different purposes in the moral landscape although they share a vague common moral motivation.

I am not saying that we should expunge the term "kindness" from our general moral or educational vocabularies, any more than we should get rid of the word "game". However, I doubt that the term is salient within (broadly) Aristotelian virtue ethics in carrying significant substantive, explanatory, or developmental weight. I worry that invoking it in virtue talk may undermine rather than enhance virtue literacy.[18] Educationally, there is also much less to learn from Aristotle about the cultivation of general moral motivations. Even after *phronesis* has developed, helping us to adjudicate upon virtue conflicts between, for instance, honesty and compassion, the primary moral motivation continues to stem from the relevant discrete virtues (Kristjánsson and Fowers 2024b). *Phronesis* may furnish us with a more general motivation to be good persons, committed to *eudaimonia*, but Aristotle is very cagey about how that general blueprint-of-the-good-life-forming motivation emerges, except noting that it is not inborn (although the capacity to develop it is) and that it forms only if we are brought up "in good habits" (Aristotle 1985, 6 [1095b4–5]).

Mutatis mutandis, if Aristotle had written about a general motivation to be kind, he would probably have been equally reticent about it. As practically minded as he was, he was mainly interested in the discrete character traits that can be inculcated, honed, and later sought and revised by the students themselves. We know that he was very pessimistic—perhaps unduly so—about radical moral conversions later in life, and seemed to believe that the general foundations of what is nowadays referred to as "moral identity" are mostly the result of the ethical environment which nurtures us, and hence deeply susceptible to moral luck. To be sure, among the "caught" methods that Aristotle mentions as forming the core of character education is the emulation of moral exemplars; so one could envisage an Aristotle-inspired character intervention focused on getting students to read about exemplars of kindness, reflect on how such folks might behave in their circumstances, and try to emulate them. Yet, in his talk about emulation, Aristotle reminds us not to copy the emulated person *qua* person, but rather to understand and emulate the specific virtuous traits that she represents (see various references in Kristjánsson 2007, ch. 7). That cannot be done without a clear grasp of the relevant virtue; and if there is no discrete virtue of kindness, as I have argued, we might end up with the counter-productive consequences that are likely to ensue when teachers try to develop positive traits in students indiscriminately and without the necessary conceptual nuance (Morgan et al. 2015). So, while there are surely some valid ways in which the meaning of cluster concepts such as kindness can be conveyed to moral learners, for instance through "caught" methods of language osmosis, they would never, on an

[18] For a spirited defence of the importance of coherent virtue language for the development of virtue, see Vasalou (2012).

Aristotelian account, be accorded the same priority as that of more discrete concepts referring to discrete or umbrella-like moral virtues.

Of course, there is no reason for contemporary virtue ethicists and character educators to take Aristotle as the last word on those issues (see Kristjánsson 2020, ch. 6). As a die-hard methodological naturalist, he would encourage us to revise his theories in light of new empirical findings. However, as the tenor of my above argument suggests, I am not sanguine about the possibility of some sort of retrieval of a virtue of kindness being able to aid us in those revisionary endeavours.

5. Concluding remark

One of Lord Rutherford's famous aphorisms is that all academic work is either science or stamp collecting. Conceptual analysis, as conducted above, can either aim at "carving nature at its joints" or arranging a "stamp collection" in a more orderly and systematic fashion. I have only aimed at the latter here. Unfortunately, conceptual analysis has fallen out of favour of late in analytic philosophy. Apart from making a substantive point about what kindness is—in the sense of being "best understood as"—in this article, I hope to have demonstrated that even "stamp arrangement" of this sort does have practical reverberations. Some arrangements are, for example, educationally productive but others much less so. I have argued that understanding kindness as a discrete moral virtue falls into the latter category. That is one of the reasons, albeit not the only one, for rejecting the view that kindness is a moral virtue.

The strongest counter-argument to the rejection of kindness as a moral virtue, as set out in this article, would be to attack the "success criteria" for an account of a disposition to count as a moral virtue set out at the end of Section 1. For example, is it necessary for an account of kindness to match our intuitions about the concept, or could a radically revisionary account of kindness (on which kindness is still a virtue) meet the challenge posed here? Moreover, if an account does need to match certain intuitions, is it obvious which intuitions we should appeal to? There are points in this article where I assumed the credibility of intuitions such as that our use of the word kindness is at least clear enough to exclude Kantian altruism and that compassion is clearly a much narrower concept than kindness. These assumptions could be questioned. To anticipate and resist such a possible counter-argument would require a much longer venture into the methodology of conceptual analyses than I have space for here. It suffices to repeat at this final stage the Aristotelian point that one of the most important aims of virtue talk is to make substantive claims relevant to moral development and moral education. It is difficult to envisage how divorcing an account of kindness as a virtue from intuitions embedded in ordinary language would further those essentially practical aims.

Acknowledgments

Thanks to two anonymous reviewers for their useful comments on earlier versions of this paper.

REFERENCES

Aristotle. 1985. *Nicomachean Ethics*, trans. T. Irwin. Indianapolis: Hackett Publishing.

Aristotle. 2007. *On Rhetoric*, trans. G. A. Kennedy. Oxford: Oxford University Press.

Batson, C. Daniel, Michelle H. Bolen, Julie A. Cross, and Helen E. Neuringer-Benefiel. 1986. “Where Is the Altruism in the Altruistic Personality?” *Journal of Personality and Social Psychology* 50 (1): 212–220. https://doi.org/10.1037/0022-3514.50.1.212.

Canter, David, Donna Youngs, and Miroslava Yaneva. 2017. “Towards a Measure of Kindness: An Exploration of a Neglected Interpersonal Trait.” *Personality and Individual Differences* 106 (1): 15–20. https://doi.org/10.1016/j.paid.2016.10.019.

