dialectica

International Journal of Philosophy

Contents

EUROPEAN
SOCIETY
FOR
ANALYTIC
PHILOSOPHY

Vol 74 : 1
March 2020

dialectica

International Journal of Philosophy
Official Organ of the European Society of Analytic Philosophy

founded in 1947 by Gaston Bachelard, Paul Bernays and Ferdinand Gonseth dialectica

vol 74
issue 1

March 2020

Contents

Editorial
Dialectica Goes Open Access

Philipp Blum

We are happy to announce that *Dialectica* is now an Open Access journal. Starting with this issue, the journal has adopted the so-called “Platinum” or “Diamond” Open Access model under which we do not charge access fees to readers nor article processing charges to authors. Thanks to the generous support of the Swiss National Science Foundation (CRSK-1-190939) and swissuniversities (OAHUBSP), all new issues are available at dialectica.philosophie.ch. In the near future, we hope to make freely available a full bibliography and a detailed submission statistics.

We hope that this audacious step will increase our readership, attract more and better submissions and reward our many industrious referees, without whom our journal would not be possible. By making all tools and techniques devised for our OA-transition freely available, by documenting the transition process itself, putting the reasoning behind our decisions out for your scrutiny and disclosing the difficulties encountered in establishing a sustainable financial model, we also hope to convince other well-established philosophy journals to free themselves of the increasingly tight grip of profit-oriented publishers and to turn the ideals of Open Science into action.

The success of our discipline’s transition to Open Access will depend on four factors:

- that established journals, in particularly the most important ones, start the transition to Open Access immediately, with the aim to sever their links to for-profit publishers;
- that the many new journals coming into existence, reflecting in theme and character the diversity of our growing discipline, are all fully (i.e. Platinum) Open Access;
- that philosophers strongly support the transition by making a collective decision to only submit to, referee for, and edit Open Access journals;

- that our funding bodies make their support contingent on publication in OA venues, and that new funding models are devised which allow universities and libraries to directly fund the running costs of OA journals – an "OA coalition" of philosophy journals should be created to press for this cause.

While we hope to have contributed to the first of these objectives, our transition has two further goals. First, to strengthen our institutional basis in Switzerland, notably by recruiting more Swiss philosophers into the Editorial Committee. Second, to make the refereeing process faster and more positive-oriented. In our new "fishpond" model, members of the Editorial Committee pick anonymized papers that they hope to promote and send them out to referees. If, based on the reports, they make a successful case to the committee, the papers are accepted. Submissions are sent back to the authors after one month if they have not been picked up. This is not to be understood as a (desk-) rejection, but simply as an acknowledgment of the limitedness of our resources. We hope that this triple-blind, positive-oriented process will shorten the turn-around time for authors and make the editorial and referee work more attractive.

The ongoing transition process has already benefitted from much help, including much needed technical advice by Denis Maier and Albert Krewinkel. Let me thank, first and foremost, my co-editor Fabrice Correia, the members of the Editorial Board and the Editorial Committee, Julien Dutant, the head of the *Dialectica* OA initiative, the members of the SNSF *Spark* project team (Jonathan Biedermann, Sharon Casu, Thomas Hodgson, Nemo Krüger, Ryan Miller, Sandro Räss, Marco Toscano, Christian Weibel), our library consultants Rebecca Iseli Büchi and Gian-Andri Töndury and the new managing editor of *Dialectica*, Marco Schori – philosophers who sacrificed some of their research time to make a practical impact in our common quest to make the world a little better.

Philipp Blum
Editor and Project Leader
philipp.blum@philosophie.ch

The Personalized A-Theory of Time and Perspective

Vincent Conitzer

A-theorists and B-theorists debate whether the “Now” is metaphysically distinguished from other time slices. Analogously, one may ask whether the “I” is metaphysically distinguished from other perspectives. Few philosophers would answer the second question in the affirmative. An exception is Caspar Hare, who has devoted two papers and a book to arguing for such a positive answer. In this paper, I argue that those who answer the first question in the affirmative—A-theorists— should also answer the second question in the affirmative. This is because key arguments in favor of the A-theory are more effective as arguments in favor of the resulting combined position, and key arguments against the A-theory are ineffective against the combined position.

In a series of unconventional but lucid works, Caspar Hare has laid out and defended a theory of *egocentric presentism* (or, in his more recent work, *perspectival realism*), in which a distinguished individual’s experiences are *present* in a way that the experiences of others are not (2007, 2009, 2010). Closely related ideas appear in the writings of others. One example is Valberg’s notion of the “personal horizon,” especially considering his discussion of “the truth in solipsism” and his insistence that “my” horizon is really “the” (preeminent) horizon (2007). Merlo’s “subjectivist view of the mental” is arguably even more closely related; he argues that “one’s own mental states are metaphysically privileged vis-à-vis the mental states of others” and discusses in detail the relationship of his view to Hare’s (2016). As another example, in a review of “The Character of Consciousness” (Chalmers 2010), Hellie (2013) argues that this work fails to do justice to the *embedded point of view* aspect of consciousness. He illustrates this with what he calls a “vertiginous question”: why, of all subjects, is *this* subject (the one corresponding to the human being Benj Hellie) the one whose experiences are “live”? In other work (2019), I explore whether the “liveness” of one particular perspective is a *further fact*—a fact that does not follow logically from the physical facts of the world—by considering the analogy to looking in on a simulated world through a virtual reality headset: besides the computer code that determines the physics of the simulated world, there must be additional code that determines which simulated agent's perspective to show on the headset.