Carr, David. 2022. “The Moral Status of Love.” *International Philosophical Quarterly* 1 (245): 99–113. https://doi.org/10.5840/ipq2022621193.

Crisp, Roger. 2008. “Compassion and Beyond.” *Ethical Theory and Moral Practice* 11 (3): 233–246. https://doi.org/10.1007/s10677-008-9114-x.

Fowers, Blaine J. 2010. “Instrumentalism and Psychology: Beyond Using and Being Used.” *Theory and Psychology* 20 (1): 102–124. https://doi.org/10.1177/0959354309346080.

Geach, Peter. 1977. *The Virtues*. Cambridge: Cambridge University Press.

Gulliford, Liz, and Robert C. Roberts. 2018. “Exploring the “Unity” of the Virtues: The Case of an Allocentric Quintet.” *Theory and Psychology* 28 (2): 208–226. https://doi.org/10.1177/0959354317751666.

Jubilee Centre for Character and Virtues. 2022. *A Framework for Character Education in Schools*. Accessed August 14, 2024.

https://www.jubileecentre.ac.uk/wp-content/uploads/2023/07/Framework-for-Character-Education.pdf

Konstan, David. 2006. *The Emotions of the Ancient Greeks: Studies in Aristotle and Classical Literature*. Toronto: University of Toronto Press.

Kristjánsson, Kristján. 2006. *Justice and Desert-Based Emotions*. Aldershot: Ashgate/Routledge.

———. 2007. *Aristotle, Emotions, and Education*. Aldershot: Ashgate/Routledge.

———. 2013. *Virtues and Vices in Positive Psychology: A Philosophical Critique*. Cambridge: Cambridge University Press.

———. 2018. *Virtuous Emotions*. Oxford: Oxford University Press.

———. 2020. *Flourishing as the Aim of Education: A Neo-Aristotelian view*. London: Routledge.

———. 2023. "Considerateness Differentiated: Three Types of Virtuousness." *Journal of the American Philosophical Association*, in press. https://doi:10.1017/apa.2023.22.

Kristjánsson, Kristján, and Blaine J. Fowers. 2024a. "*Phronesis* as Moral Decathlon: Contesting the Redundancy Thesis about *Phronesis*." *Philosophical Psychology* 37 (2): 279–298. https://doi.org/10.1080/09515089.2022.2055537.

———. 2024b. *Phronesis: Retrieving Practical Wisdom in Psychology, Philosophy, and Education*. Oxford: Oxford University Press.

Malti, Tina. 2021. "Kindness: A Perspective from Developmental Psychology." *European Journal of Developmental Psychology* 18 (5): 629–657. https://doi.org/10.1080/17405629.2020.1837617.

McDowell, John. 1979. "Virtue and Reason." *Monist* 62 (3): 331–350. https://doi.org/10.5840/monist197962319.

McGrath, Robert E. 2015. "Character Strengths in 75 Nations: An Update." *Journal of Positive Psychology* 10 (1): 41–52. https://doi.org/10.1080/17439760.2014.888580.

———. 2019. "Refining our Understanding of the VIA Classification: Reflection on Papers by Han, Miller, and Snow." *Journal of Positive Psychology* 14 (1): 41–50. https://doi.org/10.1080/17439760.2018.1528382.

———. 2022. "The VIA Virtue Model: Half-baked or Brilliant?" *Journal of Positive Psychology* 17 (2): 250–256. https://doi.org/10.1080/17439760.2021.2016905.

Miller, Kori D. 2019. "What is Kindness in Psychology?" Accessed August 14, 2024. https://positivepsychology.com/character-strength-kindness/

Morgan, Blaire, Liz Gulliford, and David Carr. 2015. "Educating Gratitude: Some Conceptual and Moral Misgivings." *Journal of Moral Education* 44 (1): 97–111. https://doi.org/10.1080/03057240.2014.1002461.

Ng, Vincent, and Louis Tay. 2020. "Lost in Translation: The Construct Representation of Character Virtues." *Perspectives on Psychological Science* 15 (2): 300–326. https://doi.org/10.1177/1745691619886014.

Nussbaum, Martha C. 1988. "Non-Relative Virtues: An Aristotelian Approach." *Midwest Studies in Philosophy* 13 (1): 32–53. https://doi.org/10.1111/j.1475-4975.1988.tb00111.x.

———. 1996. "Compassion: The Basic Social Emotion." *Social Philosophy and Policy 13* (1): 27–58. https://doi.org/10.1111/j.1475-4975.1988.tb00111.x.

Peterson, Christopher, and Martin E. P. Seligman. 2004. *Character Strengths and Virtues: A Handbook and Classification*. Oxford: Oxford University Press.

Philips, Adam, and Barbara Taylor. 2009. *On Kindness*. London: Penguin.

Suits, Bernard. 2005. *The Grasshopper: Games, Life, and Utopia*. Ontario: Broadview Press.

Vasalou, Sophia. 2012. "Educating Virtue as a Mastery of Language." *Journal of Ethics* 16 (1): 67–87. https://doi.org/10.1007/s10892-011-9111-5.

Wilson, Alan T. 2016. "Modesty as Kindness." *Ratio* 29 (1): 74–88. https://doi.org/10.1111/rati.12045.

———. 2017. "Avoiding the Conflation of Moral and Intellectual Virtues." *Ethical Theory and Moral Practice* 20 (5): 1037–1050. https://doi.org/10.1007/s10677-017-9843-9.

Wittgenstein, Ludwig. 1973. *Philosophical Investigations*, trans. G. E. M. Anscombe. New York: Prentice-Hall.