In any case, Hare's exposition of these ideas is clearest for the present purpose, so I will focus on it. In an effort, possibly with limited success, to avoid misrepresenting his position, as well as to clarify the relation to other work, let me introduce my own terminology. Let us refer to the theory that states that there is a metaphysically (rather than merely epistemically) distinguished[1] *I* (or *Self*[2]) as the α-theory. The intent is to emphasize the analogy with how the A-theory (McTaggart 1908) states that there is a metaphysically distinguished *Now*.[3] Similarly, I will refer to the theory that contradicts the existence of any metaphysically distinguished *I* as the β-theory. Hare is thus defending the

[1] Throughout the paper, I will be deliberately noncommittal about the exact nature of such a metaphysical distinction. The reason is that the arguments presented here do not depend on what this distinction consists in. In the analogous case of a metaphysically distinguished *time* (rather than a metaphysically distinguished subject), by not committing to any particular interpretation, I can simultaneously address all varieties of A-theorists—presentists, moving-spotlight theorists, growing-block theorists, etc.—even though they disagree about the exact nature of the Now's metaphysical distinction. Of course, there is disagreement even about how to define the individual varieties. Deasy (2017) discusses this at length, and proposes to define each of the main varieties as the conjunction of the A-theory (which he takes to mean "There is an absolute, objective present instant") and a proposition about whether things begin and/or cease to exist. For example, for the growing-block theorist, that proposition is "Sometimes, something begins to exist and nothing ever ceases to exist." While the distinctions between the various definitions are significant, again, my aim is to steer clear of this debate here and stick to arguments that work for any of these definitions. The same is true for the case of a metaphysically distinguished subject.

[2] Again, what exactly the distinguished entity is—a human being, a brain, an experience—is not essential to my arguments, so I will remain deliberately noncommittal.

[3] Is a commitment to a distinguished *Now* what defines the A-theory, or is it a commitment to tensed facts? (And in the latter case, should the α-theory's defining commitment instead be to first-personal facts?) To the extent that these commitments are not equivalent, in this paper, I will stick with the commitment to a distinguished *Now* (or *I*), as others have done—e.g. Cameron (2015, 89). For what it is worth, while a detailed analysis is outside the scope of this paper, I believe that they are in fact equivalent. I believe that a distinguished *Now* implies tensed facts, such as the fact that today is July 3, 2019. The other direction is perhaps more controversial, but I believe it holds as well: tensed facts such as the fact that today is July 3, 2019 distinguish a specific time, to which we may refer as the *Now*. A theory such as fragmentalism (Fine 2005) might be used to dispute the second direction: if we consider *all* tensed facts, including those for other times, then no specific time is distinguished. But, of course, the set of all tensed facts taken together is full of contradictions, as it also contains, for example, the fact that today is not July 3, 2019. Avoiding such contradictions means restricting attention to a consistent fragment—but

α-theory. It is not entirely clear to me whether the specific version he defends is intended to be analogous to presentism [or actualism—I will refrain from discussing modality in this paper, but the parallels between time/subjectivity and modality are well recognized; see Prior and Fine (1977); Bergmann (1999)], or rather to something like a spotlight theory (or possibilism). In fact, his writing suggests different answers to this question in different places, and I will not attempt to resolve this small mystery here.

Others have commented on the idea of a metaphysically distinguished *I*—or, similarly but not equivalently,[4] a metaphysically distinguished *Here*—in the context of the philosophy of time. (While the differences between a metaphysically distinguished *I* and a metaphysically distinguished *Here* will not matter for some of the arguments presented in this paper, it is useful to note that, in the context where a distinguished *I* is combined with a distinguished *Now*, the combination of these two immediately implies a distinguished *Here* as well—namely, the location of the distinguished individual at the distinguished time.[5]) However, they have usually dismissed it rather quickly, in order to move on with the case of a metaphysically distinguished *Now* (whether or not they support the latter). For example, Zimmerman (2005, 422) writes:

> An egocentric analogue of actualism ('personalism', to steal and abuse a term) is very hard to imagine. Perhaps there is some kind of not-merely-epistemological solipsism that would qualify. In any case, only the maniacally egocentric could be this sort of personalist.

Further back, Williams (1951, 458) writes:

> Perhaps there exists an intellectualistic solipsist who grants the propriety of conceiving a temporal stretch of events, to wit, his own whole inner biography, while denying that the spatial scheme is a literal truth about anything. Most of the disparagers of the manifold, however, are of opposite bias. Often ready enough to

this in turn distinguishes a specific time. For further discussion of problems that fragmentalism faces, see Cameron (2015, 86–102).

[4] For a discussion of the differences and their implications, in the related context of the Lewisian and Quinean accounts of centered worlds, see Liao (2012).

[5] The combination similarly implies a distinguished observational frame of reference corresponding to the distinguished individual's state of motion. All of this does, of course, require the distinguished individual to be spatially located and to move through time and space, rather than, say, an immaterial soul or something existing for only an instant.

doi: 10.48106/dial.v74.i1.02

> take literally the spatial extension of the world, they dispute the codicil which rounds it out in the dimension of time.

Fine (2005, 285) treats the case of first-personal realism in detail, but advocates for adopting a nonstandard variety of realism, either taking reality to be relative to a standpoint, or (his preferred option) considering it to be fragmented.[6] He notes:

> It has seemed evident that, of all the possible worlds, the actual world is privileged; it is the standpoint of reality, as it were, and the facts that constitute reality are those that obtain in this world. On the other hand, if we ask, in the first-personal case whether we should be a nonstandard realist (given that we are going to be first-personal realists in the first place), then the answer to most philosophers has seemed to be a clear 'yes'. It has seemed metaphysically preposterous that, of all the people there are, I am somehow privileged - that my standpoint is *the* standpoint of reality and that no one else can properly be regarded as a source of first-personal facts. The case of time is perplexing in a way that these other cases are not.

I believe that there is value in exploring the α-theory more thoroughly, rather than dismissing it summarily for being repugnant in one way or another. The words "egocentrism" and "solipsism" are both loaded with too much baggage. While "egocentric," taken literally, aptly describes the α-theory, the common interpretation of the word carries various negative connotations, and it is not clear to me that these are fair to apply to every possible α-theorist. Just as A-theorists can take great interest in times other than their own (otherwise why would they bother to write papers?), the α-theorist can presumably take great interest in people other than herself.[7] The relation to solipsism is also

[6] Lipman (2015) discusses fragmentalism in more detail.

[7] It should be noted here that, on the face of it, Hare (2007, 2009) does introduce his theory to justify placing greater weight on oneself than on others in making decisions. However, he also points out that the (distinguished) presence of an experience is only one factor in making decisions ("It is better that there be present suffering from a hangnail than absent suffering of leg-crushing."). Perhaps more importantly, key examples that Hare uses in these works to support his theory are preferential in nature, such as an example where one knows that CJH (Hare) and Joe Bloggs have been in a train crash, CJH is about to have a painful operation, the subject knows he is one of these two but cannot remember which one, and so the subject hopes to not be CJH (2007). Such preferential examples are quite helpful to illustrate and motivate these types of theories—similar

not obvious. Hare intends for his theory to be only a weak and subtle version of solipsism that does not deny the existence of others' consciousness (Hare 2009, 41–46), and others have granted him as much (e.g. Smith 2011; and Mark Johnston in the introduction to Hare 2009).[8]

Indeed, a key point is that, just as there are multiple versions of the A-theory, there are also multiple versions of the α-theory, and these vary in the status they accord to other individuals. Perhaps more importantly—and this is the main focus of this paper—something is lost when attempting to study the A vs. B question separately from the α vs. β question; the two are very much interrelated. To illustrate this, consider a theory that allows a distinguished *I* that is not alive at the time of the distinguished *Now*, thereby treating the two types of distinction as independent. Many of the arguments that I give in what follows would do little to support such a theory. Hence, in what follows I will not take the αA-theory—the label that I will use for a view that combines the α-theory with the A-theory—to allow this possibility; what I have in mind is that a single *(living-)person-stage* is distinguished. This interrelation is relevant to the previous point. For example, the αA-theorist may accord to other persons the same metaphysical status as she does to herself in past and future time slices.

After presenting, for the sake of illustration, some versions of the αA-theory in section 1, I will argue in section 2 that key arguments that have been given to support the A-theory support the α-theory just as well, and in fact support the combined αA-theory especially strongly, placing the onus on the βA-theorist to explain why she accepts the A-theory but not the α-theory. (It would seem that most A-theorists, at least publicly, are βA-theorists in my terminology.) Specifically, in subsection 2.1 I will discuss the argument from presence *simpliciter*, and in subsection 2.2 the argument from the appropriateness of sentiments such as those expressed by "Thank goodness that's over!" I will also argue, in section 3, that some serious challenges that the βA-theorist faces are much less problematic for the αA-theorist. Specifically, in subsection 3.1 I will discuss the argument from special relativity, in subsection 3.2 the argument that the direction of time may be a local matter, in subsection 3.3

ones can be given to motivate the A-theory, as Hare does and others have done before him—even if one does not wish to normatively endorse the preferences used in the example. I will also discuss such examples later in this paper.

[8] Others have tried to distinguish between more and less defensible versions of solipsism along similar lines; a particularly notable example is Valberg (2007). Similar ideas also appear in Johnston (2010).

doi: 10.48106/dial.v74.i1.02

the argument that asks for the rate at which time passes, and in subsection 3.4 the argument from time travel and Gödelian universes.

Overall, my main objective is to argue that the αA-theory is superior to the βA-theory.[9] I would similarly argue that the αA-theory is superior to the αB-theory, but I do not expect many to defend the latter view.[10] This would leave the αA-theory and the βB-theory as the remaining candidates. The reader might expect that my next step will be simply to argue that the α-theory is so unappealing that we should accept the βB-theory, and hence, *a fortiori*, the B-theory. However, I believe that that conclusion is too hasty; an effective discussion of the relative merits of the αA-theory and the βB-theory requires arguments of a different type than what I will present here. So, I will be content to let both theories stand for now.

1 Some Versions of the α-Theory

The A-theory counts among its supporters presentists, moving-spotlight theorists, and growing-block theorists. Can we conceive of similar distinctions among α-theorists? Rather than studying this in isolation from the A vs. B question, it seems more enlightening to ask what natural versions of the αA-theory there are. (Common versions of the A-theory and the B-theory can straightforwardly be reinterpreted as versions of the βA-theory and the βB-theory.) I will present some versions in this section. My aim here is not to defend specific versions or to reach any definitive conclusion about which version is best. I also make no claim that this list is exhaustive, though I believe that it includes the versions that are most natural to discuss in the context of the existing literature on the A-theory. The aim of this exercise is merely to clarify some aspects of the αA-theory and prevent overly narrow interpreta-

[9] Of course, to accept this conclusion, it is not necessary to agree with every single argument presented here.

[10] In fact, Hare (2009, 48) writes that “If you think that theories that dignify a slice of history do not survive sustained critical inspection, then you can still be a four-dimensionalist egocentric presentist. Indeed, I find that an attractive position.” This may appear to put him in the αB-camp. However, on the whole in this section on the relationship to positions in the philosophy of time (Hare 2009, 46–50), he is clear that egocentric presentism does not commit one to a particular view on time, while also stating that the moving-spotlight theory is the most analogous one. Elsewhere (Hare 2010), he writes, “If you find yourself sympathetic to [the central tense realist idea] then I recommend that you consider *going the whole hog*, and becoming a perspectival realist” (emphasis mine), which might be interpreted to imply that perspectival realism is a stronger position than tense realism. In any case, as I hope will become clear from this paper, the αA-theory does not at all require a dignified *slice* of history.

tions of it. Moreover, it will be helpful to refer to some of these versions in what follows. I will also contrast these versions with some scenarios from the literature.

> Personalized Presentism. This is the most natural way to adapt presentism into an αA-theory. In this version, there is a single distinguished individual whose experience at a single distinguished point in time is, in some sense, "present." (I hope that the intended meaning of "presence" is at least somewhat clear at this point; I will discuss it in more detail in subsection 2.1.) Beyond this present experience, nothing exists. Or, perhaps, some part of the outer world can be granted some type of existence; but other experiences do not exist. However, presumably, the present experience can change (more on this below), just as presentists typically consider it possible for the Now to change.
>
> Personalized Moving Spotlight. As in the classical moving-spotlight theory, a spotlight moves over the four-dimensional block universe, except now this spotlight shines on a single individual (or that individual's experience) at a single point in time. For the personalized moving spotlight, it is less obvious how it moves (more on this below).
>
> Personalized Growing Block. In the classical growing block theory, time slices are added to the block that contain all the events in the universe at that point in time. In the personalized growing block theory, only those parts of spacetime are added that are experienced by a distinguished individual (and, perhaps, their past light cones).