ABSTRACTS (SAŽECI)

Arbiters of Truth and Existence

Nathaniel Gan
National University of Singapore, Singapore

ABSTRACT

Call the epistemological grounds on which we rationally should determine our ontological (or alethiological) commitments regarding an entity its arbiter of existence (or arbiter of truth). It is commonly thought that arbiters of existence and truth can be provided by our practices. This paper argues that such views have several implications: (1) the relation of arbiters to our metaphysical commitments consists in indispensability, (2) realist views about a kind of entity should take the kinds of practices providing that entity's arbiters to align with respect to their metaphysical dependencies, (3) if realists take a kind of practice to provide grounds on which to affirm the existence of a kind of entity, they should turn to those same grounds when seeking to provide an epistemology of the relevant domain.

Keywords: naturalism; Carnapian realism; indispensability arguments; epistemic problems.

Arbitri istine i postojanja

Nathaniel Gan
National University of Singapore, Singapore

SAŽETAK

Nazovi epistemološke osnove na kojima bismo racionalno trebali odrediti naše ontološke (ili aletiološke) obveze u vezi s entitetom njegov arbitar postojanja (ili arbitar istine). Uobičajeno je mišljenje da arbitri postojanja i istine mogu biti dani putem naših praksi. Ovaj rad tvrdi da takva gledišta imaju nekoliko implikacija: (1) veza između arbitara i naših metafizičkih obveza sastoji se u neophodnosti, (2) realistička gledišta o vrsti entiteta trebala bi se podudarati s vrstama praksi koje pružaju arbitre za tu vrstu entiteta s obzirom na njihove metafizičke ovisnosti, (3) ako realisti smatraju da vrsta prakse pruža temelje na kojima se potvrđuje postojanje vrste entiteta, trebali bi se referirati na iste temelje koje koriste kada pokušavaju pružiti epistemologiju relevantnog područja.

Ključne riječi: naturalism; Carnapovski realizam; argument neizbježivosti; epistemički problemi.

Is L.A. Paul's Essentialism Really Deeper than Lewis's?

Cristina Nencha
University of Bologna and University of Bergamo, Italy

ABSTRACT

L.A. Paul calls "deep" the kind of essentialism according to which the essential properties of objects are determined independently of the context. Deep essentialism opposes "shallow essentialism", of which David Lewis is said to be a prominent advocate. Paul argues that standard forms of deep essentialism face a range of issues (mainly based on an interpretation of Quinean skepticism) that shallow essentialism does not. However, Paul claims, shallow essentialism eliminates the very heart of what motivates essentialism, so it is better to be deep than shallow. Accordingly, she proposes a very sharp novel account of essentialism, which, while attempting to preserve some of the advantages of shallow essentialism over the classical forms of deep essentialism, can be deemed to be deep. In this paper, I compare Paul's proposal for a kind of deep essentialism with Lewis's account, as it is presented by Paul. My aim is to show that the differences between the two approaches are not as significant as Paul takes them to be, and that Paul's account can be taken to be deeper than Lewis's only at the cost of sacrificing the very idea at the bottom of deep essentialism.

This might be taken to suggest that, if Paul is correct in asserting that shallow essentialism is better equipped to address some skeptical challenges, but it is generally preferable to be deep than shallow, then Lewis's account should be re-evaluated, since, as shallow as it can be, it might be deeper than it looks.

Keywords: David Lewis; essentialism; L.A. Paul; context-sensitivity.

Je li esencijalizam L.A. Paul zaista dublji od Lewisovog?

Cristina Nencha
University of Bologna and University of Bergamo, Italy

SAŽETAK

L.A. Paul naziva "dubokim" onaj tip esencijalizma prema kojem su bitne osobine predmeta određene neovisno o kontekstu. Duboki esencijalizam suprotstavlja se "plitkom esencijalizmu", te se smatra da ga je zastupao David Lewis. Paul tvrdi da, za razliku od plitkog esencijalizma, standardni oblici dubokog esencjializma se sučavaju s nizom problema (uglavnom se temelje na određenoj interpretaciji Quineanskog skepticizma). Međutim, Paul tvrdi da plitki esencijalizam eliminira samu srž onoga što motivira esencijalizam što ga čini manje privlačnim od dubokog. U skladu s tim, predlaže vrlo oštru novu teoriju esencijalizma koja, iako zadržava neke od prednosti plitkog esencijalizma nad klasičnim oblicima dubokog esencijalizma, može se smatrati dubokim.
U ovom radu, uspoređujem Paulinu varijantu dubokog esencijalizma s Lewisovim opisom, kako ga predstavlja Paul. Moj cilj je pokazati da, unatoč Paulinom mišljenju, razlike između ta dva pristupa nisu toliko značajne, te da se Paulin opis može smatrati dubljim od Lewisovog samo uz žrtvovanje same ideje na kojoj počiva duboki esencijalizam. To bi se moglo shvatiti kao sugestija da, ako je Paul u pravu kada tvrdi da plitki esencijalizam može bolje odgovoriti na neke skeptične izazove, ali je općenito poželjnije biti dubok nego plitak, tada bi Lewisova teorija trebala biti ponovno procijenjen, jer, koliko god plitak bio, možda je dublji nego što izgleda.

Ključne riječi: David Lewis; esencijalizam; L.A. Paul; osjetljivost na kontekst.

The Logical Possibility of Moral Dilemmas in Expressivist Semantics: A Case Study

Ryo Tanaka
University of Tokyo, Japan

ABSTRACT

In this paper, using Mark Schroeder’s (2008a) expressivist semantic framework for normative language as a case study, I will identify difficulties that even an expressivist semantic theory capable of addressing the Frege-Geach problem will encounter in handling the logical possibility of moral dilemmas. To this end, I will draw on a classical puzzle formulated by McConnell (1978) that the logical possibility of moral dilemmas conflicts with some of the prima facie plausible axioms of the standard deontic logic, which include obligation implies permission. On the tentative assumption that proponents of ethical expressivism should be generally committed to securing the logical possibility of moral dilemmas in their semantic theories, I will explore whether and how expressivists can successfully invalidate obligation implies permission within the framework developed by Schroeder. The case study eventually reveals that this can indeed be a hard task for expressivists. Generalizing from the case study, I will suggest that the source of the difficulty ultimately lies in the mentalist assumption of the expressivist semantic project that the logico-semantic relations exhibited by normative sentences should be modeled in terms of the psychological attitudes that speakers express by uttering them. My final goal will be to show that the difficulty expressivists face in dealing with the logical possibility of moral dilemmas is a reflection of the more general problem that their commitment to the mentalist assumption prevents them from flexibly adopting or dropping axioms in their semantic theories to get the right technical results.

Keywords: expressivism; moral dilemmas; metaethics; semantics; deontic logic.

Logička mogućnost moralnih dilema u ekspresivističkoj semantici: Studija slučaja

Ryo Tanaka
University of Tokyo, Japan

SAŽETAK

U ovom radu, koristeći Mark Schroederov (2008a) semantički okvir za ekspresivistički normativni jezik kao studiju slučaja, identificirat ću poteškoće s kojima će se čak i ekspresivistička semantička teorija sposobna za rješavanje Frege-Geach problema susresti pri objašnjenju logičke mogućnosti moralnih dilema. U tu svrhu, oslonit ću se na klasičnu zagonetku koju je formulirao McConnell (1978)a pokazuje da se logička mogućnost moralnih dilema sukobljava s nekim od naizgled opravdanih aksioma standardne deontičke logike, među kojima je i aksiom da obaveza implicira dopuštenost. Na temelju tentativne pretpostavke da zastupnici etičkog ekspresivizma trebaju općenito biti posvećeni osiguravanju logičke mogućnosti moralnih dilema u svojim semantičkim teorijama, istražit ću da li i kako ekspresivisti mogu uspješno opovrgnuti aksiom da obaveza implicira dopuštenje unutar okvira koji je razvio Schroeder. Studija slučaja konačno otkriva da to može biti zaista težak zadatak za ekspresiviste. Generalizirajući iz studije slučaja, sugerirat ću da izvor poteškoće leži u mentalističkoj pretpostavci ekspresivističkog semantičkog projekta da bi logičko-semantički odnosi prikazani normativnim rečenicama trebali biti modelirani pomoću psiholoških stavova koje govornici izražavaju izgovarajući ih. Moj konačni cilj će biti pokazati da je poteškoća s kojom se ekspresivisti susreću u suočavanju s logičkom mogućnosti moralnih dilema odraz općeg problema da njihova predanost mentalističkoj pretpostavci sprječava fleksibilno usvajanje ili odbacivanje aksioma u njihovim semantičkim teorijama kako bi dobili ispravne tehničke rezultate.

Ključne riječi: ekspresivizam; moralne dileme; metaetika; semantika; deontička logika.

Two Problems About Moral Responsibility in The Context of Addiction

Federico Burdman
Alberto Hurtado University, Chile

ABSTRACT

Can addiction be credibly invoked as an excuse for moral harms secondary to particular decisions to use drugs? This question raises two distinct sets of issues. First, there is the question of whether addiction is the sort of consideration that could, given suitable assumptions about the details of the case, excuse or mitigate moral blameworthiness. Most discussions of addiction and moral responsibility have focused on this question, and many have argued that addiction excuses. Here I articulate what I take to be the best argument for this view, based on the substantial difficulty that people with severe addiction experience in controlling drug-related behavior. This, I argue, may in some cases be sufficient to ground a mitigating excuse, given the way in which addiction undermines agents’ responsiveness to relevant moral reasons to do otherwise. Much less attention has been devoted to a second set of issues that critically affect the possibility of applying this mitigating excuse in particular cases, derived from the ambivalent nature of agential control in addiction. In order to find a fitting response to moral harm, the person with the right standing to blame must make a judgment about the extent to which the agent possessed certain morally relevant capacities at the time of the act.

In practice, this will often prove tremendously difficult to assess. The ethical challenge for the person with the right standing to blame is fundamentally one of making a judgment about matters that seem underdetermined by the available evidence.

Keywords: addiction; moral responsibility; behavioral control; mitigation; degrees of blameworthiness.

Dva problema moralne odgovornosti u kontekstu ovisnosti

Federico Burdman
Alberto Hurtado University, Chile

SAŽETAK

Predstavlja li ovisnost uvjerljiv izgovor za moralne štete koje rezultiraju iz odluke da se koristi droga? Ovo pitanje upućuje na dva različita skupa problema. Prvo, postavlja se pitanje je li ovisnost vrsta razmatranja koja bi, uz odgovarajuće pretpostavke o pojedinostima slučaja, mogla opravdati ili ublažiti moralnu krivnju. Većina se rasprava o ovisnosti i moralnoj odgovornosti fokusirala na ovo pitanje, te su mnogi tvrdili da je ovisnost ispričavajuća. Ovdje artikuliram ono što smatram najboljim argumentom za ovo gledište, a temelji se na značajnoj teškoći koju ljudi s ozbiljnom ovisnošću doživljavaju pri kontroliranju ponašanja povezanog s drogama. Tvrdim da ovo u nekim slučajevima može biti dovoljno da služi kao olakotna okolnost, s obzirom na način na koji ovisnost umanjuje djelatnikovu prijemčivost na relevantne moralne razloge za postupanje drugačije. Mnogo manje pažnje posvećeno je drugom skupu problema koji kritički utječu na mogućnost primjene ovog izgovora u vidu olakotne okolnosti u pojedinim slučajevima, izvedenih iz ambivalentne prirode djelatničke kontrole u ovisnosti. Kako bi pronašao prikladan odgovor na moralnu štetu, osoba koja ima pravo kriviti nekoga mora donijeti sud o tome u kojoj mjeri je djelatnik posjedovao određene moralno relevantne sposobnosti u vrijeme čina. U praksi, to će često biti izuzetno teško procijeniti. Etički izazov za osobu s pravom na krivljenje se u suštini odnosi na donošenje suda o stvarima koje se čine pododređenima dostupnim dokazima.