Every one of these versions of the αA-theory leaves several possibilities for *how* the point of present experiences—the "I-Now"—could change or move (if it changes or moves at all).[11] These include the following variants:

[11] The word "I-Now" sounds more mystical than I would like, but we will need such a word. The word "spotlight," when interpreted as shining on a single individual's experience at a single point in time, would give the right idea, except it seems to commit the discussion to a view that all of the four-dimensional spacetime block exists, but not all of it is illuminated. While I do not want to dismiss such a view, in what follows we will not require this as an assumption. In contrast, the awkward word "I-Now" does not seem to rule out any of the possibilities. (Similarly, Hellie 2013 uses "me-now.")

doi: 10.48106/dial.v74.i1.02

> Single Individual Overall. The I-Now moves along with a single individual throughout his or her lifetime. It is never associated with any other individual.
>
> Changing Individual (αA-reincarnation). At the end of the distinguished individual's lifetime, the I-Now jumps to another individual. We can consider various subvariants. For example: (1) the I-Now cannot jump backwards in time; (2, a relativistic subvariant) the I-Now can jump anywhere that is outside of all the past light cones of points in spacetime that the I-Now occupied earlier; (3) the I-Now can jump anywhere it has not previously been; (4) the I-Now is not constrained in where it can jump.[12]
>
> Rapidly Changing Individual. The I-Now can jump from one individual to another even before the former's demise, and then jump back to the previous individual as well. We can consider the same subvariants as for αA-reincarnation.

It is admittedly odd to propose all these different versions of the α-theory without making any serious attempt to justify them individually or to claim to be exhaustive.[13] Again, my goal in doing so is merely to illustrate some of the possibilities that the theory leaves open. The availability of multiple distinct interpretations should not be surprising given the analogy and interrelation with the A-theory. It is also clear that some of these versions are much more solipsistic than others, or, at least, fit the negative connotations of solipsism more than others.

Moreover, in earlier work on theories resembling the α-theory, scenarios are often sketched that fit much better with some of these versions than with others. Usually, this is done without much discussion of why the author prefers

[12] The last two subvariants seem more difficult to reconcile with the personalized growing block theory, and might also have negative implications for free will.

[13] For example, perhaps it is not even necessary for the I-Now to change only in a sequential manner as in these variants; perhaps it can change along multiple dimensions, corresponding to changes across time and changes across space or individuals. Skow's (2009) relativistic moving-spotlight theory, in which individual points in spacetime are "lit up" from the perspective of points in *superspacetime*, seems very much in line with such a view. This also raises important questions about how these dimensions interact: Is temporal change objective or subjective? Is subjectivity eternal or temporary? For related questions on the interaction of time and modality, see Dorr and Goodman (2020).

such a version or even of what the alternatives might be. This has the effect of opening up the theory to criticisms that another version of the α-theory might have avoided. Consider the following passage by Hare (2009, 51) (discussing a thought he had as a child), corresponding to a single-individual-overall theory:

> Isn't it amazing and weird that for millions of years, generation after generation of sentient creatures came into being and died, came into being and died, and all the while there was this absence, and then one creature, CJH, unexceptional in all physical and psychological respects, came into being, and POW! Suddenly there were present things!

Later on, Hare (2009, 83) considers a type of reincarnation:

> Is it necessary that only one person ever have present experiences? Again, the natural thing is to say no. Egocentric presentism gives me conceptual resources to imagine being one sentient creature, and then, later, being another sentient creature. So (recall Nagel's "fantasy of reincarnation without memory") I can imagine that, after a lifetime of oblivious egg consumption, I die a happy philosopher, then find myself in a cage eighteen inches tall by twelve inches wide, my beak clipped to its base. This need not involve imagining that CJH dies a happy philosopher and then becomes a battery chicken. It may only involve imagining that after CJH's death there are again present experiences, and they are the experiences of a battery chicken. Once again this is a real, real nasty, metaphysical possibility.
>
> So "the one with present experiences" is a definite description that may be satisfied by different things at different times. Like all such descriptions, it behaves as a *temporally nonrigid referring term*.

Similarly, Valberg (2013, 366) writes:

> We can, however, give sense to the possibility that a human being other than JV in the past was "me," or that a human being other JV [sic] might be "me" in the future. That is, it makes sense experientially (as a way things might be or develop from within my experience) that, in the past, a human being other than JV

> occupied the position at the center of my horizon, or that a human being other than JV will occupy this position in the future.

Again, the main point here is to make clear how many possibilities the α-theory leaves open and thereby to prevent overly specific interpretations. The discussions in the remainder of the paper generally apply to all of the above versions of the αA-theory. A reader who wants to keep just a single version in mind might focus on, for example, personalized presentism or a personalized moving spotlight theory, with a single individual overall.

2 Revisiting Arguments in Favor of the A-theory

In this section, I will revisit some well-known arguments in favor of the A-theory. Subsection 2.1 concerns the argument from presence *simpliciter* and subsection 2.2 concerns the argument from the appropriateness of sentiments such as those expressed by "Thank goodness that's over!" In both cases, the argument will be shown to support the αA-theory more strongly than the βA-theory, because the argument supports a distinguished *I* just as it supports a distinguished *Now*. Whether these arguments are indeed effective against the B-theory is not the topic of this paper, so I will not review responses that B-theorists may give to these arguments here.

2.1 *Presence* Simpliciter

Arguably the most basic argument in favor of the A-theory is that of "the presence of experience." Many have made such an argument; a good exposition of one is given by Balashov (2005). The argument is that my current experience of writing this paper is *present* (or *occurs*[14]) in a way that my going through security at the airport yesterday is not present. This is not to be taken as a relative statement; everyone will agree that the writing experience at 5:50pm on March 18, 2019 is present *at 5:50pm on March 18, 2019* in a way that the airport security experience at 8:15am on March 17, 2019 is not present *at 5:50pm on March 18, 2019*. Rather, the writing experience seems present in an

[14] Balashov (2005) uses "presence" and "occurrence" to refer to different concepts, but it seems to me that others have used "presence" to refer to a concept that is closer to Balashov's "occurrence". In any case, this latter concept is what I am after, and I hope that the use of "*simpliciter*" makes this clear.

absolute sense that does not require the boldface phrases, and this is referred to as presence *simpliciter*.