Ključne riječi: ovisnost; moralna odgovornost; kontrola ponašanja; umanjenje; stupnjevi krivnje.

Nontrivial Existence in Transparent Intensional Logic

Miloš Kosterec
Institute of Philosophy, Slovak Academy of Sciences, Slovakia

ABSTRACT

The paper analyses the validity of arguments supporting the assumption of a constant universe of individuals over all possible worlds within Transparent Intensional Logic. These arguments, proposed by Tichý, enjoy widespread acceptance among researchers working within the system. However, upon closer examination, this paper demonstrates several weaknesses in the argumentation, suggesting that there is an open possibility to incorporate a variable universe of individuals even in models within this system.

Keywords: individual; existence; non-trivial property; existence test.

Netrivijalno postojanje u transparentnoj intenzionalnoj logici

Miloš Kosterec
Institute of Philosophy, Slovak Academy of Sciences, Slovakia

SAŽETAK

U radu se analizira valjanost argumenata koji podupiru pretpostavku o stalnom univerzumu pojedinaca kroz sve moguće svjetove unutar Transparentne intenzionalne logike. Ovi argumenti, koje je predložio Tichý, su naširoko prihvaćeni među istraživačima koji rade unutar ovog sustava. Međutim, putem detaljanijeg ispitivanja, ovaj rad pokazuje nekoliko slabosti u argumentaciji, sugerirajući da postoji otvorena mogućnost uključivanja varijabilnog univerzuma pojedinaca čak i u modelima unutar ovog sustava.

Ključne riječi: individua; postojanje; netrivijalno svojstvo; test postojanja.

Can We Defend Normative Error Theory?

Joshua Taccolini
Saint Louis University, USA

ABSTRACT

Normative error theorists aim to defend an error theory which says that normative judgments ascribe normative properties, and such properties, including reasons for belief, are never instantiated. Many philosophers have raised objections to defending a theory which entails that we cannot have reason to believe it. Spencer Case objects that error theorists simply cannot avoid self-defeat. Alternatively, Bart Streumer argues that we cannot believe normative error theory but that, surprisingly, this helps its advocates defend it against these objections. I think that if Streumer's argument is successful, it provides error theorists an escape from Case's self-defeat objection. However, I build upon and improve Case's argument to show that we could never even successfully defend normative error theory whether we can believe it or not. So, self-defeat remains. I close by offering some reasons for thinking our inability to defend normative error theory means that we should reject it, which, in turn, would mean that it's false.

Keywords: Normative Error Theory; self-defeat; theory defense.

Možemo li braniti teoriju normativne pogreške?

Joshua Taccolini
Saint Louis University, USA

SAŽETAK

Normativni teoretičari pogreške nastoje braniti teoriju pogreške koja kaže da normativni sudovi pripisuju normativna svojstva, a takva svojstva, uključujući razloge za vjerovanje, nikada nisu instancirana. Mnogi filozofi su iznijeli prigovore obrani teorije koja podrazumijeva da ne možemo imati razloga vjerovati u nju. Spencer Case prigovara da teoretičari pogreške jednostavno ne mogu izbjeći samopobijanje. S druge strane, Bart Streumer tvrdi da ne možemo vjerovati u normativnu teoriju pogreške, ali da to, pomalo iznenađujuće, pomaže njenim zagovornicima da je obrane od ovih prigovora. Smatram da, ako je Streumerov argument uspješan, on omogućuje teoretičarima pogreške izbjegavanje Caseovog prigovora o samopobijanju. Međutim, nadograđujem i poboljšavam

Caseov argument kako bih pokazao da nikada ne bismo mogli uspješno obraniti normativnu teoriju pogreške, bez obzira na to možemo li vjerovati u nju ili ne. Dakle, samopobijanje ostaje. Rad zaključujem nudeći neke razloge za mišljenje da naša nesposobnost da obranimo normativnu teoriju pogreške znači da bismo je trebali odbaciti, što bi posljedično značilo da je ona neistinita.

Ključne riječi: teorija normativne pogreške; samopobijanje; obrana teorije.

Better-Making Properties and the Objectivity of Value Disagreement

Erich H. Rast
Nova University Lisbon, Portugal

ABSTRACT

A light form of value realism is defended according to which objective properties of comparison objects make value comparisons true or false. If one object has such a better-making property and another lacks it, this is sufficient for the truth of a corresponding value comparison. However, better-making properties are only necessary and usually not sufficient parts of the justifications of value comparisons. The account is not reductionist; it remains consistent with error-theoretic positions and the view that there are normative facts.

Keywords: values; axiology; better than; the good; objectivity; value disagreement.

Poboljšavajuća svojstva i objektivnost neslaganja u pogledu vrijednosti

Erich H. Rast
Nova University Lisbon, Portugal

SAŽETAK

Brani se lagani oblik realizma vrijednosti prema kojem objektivna svojstva predmeta usporedbe čine vrijednosne usporedbe istinitima ili lažnima. Ako jedan predmet ima svojstvo koje ga čini boljim, a drugi ga nema, to je dovoljno za istinitost odgovarajuće vrijednosne usporedbe.

Međutim, svojstva koja predmet boljim samo su nužni i obično nisu dovoljni dijelovi opravdanja vrijednosnih usporedbi. Ovo objašnjenje nije redukcionističko; ostaje dosljedno sa stajalištima poput teorije pogreške i gledištem da postoje normativne činjenice.

Ključne riječi: vrijednosti; aksiologija; bolje od; dobro; objektivnost; neslaganje u pogledu vrijednosti.

Integrative Bioethics: A Blind Alley of European Bioethics

Tomislav Bracanović
Institute of Philosophy, Croatia

ABSTRACT

Integrative bioethics is a predominantly Croatian school of thought whose proponents claim to have initiated an innovative and recognizably European concept of bioethics capable of dealing with the most pressing issues of our time. In this paper, a critical overview of the integrative bioethics project is undertaken to show that it is, in fact, a poorly articulated and arguably pseudoscientific enterprise fundamentally incapable of dealing with practical challenges. The first section provides the basic outline of integrative bioethics: its historical development, major proponents, geographical context and philosophical foundations. The second section considers its main theoretical shortcomings: the absence of normativity, collapse into ethical relativism and frequent intratheoretical inconsistencies. The third section addresses the issue of typically pseudoscientific features of integrative bioethics: verbose language, constant self- glorification and isolation from mainstream science. The fourth and concluding section of the paper argues that integrative bioethics—regarding its quality, reception and identity—does not merit the “European bioethics” label and is better described as a blind alley of European bioethics.

Keywords: integrative bioethics; pluriperspectivism; inconsistency; ethical relativism; pseudoscience; European bioethics.

Integrativna bioetika: slijepa ulica europske bioetike

Tomislav Bracanović
Institute of Philosophy, Croatia

SAŽETAK

Integrativna bioetika je pretežno hrvatska škola mišljenja čiji zagovornici tvrde da su inicirali inovativan i prepoznatljivo europski koncept bioetike sposoban nositi se s najhitnijim problemima našeg vremena. U ovom radu se daje kritički pregled projekta integrativne bioetike kako bi se pokazalo da je to, zapravo, loše artikuliran i vjerojatno pseudoznanstveni pothvat koji je temeljno nesposoban nositi se s praktičnim izazovima. Prvi dio rada pruža osnovni pregled integrativne bioetike: njen povijesni razvoj, glavne zagovornike, geografski kontekst i filozofske temelje. Drugi dio razmatra njene glavne teorijske nedostatke: nedostatak normativnosti, urušavanje u etički relativizam i česte unutarteorijske nedosljednosti. Treći dio bavi se pitanjem tipično pseudoznanstvenih obilježja integrativne bioetike: opširnog jezika, stalne samoglorifikacije i izolacije od "mainstream" znanosti. Četvrti i zaključni dio rada tvrdi da integrativna bioetika—s obzirom na svoju kvalitetu, recepciju i identitet––ne zaslužuje oznaku „europska bioetika“ te ju je bolje opisati kao slijepu ulicu europske bioetike.

Ključne riječi: integrativna bioetika; pluriperspektivizam; nedosljednost; etički relativizam; pseudoznanost; europska bioetika.

Are Composite Subjects Possible? A Clarification of the Subject Combination Problem Facing Panpsychism

Siddharth S
Sai University, India

ABSTRACT

Panpsychism, the view that phenomenal consciousness is present at the fundamental physical level, faces the subject combination problem—the question of whether (and how) subjects of experience can combine. While various solutions to the problem have been proposed, these often seem to be based on a misunderstanding of the threat posed by the subject combination problem. An example is the exchange in this journal between Siddharth (2021) and Miller (2022). Siddharth argued that the phenomenal bonding solution failed to address the subject combination problem, while Miller responded that Siddharth had (among other things) misunderstood the problem that the phenomenal bonding solution was trying to solve. In this paper, I seek to clarify the real subject combination problem facing panpsychism, and on this basis, evaluate the various attempts at defending the possibility of subject composition.

Keywords: panpsychism; combination problem; subject composition; consciousness.

Jesu li složeni subjekti mogući? Pojašnjenje problema kombinacije subjekata u panpsihizmu

Siddharth S
Sai University, India

SAŽETAK

Panpsihizam, gledište da je fenomenalna svijest prisutna na fundamentalnoj fizičkoj razini, suočava se s problemom kombinacije subjekata—pitanjem mogu li se (i kako) subjekti iskustva kombinirati. Iako su predložena različita rješenja za taj problem, često se čini da se temelje na pogrešnom razumijevanju izazova koji predstavlja problem kombinacije subjekata. Primjer je razmjena u ovom časopisu između Siddhartha (2021) i Millera (2022). Siddharth je tvrdio da rješenje fenomenalnog povezivanja nije uspjelo riješiti problem kombinacije subjekata, dok je Miller odgovorio da je Siddharth (između ostalog) pogrešno razumio problem koji je rješenje fenomenalnog povezivanja pokušavalo riješiti. U ovom radu nastojim razjasniti stvarni problem kombinacije subjekata s kojim se suočava panpsihizam i na temelju toga procijeniti različite pokušaje obrane mogućnosti sastava subjekata.

Ključne riječi: panpsihizam; problem kombinacije; sastav subjekta; svijest.

Is Kindness a Virtue?