I argue that, if we are to entertain such a notion, for it to be at all palatable, it must be personalized, for the following reason. Just as my earlier airport security experience is not present *simpliciter*, neither is David's experience of eating breakfast in Australia present *simpliciter*, even if this event happens to take place at the same time.[15] Let me first attempt to explain what I mean by this, and then argue for it. In order to clarify what I mean, it is tempting to write that David's breakfast experience is not present *simpliciter to me*. But to do so would undermine the argument, in the exact same way that it would undermine the purely temporal version of the argument to say that my airport security experience is not present *simpliciter right now*. In the latter sentence, "*simpliciter*" is clearly at odds with the indexical "right now." The exact same is true about the juxtaposition of "*simpliciter*" and "to me." If an experience takes place *simpliciter*, then to capture this we should not add any relativizing indexical phrases.

Moreover, it seems that only an experience can be present *simpliciter* in this way.[16] For example, it is not at all clear to me what it would mean for a chair to *itself* be present *simpliciter*. My *experience* of a chair—visual, tactile, and the result of significant cognitive processing—can be present *simpliciter*. Such an experience is the kind of thing that can have the "liveness" that past and future experiences do not, and that others' experiences do not. But I cannot imagine what it would mean for the chair to *itself* be "live" in this way. If we are willing to be a bit loose with our language, in most cases it will not cause confusion to, as a shorthand, say that the chair is present *simpliciter* when we really mean to refer to my experience of the chair. But if we are being strict, the experience is not the chair itself. Moreover, it seems that an experience can only be had by a single person[17] at a single time,[18] and it does not seem

[15] There is, of course, the question of what "at the same time" even really means given that in special relativity, simultaneity depends on the frame of reference. I will discuss relativity later; for the purpose of the current argument, we may assume a Newtonian universe.

[16] Merlo (2016, 326–27) makes a similar point.

[17] I use "person" here, and throughout, in a broad sense; presumably animals and perhaps artificial intelligence can similarly have experiences. Also, in common parlance, of course two people can "share an experience," but I use "experience" here more narrowly in its phenomenological sense.

[18] Along the same lines, Hare (2009, 49) describes the distinguished nature of his current experience and emphasizes that it is an easy-to-make "big mistake" to extend this to other current experiences. Hare (2010) presents an argument with strong similarities to the one presented here. Finally, at the end of his paper, Skow (2009) also discusses the vivid nature of present experiences and

that two distinct experiences, corresponding to different individuals and/or times, can be co-present *simpliciter* in this way. So, if anything, the argument would suggest the existence of a metaphysically distinguished (I, Now) pair.

Is this argument equivocating between "presence" in the temporal sense and "presence" in the experiential sense? Indeed both meanings of the word seem to play a role, and I believe that this is revealing rather than misleading. Insofar as the current moment in time has a "liveness" that other moments do not, it has it only through my own experience; the same moment elsewhere, even if experienced by someone else, lacks this liveness just as a past moment here, even if experienced by me, lacks it. In this way, the two meanings of the word are inextricably linked. Hare (2009, 100) similarly argues that it is in fact advantageous that the word "present" has multiple readings.

It is also important here not to be misled by how we use language. The sentence "David is eating breakfast" is, in a sense, simpler than "I went through airport security yesterday morning." Both sentences explictly refer to their subject ("David" and "I"), but only the latter needs to explicitly refer to when the event took place ("yesterday morning") in order to place it in time. So the first sentence has a type of simplicity that the second one lacks; we could add "now" to the former, but it is not needed. On the other hand, dropping "I" from the second sentence leaves it grammatically mangled. From this asymmetry between "I" and "now" one might be tempted to conclude that the word "simpliciter" more naturally corresponds to what is happening *now*—since the word "now" is usually not needed for sentences concerning the present—than it would correspond to what is happening to *me*—since a word such as "I" or "me" is usually needed for a sentence concerning the first person.

However, I would argue that the significance of this asymmetry is not metaphysical, but rather entirely linguistic. So many of our spoken sentences concern the present that, pragmatically, it would be inefficient to require adding a word like "now" to all these sentences. On the other hand, usually a conversation concerns multiple actors, so it is important to make it clear who is the subject in each sentence. To see that this is the driving force behind the asymmetry, consider a different context: my planner. In my planner, I write entries such as "attend faculty meeting at noon." It would be an inefficient use of my time to add "I" (or "I will") to the beginning of the sentence, because I would have to do so for almost every entry in my planner! In contrast, naturally,

argues that a local spotlight shining on a single individual explains this just as well as a global one (though he does not argue that it actually explains it *better*).

each of my planner entries *must* have a time associated with it; after all, if the event were happening right now, I would not have to add an entry to my planner. So, in the context of my planner, the roles that subject and time play in the pragmatic issue at hand are reversed: the former is generally implicit but the latter is not.[19] This appears to confirm that the asymmetry is due to pragmatic reasons.

2.2 *The Appropriateness of Wanting Things to (not) be Past*

Another well-known argument (Prior 1959; Zimmerman 2007) in favor of the A-theory (and presentism in particular) concerns the appropriateness of statements such as "Thank goodness that's over!" Here, "that" might refer to something like a headache the speaker was experiencing. It is often argued that the B-theory does not provide the resources to capture the full significance of this statement. Prior argues that the meaning of such a statement is not that it is good that the headache takes place at a point in spacetime earlier than the point at which the statement is uttered; in his words, "Why should anyone thank goodness for that?" Instead, what the statement is getting at is that the headache is simply *over*, and the A-theory provides the resources to capture this. But one might similarly argue in favor of the α-theory, for example appealing to the appropriateness of statements such as "Thank goodness that is not happening *to me*!" This is closely related to the question of whether self-bias could be metaphysically justified, as studied by Hare (2007, 2009). The βA-theorist is likely to complain that the analogy is not apt, because the second statement merely reflects a selfish disposition rather than something more fundamental. It is not clear to me why the same could not be said of the first statement, that the statement merely reflects the speaker's callousness towards her past self. To avoid this criticism, perhaps one can make the first statement about someone else ("Thank goodness John's headache is over!"), but, and I believe this is telling, the argument seems to lose force with this move.