Kristján Kristjánsson
University of Birmingham, UK

ABSTRACT

This article swims against the stream of academic discourse by answer the title question in the negative. This contrarian answer is not meant to undermine the view that kindness is a good thing; neither is it, however, an example of a mere philosophical predilection for word play. I argue that understanding kindness as a virtue obscures rather than enlightens, for the reason that it glosses over various distinctions helping us make sense of moral language and achieving "virtue literacy". I survey some of the relevant psychological literature before moving on to philosophical sources. I subsequently delineate the alternative ways in which coherent virtue ethicists can say everything that they want to say about kindness by using much better entrenched and less bland terms. I offer a view of kindness as a cluster concept in the same sense as the Wittgensteinian concept of a game. Finally, I elicit some implications of this view for practical efforts at character education.

Keywords: virtue ethics; Aristotle; kindness; moral virtue; umbrella concept; cluster concept.

Je li ljubaznost vrlina?

Kristján Kristjánsson
University of Birmingham, UK

SAŽETAK

Ovaj članak ide protiv struje akademskog diskursa odgovarajući na naslovno pitanje negativno. Ovaj suprotni odgovor nije zamišljen da potkopa stav da je ljubaznost dobra stvar; niti je, međutim, primjer puke filozofske sklonosti za igru riječima. Tvrdim da shvaćanje ljubaznosti kao vrline više zamagljuje nego rasvjetljuje, iz razloga što zanemaruje različite razlike koje nam pomažu razumjeti moralni jezik i postići „kreposnu pismenost". U radu dajem pregled relevantne psihološke literature prije nego što se prebacim na filozofske izvore. Nakon toga ocrtavam alternativne načine na koje dosljedni etičari vrline mogu reći sve što žele reći o ljubaznosti, koristeći mnogo bolje utemeljene i manje nejasne pojmove. Nudim pogled na ljubaznost kao klasterski pojam u istom smislu kao što je Wittgensteinov pojam igre. Na kraju, iznosim neke implikacije ovog stajališta za praktične izazove za razvoj karaktera.

Ključne riječi: etika vrline; Aristotel; ljubaznost; moralna vrlina; krovni pojam; klasterski pojam.

Translated by Marko Jurjako (Rijeka) and Iva Martinić (Rijeka)

Proofread by Iva Martinić (Rijeka)

AUTHOR GUIDELINES

Publication ethics

EuJAP subscribes to the publication principles and ethical guidelines of the Committee on Publication Ethics (COPE).

Submitted manuscripts ought to:

- be unpublished, either completely or in their essential content, in English or other languages, and not under consideration for publication elsewhere;
- be approved by all co-Authors;
- contain citations and references to avoid plagiarism, self-plagiarism, and illegitimate duplication of texts, figures, etc. Moreover, Authors should obtain permission to use any third party images, figures and the like from the respective copyright holders. The pre-reviewing process includes screening for plagiarism and self-plagiarism by means of internet browsing and software Turnitin;
- be sent exclusively electronically to the Editors (eujap@ffri.uniri.hr) (or to the Guest editors in the case of a special issue) in a Word compatible format;
- be prepared for blind refereeing: authors' names and their institutional affiliations should not appear on the manuscript. Moreover, "identifiers" in MS Word Properties should be removed;
- be accompanied by a separate file containing the title of the manuscript, a short abstract (not exceeding 300 words), keywords, academic affiliation and full address for correspondence including e-mail address, and, if needed, a disclosure of the Authors' potential conflict of interest that might affect the conclusions, interpretation, and evaluation of the relevant work under consideration;
- be in American or British English;
- be no longer than 9000 words, including references (for Original and Review Articles).
- be between 2000 and 5000 words, including footnotes and references (for Discussions and Critical notices)

We ask authors to submit only one manuscript at a time. A second submission by the same author is allowed only after a final decision has been made on their previously submitted manuscript.

Norms for publishing with AI

The Journal does not exclude the use of AI generated text. However, all authors (including reviewers and editors) take full responsibility for its factual accuracy and the proper acknowledgement of sources. In the acknowledgement section of your manuscript or the title page (depending on the submission/publication stage) or in other kind of reports you must identify the AI that was used, and the extent of the contribution. For instance, ChatGPT (version or the date when the AI was used).

The contribution level of the AI can be defined as follows:

- negligible – means the AI only made minor changes to the manuscript's style or grammar (this includes using AI for copyediting and similar services);
- modest – means the AI made important suggestions but was not the primary driver of the research or had an essential role in writing the manuscript;
- substantial – means the AI made several crucial suggestions that shaped the research and the manuscript could not have been completed without it.

If the contribution of the AI is "negligible", there is no requirement to mention its usage during the submission or review and publication processes. However, for any other level of contribution, it is expected that authors will report the extent of AI usage. In cases where the AI contribution is "substantial", authors, reviewers, and editors should provide a comprehensive description of the AI usage and its contributions in a narrative format.

Initial submission

When first submitting a manuscript it is not required that the manuscript conforms to EuJAP's style guidelines. Only after a manuscript has been accepted for publication we expect the authors to format the manuscript in accordance with EuJAP's style guidelines.

Submitting revised manuscripts

When submitting a revised manuscript, please include also a separate document where it is explained how revisions were made in response to reviewers’ comments.

Policy for submitted manuscripts

If the submitted manuscript is authored by more than one person, there should be a brief explanation in the title page of the contribution of each Author with respect to the conception and design of the argument, study, etc. and writing of the paper.

To preserve the anonymous status of the review process, we prefer (but do not require) that submitted versions of manuscripts are not deposited in open access article repositories.

Policy for accepted and published manuscripts

Accepted and published versions of the manuscript can be deposited in institutional or personal repositories without an embargo period. In case of published manuscripts, a link (with DOI) to the journal’s web pages and/or HRCAK should be added.

Malpractice statement

If the manuscript does not match the scope and aims of EuJAP, the Editors reserve the right to reject the manuscript without sending it out to external reviewers. Moreover, the Editors reserve the right to reject submissions that do not satisfy any of the previous conditions.

If, due to the authors' failure to inform the Editors, already published material will appear in EuJAP, the Editors will report the authors' unethical behaviour in the next issue and remove the publication from EuJAP web site and the repository HRČAK.

In any case, the Editors and the publisher will not be held legally responsible should there be any claims for compensation following from copyright infringements by the authors.

For additional comments, please visit our web site and read our Publication ethics statement (https://eujap.uniri.hr/publication-ethics/). To get a sense of the review process and how the referee report ought to look like, the prospective Authors are directed to visit the *For Reviewers* page on our web site (https://eujap.uniri.hr/instructions-for-reviewers/).

Style

Accepted manuscripts should:

- follow the guidelines of the most recent Chicago Manual of Style
- contain footnotes and no endnotes
- contain references in accordance with the author-date Chicago style, here illustrated for the main common types of publications (T = in text citation, R = reference list entry)

Book
T: (Nozick 1981, 203)
R: Nozick, R. 1981. *Philosophical Explanations.* Cambridge: Harvard University Press.

Book with multiple authors

T: (Hirstein, Sifferd, and Fagan 2018, 100)

R: Hirstein, William, Katrina Sifferd, and Tyler Fagan. 2018. *Responsible Brains: Neuroscience, Law, and Human Culpability*. Cambridge, Massachusetts: The MIT Press.

Chapter or other part of a book
T: (Fumerton 2006, 77-9)
R: Fumerton, Richard. 2006. 'The Epistemic Role of Testimony: Internalist and Externalist Perspectives'. In *The Epistemology of Testimony*, edited by Jennifer Lackey and Ernest Sosa, 77–91. Oxford: Oxford University Press.
https://doi.org/10.1093/acprof:oso/9780199276011.003.0004.

Edited collections
T: (Lackey and Sosa 2006)
R: Lackey, Jennifer, and Ernest Sosa, eds. 2006. *The Epistemology of Testimony*. Oxford: Oxford University Press.

Article in a print journal
T: (Broome 1999, 414-9)
R: Broome, J. 1999. "Normative requirements." *Ratio* 12: 398-419.

Electronic books or journals
T: (Skorupski 2010)

R: Skorupski, John. 2010. "Sentimentalism: Its Scope and Limits." Ethical Theory and Moral Practice 13 (2): 125–36. https://doi.org/10.1007/s10677-009-9210-6.

Article with multiple authors in a journal
T: (Churchland and Sejnowski 1990)
R: Churchland, Patricia S., and Terrence J. Sejnowski. 1990. "Neural Representation and Neural Computation." *Philosophical Perspectives 4*. https://doi.org/10.2307/2214198

T: (Dardashti, Thébault, and Eric Winsberg 2017)
R: "Dardashti, Radin, Karim P. Y. Thébault, and Eric Winsberg. 2017. Confirmation via Analogue Simulation: What Dumb Holes Could Tell Us about Gravity." *The British Journal for the Philosophy of Science* 68 (1): 55–89.
https://doi.org/10.1093/bjps/axv010

Website content
T: (Brandon 2008)
R: Brandon, R. 2008. Natural Selection. *The Stanford Encyclopedia of Philosophy*. Edited by Edward N. Zalta. Accessed September 26, 2013.
http://plato.stanford.edu/archives/fall2010/entries/natural-selection

Forthcoming
For all types of publications followed should be the above guideline style with exception of placing 'forthcoming' instead of date of publication. For example, in case of a book:
T: (Recanati forthcoming)
R: Recanati, F. forthcoming. *Mental Files*. Oxford: Oxford University Press.

Unpublished material
T: (Gödel 1951)
R: Gödel, K. 1951. *Some basic theorems on the foundations of mathematics and their philosophical implications*. Unpublished manuscript, last modified August 3, 1951.

Final proofreading

Authors are responsible for correcting proofs.

Copyrights

The journal allows the author(s) to hold the copyright without restrictions. In the reprints, the original publication of the text in EuJAP must be acknowledged by mentioning the name of the journal, the year of the publication, the volume and the issue numbers and the article pages.

EuJAP subscribes to Attribution-ShareAlike 4.0 International (CC BY-SA 4.0). Users can freely copy and redistribute the material in any medium or format, remix, transform, and build upon the material for any purpose. Users must give appropriate credit, provide a link to the license, and indicate if changes were made. Users may do so in any reasonable manner, but not in any way that suggests the licensor endorses them or their use. Nonetheless, users must distribute their contributions under the same license as the original.

Archiving rights

The papers published in EuJAP can be deposited and self-archived in the institutional and thematic repositories providing the link to the journal's web pages and HRČAK.

Subscriptions

A subscription comprises two issues. All prices include postage.

Annual subscription:
International:
individuals € 50
institutions € 100
Croatia:
individuals € 30
institutions € 60

Bank: Zagrebačka banka d.d. Zagreb
SWIFT: ZABAHR 2X
IBAN: HR9123600001101536455
Only for subscribers from Croatia,
please add: “poziv na broj”: 0015-03368491

European Journal of Analytic Philosophy is published twice per year.

The articles published in the European Journal of Analytic Philosophy are indexed and abstracted in SCOPUS, SCImago, Web of Science (Emerging Sources), The Philosopher’s Index, European Reference Index for the Humanities (ERIH PLUS), Dimensions, Directory of Open Access Journals (DOAJ), PhilPapers, and Portal of Scientific Journals of Croatia (HRČAK), ANVUR (Italy), Sherpa Romeo