Let us explore this in a bit more depth. Suppose all headaches last exactly one or two days with no ill effects afterwards, and consider the following two statements:

[19] For additional discussion of the linguistic asymmetry between time and space, and how this asymmetry is driven by pragmatic concerns in communication, see Butterfield (1984).

doi: 10.48106/dial.v74.i1.02

S_1: Thank goodness John's headache, which started yesterday, ended yesterday as well, rather than continuing into today.

S_2: Thank goodness John's headache, which started the day before yesterday, ended the day before yesterday as well, rather than continuing into yesterday.

Here, we imagine caring a great deal about John and preferring him not to suffer. Under the βA-theory, one would expect S_1: to have a significance not shared by S_2:, as the former concerns a difference in what is happening *now*, whereas the latter concerns a difference that is in any case entirely in the past. It is not clear to me that such a difference in significance is really there. Is it not just as reasonable to appreciate that John did not suffer yesterday, as it is to appreciate that he is not suffering today?

Yet, one may have an intuition that indeed, S_1: has a significance that S_2: does not. I believe that the likely grounds for this intuition are not germane to the issues under discussion here, and we can modify the scenario to remove these grounds. First, in the first situation, if John were still having a headache, I might feel compelled to try to *do* something to alleviate his suffering. However, this is easily addressed by postulating that it is common knowledge that I can do nothing of the sort. Second, if John is in my immediate environment and I see him suffering, this may cause me to suffer as well, for example due to the mirror neurons in my brain. But this is merely returning us to an example where I myself suffer, which is precisely what we were trying to avoid by introducing John. Hence, we should postulate that John is somewhere else entirely.

To make all this concrete, suppose that John has decided to go on a two-month retreat in a faraway country. He will not communicate until he gets back. Halfway into his retreat, I realize that around this time of year, he always gets a headache, which may last one or two days. I care for him and so I hope that it is just a one-day headache this time. But I will not find out until he comes back and tells me. Imagining this scenario, I do not find myself concerned specifically about whether his headache happens to be taking place right now, or not.[20]

[20] In this example, there is nothing to synchronize John's experience with mine; his life is unfolding in parallel to mine and it is hard to see why it would matter which events are contemporaneous. As we will discuss in subsection 3.1, we can make the example even more extreme by having John fly far off into space somewhere, so that, as far as the theory of relativity is concerned, there really is no absolute answer to the question whether his headache is taking place at the same

Hence, given that the scenario is set up appropriately, I remain unconvinced that there is any significant difference between S_1: and S_2:, and this seems to deal a blow to the βA-theory. Naturally, the βB-theory avoids this blow; but I believe the αA-theory also avoids it, in that John today is just as much "outside the I-Now" as John yesterday, because I am not John. In fact, compared to the βB-theory, the αA-theory does a better job explaining why something about the example seems to change when I myself am brought into it. That is, if we replace "John's" with "my" in the statements above to obtain S_1' and S_2', then it does seem that S_1' has a significance that S_2' does not. S_2' is not an unreasonable statement—it makes sense to appreciate having suffered less than one might have, just as it makes sense to appreciate someone else suffering less than he might have—but only S_1' concerns the immediate presence or absence of suffering, which is the vivid characteristic that imbues "Thank goodness that's over!" examples with their intended significance.[21]

Indeed, both Suhler and Callender (2012) and Green and Sullivan (2015) report on an experimental study by Caruso et al. (2008) in which subjects were asked what would be fair compensation for a particular task. The study found that when subjects were asked to imagine themselves doing the task in the future, they felt that they should be compensated significantly more than when they imagined themselves doing the task in the past; but this effect disappeared when they were asked to imagine someone *else* doing it. Suhler and Callender (2012) take this to invalidate the "Thank goodness that's over" argument, and Greene and Sullivan (2015) argue for complete temporal neutrality in making decisions. (The argument for temporal neutrality is worked out in detail in Sullivan (2018). Hurka (1993, 61) argues that temporal neutrality is appropriate for certain non-hedonic goods, but is convinced that it is not for avoiding pain, by the example from Parfit (1984, 165) that

time as my current experience. If so, caring about simultaneity seems to require a very strong commitment to the βA-theory, as it requires that there be an additional fact about simultaneity, over and above the theory of relativity, that is important for what we should care about, even though no physical measurement could ever tell us whether two events actually were or were not simultaneous in this sense.

[21] Some of this is reminiscent of Turri's (2013) "That's outrageous!" example. Turri argues that just as the appropriateness of statements such as "Thank goodness that's over!" can be used to support presentism, the appropriateness of statements such as "That's outrageous!" can be used to attack it, because it seems perfectly legitimate to be outraged by, say, a past genocide. I consider it telling that "Thank goodness that's over!" examples typically involve oneself and "That's outrageous!" examples typically involve others; this may well be what is driving the difference in conclusions from these examples.

doi: 10.48106/dial.v74.i1.02

we would prefer a more painful operation in the past to a less painful one in the future.) The analysis above suggests that while indeed, the results of the Caruso et al. (2008) study cast doubt on whether the "Thank goodness that's over" argument effectively supports the βA-theory, they are perfectly consistent with this argument supporting the αA-theory.

3 Revisiting Arguments Against the A-theory

In this section, I will revisit some well-known arguments against the A-theory. Subsection 3.1 concerns the argument from special relativity, subsection 3.2 concerns the argument that the direction of time may be a local matter, subsection 3.3 concerns the argument that asks for the rate at which time passes, and subsection 3.4 concerns the argument from time travel and Gödelian universes. In all cases, the αA-theory will be shown to avoid most of the bite that these arguments inflict on the βA-theory, roughly because the arguments hinge on the fact that the Now is global in nature—that is, it stretches across all of space. Because the I-Now is local in nature, the arguments are ineffective against the αA-theory.

3.1 *Special Relativity*

Einstein's theory of relativity has often been invoked to criticize the A-theory. Unlike in a Newtonian universe, in the special theory of relativity, simultaneity is not absolute; rather, whether two events are simultaneous depends on the reference frame. But if there is no absolute simultaneity, then how can there be an absolute Now? Special relativity can also be used to cast doubt on specific arguments in favor of the A-theory—or at least, the βA-theory. For example, let us modify the example from subsection 2.2 by putting John on a faraway planet, so that whether his headache is earlier or later than our own time depends on the reference frame. This seems to make it difficult to hold the position that, in order to know how we should feel about John's headache, it is important to know whether it is in the past or in the future. Now, perhaps there may still be a separate, absolute sense in which John's headache is in the past, even if this is not implied by the theory of relativity. But if there is not, this poses a problem for using the "Thank goodness that's over!" argument in support of the βA-theory—but, importantly, not for using it in support of the αA-theory, because, as discussed in subsection 2.2, in that case the argument is only made about one's own pains rather than those of someone on a faraway planet. Still, we must investigate the implications of relativity for the αA-theory more broadly.

Some (e.g. Markosian 2004) have argued that, in fact, a philosophically austere version of the theory of relativity could explain the empirical evidence without implying that there is no absolute simultaneity. The relation of absolute simultaneity could be added on top of the theory of relativity. For example, one might suppose that there exists a distinguished frame of reference that determines which events are absolutely simultaneous. Positing such a distinguished frame seems a rather awkward and inelegant addition to the theory, one that is rather contrary to the spirit of the theory of relativity and perhaps more in line with older theories of a stationary aether. But, Zimmerman (2007) has argued that such an addition to the physical theory is no different in kind from the addition of a distinguished Now in the first place. That may be so, but it is a further addition, and it seems that, for the sake of parsimony, each addition should at least count against the resulting theory. The analogy is also imperfect. It can at least be argued that we know when the Now is; in contrast, it is not clear whether and how we could ever know what the distinguished frame of reference is. Zimmerman (2011) discusses and responds to all these concerns in far more detail than I can do here, and argues well that they are not fatal to the βA-theory, but it is clear that at least they pose significant challenges.

In any case, the above arguments only concern the βA-theory. In the αA-theory, there is no need for any observer-independent simultaneity at all. While the Now in the βA-theory must be global—in the sense that everywhere in the universe, there are events happening Now, thereby introducing an observer-independent simultaneity relation across all of space—the I-Now in the αA-theory is local. The precise nature of this locality—for example, whether the I-Now is spatially extended—does not matter much for the arguments at hand; what matters is that the I-Now is associated with an observer, and that that observer can be localized in spacetime. Specifically, this ties the I-Now to the frame of reference associated with that observer;[22] if so desired, simultaneity could be determined based on this frame of reference according to the theory of relativity. For that matter, no notion of simultaneity across space is even required for the theory to make sense. While the βA-theory necessitates such a notion—whatever is happening Now across space must

[22] The definition of what constitutes a frame of reference varies. Here, we consider a frame of reference to be determined purely by its state of motion, rather than to also include a coordinate system.

be simultaneous, in an objective sense—it does not seem to pose any problem for the αA-theorist to hold that there is no absolute notion of simultaneity. As far as the αA-theorist is concerned, we can define a notion of simultaneity for convenience, for example the one based on the theory of relativity and the distinguished frame of reference corresponding to the I-Now as just suggested, but none is truly needed. In fact, the problems that the theory of relativity poses for the A-theory have already led to at least one proposal similar to the αA-theory, namely Skow (2009)'s relativistic spotlight theory,[23] in which the spotlight shines locally, not globally.[24]

3.2 *The Direction of Time*

For any version of the βA-theory in which time flows, there needs to be an objective *direction* in which time flows. Presumably, it flows from what we perceive as the past to what we perceive as the future. But if the laws of physics are invariant to time reversal, then these laws do not naturally provide such a direction. It is commonly held that what we perceive as the direction of time is tied to the entropy gradient, and that this entropy gradient may well be reversed in other parts of spacetime. If so, we may imagine a Doppelgänger being that is otherwise very much like ourselves, living its life in such a part, backwards in time from our perspective (Williams 1951; Maudlin 2002). The Doppelgänger would presumably think that *we* have it backwards, that the direction of time's flow is opposite from what we think it is. So what gives us reason to believe that we are the ones to have it right? A key issue here is that presumably, the βA-theory requires time to flow in the same direction everywhere; the direction should be *globally* consistent.[25] It has been argued that we have no reason to believe that the Doppelgänger even has mental

[23] In earlier work, Stein (1968, 18) hints at a similar theory when he contemplates what would result from an argument by Putnam (1967) if one tried to preserve a different intuition about the relationship between what is present and what is real. It is not clear whether he intends at all to defend such a theory.

[24] Hare (2010) and Hare (2009, 48) also make some of the points that I made in this subsection. Fine (2005, 2006) similarly gives a detailed discussion of what, for the realist, should replace the role of times when we take into account special relativity, and concludes that most plausibly frame-time pairs should take their role, in combination with a nonstandard type of realism in which either realities are indexed to different frame-times or reality is fragmented.

[25] The Now is not localized under the βA-theory, so that there is a single Now across space; but if it moves in one direction in one location and in the opposite direction elsewhere, it is hard to imagine that after moving in these opposite directions it remains the *same* Now across these locations.

states at all, by virtue of the fact that the way its life proceeds is so unlike the way ours proceeds (Maudlin 2002). But this seems a rather odd conclusion, since we have supposed that, *mutatis mutandis* for the difference in direction, the Doppelgänger's life is entirely like ours. For a more detailed discussion of this point and these issues more generally, see Price (2011) and references cited therein.

In contrast, the putative existence of persons living in parts of spacetime with a reversed entropy gradient, living their lives backwards in time (from our perspective), poses no problem for the αA-theory. This is because the I-Now is inherently *local* (in both a spatial and a temporal sense), so it does not matter if the entropy gradient is reversed elsewhere; all that matters is what the entropy gradient is *here* (and *now*), because that is what determines the direction in which the I-Now moves. If the I-Now actually tracks a Doppelgänger at some point, it does not appear to pose any problem for the theory for it to then move in the opposite direction. (This may pose problems for some of the specific illustrative versions presented earlier in section 1, but it poses no problem for the other versions.) We can view *external* time as nothing more than a dimension through which the I-Now travels.

Taking this to an extreme, we may even imagine a machine that transports you to another region of space where the entropy gradient is reversed relative to ours, and that transforms you into a Doppelgänger there. You will, in some sense, continue your life there uninterrupted, except moving in the opposite temporal direction. Of course being transported to another region of space is likely to be a bit shocking; but, if such scenarios are possible at all, there seems to be no reason to believe that your experiences will be any different than they would have been if instead you had been transported to a region of space that happens to have the same entropy gradient (and not been transformed into a Doppelgänger). Accommodating this intuition is easy under the αA-theory; for example, the I-Now could simply jump along with you and then start moving backwards (from our initial perspective). On the other hand, this example appears problematic for versions of the βA-theory that require a globally defined direction of time, because such a theory would lead to the conclusion that one of the two halves of your life is lived, in an *absolute* sense, backward. If we believe Maudlin (2002)'s argument, we would then conclude that you had real mental states in only half of your life. This seems to be an odd conclusion. If near the end of your life you were transported back to the original spacetime region, the suggestion that you had not had any real mental states since the original transportation event would seem utterly bizarre to you!

3.3 *The Rate of Time's Passage*

Opponents of the A-theory (or βA-theory) have also criticized it as follows: if the Now moves, what is the rate at which it moves? It has been argued that if one says that it moves at 1 second per second, this poses a problem for the theory, because one can cancel the units of seconds and conclude that the rate is simply 1, and (supposedly) 1 is not a rate (e.g. Olson 2009). Now, the idea that a unitless rate is not a rate is simply nonsense. This has been convincingly argued elsewhere: Skow (2011) uses the example of sociologists tracking what the "most common birth year" in the population is. One would expect the most common birth year to generally increase by roughly 1 year every year, though the rate may be higher or lower than 1 depending on demographic phenomena. In any case, the rate is unitless (one might just as well say the rate is approximately 1 decade per decade). The example is convincing to me, and clearly many other examples of sensible unitless rates can be provided. One such example is particularly relevant here: due to relativity, satellites and astronauts on the International Space Station age at a slightly different rate than objects and people on the surface of the Earth. The amount of time that such a satellite or astronaut experiences per unit of Earth surface time is a unitless rate.[26] This example actually seems to pose a more serious problem for the answer that time moves at "1 second per second"—if the idea is to think of time as moving globally rather than just locally, then in just *whose* seconds are we measuring this rate? In any case, a weaker version of the original criticism seems to hold up: the question only allows uninformative answers. The answer that it moves at "1 second per second" seems tautological. We

[26] One might counter that these conditions in fact correspond to different units, namely Earth surface seconds and ISS seconds, so that we in fact do not obtain a unitless rate. But this misses the point that a second denotes the same amount of aging for the people in each condition. The unitless rate indicates how much faster people in one condition age than those in the other, and for this comparison no units are needed. Similarly, we need no units to say that one person is 1.2 times as tall as another. That the rate being unitless is meaningful is further illustrated by the fact that it can be both above and below 1, because of the opposing effects of relative velocity time dilation and gravitational time dilation; there is an orbit, about half the radius of the Earth above the surface, at which the rate is 1 (Ashby 2002). The rate being 1 at this orbit is not just a meaningless consequence of how we defined the units; it is the orbit at which astronauts age equally fast as those on the surface.

could instead introduce the concept of *supertime* to track the Now's motion through time, so that at different points in supertime, the Now is at a different time. (For a detailed discussion of the metaphor of supertime, see Skow 2012.) Then, we can ask how many seconds pass per supersecond. However, there seems to be every reason to simply define the supersecond so that the answer becomes "1 second per supersecond," which remains uninformative.

In the αA-theory—or, at least, in versions of it where the I-Now moves along with a person through time (see section 1)—the question of how fast the I-Now moves does not pose such problems. First, the fact that on a space station, a different amount of time is experienced to pass no longer poses any problem, because the I-Now is local, so there is no requirement that time passes at the same rate everywhere. Moreover, the question of how fast the I-Now moves can have more interesting answers. In the relativistic example above, it is natural to respond that the I-Now moves at a different rate when it is associated with an astronaut in orbit than it does when it is associated with a person on the surface. Alternatively, let us put relativity aside for a moment and focus on the I-Now's experiential aspect instead. One might reasonably hold that the I-Now moves through external (i.e. clock) time at a different rate when it is associated with a person who is under anesthesia than it does when it is associated with someone who is highly alert.

If we allow ourselves to speculate, a computational[27] theory might be used to unify these two examples: consider a person's "clock speed"—the number of mental operations, according to some suitable definition, per (Earth surface) second—and take this to determine the rate at which the I-Now moves. Specifically, let us define a supersecond so that there is always exactly one mental operation per supersecond. Then, the number of (Earth surface) seconds per supersecond—which is just the reciprocal of the clock speed defined above—will vary in the different scenarios above, in a way that conforms with our intuitions. Focusing on Earth surface seconds per supersecond (regardless of the location of the person) simultaneously addresses both the relativistic and the experiential components of the scenarios, and also allows us to handle mixed cases, such as a space station inhabitant who is under anesthesia. In such a case, the number of mental operations per Earth surface second can be written as the number of mental operations per space station second, multiplied by the number of space station seconds per Earth surface second,

[27] It is important to hold a sufficiently broad view of "computation" here; such broad views are common among those working on the theory of computation. Alternatively, and less ambitiously, the reader may just view this as a suggestive analogy to the clock speed of a computer.

doi: 10.48106/dial.v74.i1.02

thereby separating out the experiential and relativistic components, respectively. This shows that these two components are compatible. Per the theory of relativity, there is nothing special about Earth surface seconds, as opposed to space station seconds or Mars surface seconds; they are just different ways to measure external time.

Supertime, so defined, perhaps more naturally corresponds to our sense of passage, leaving regular time (as tracked by clocks) in the more modest role of a dimension through which we happen to pass, as noted earlier. That is, this notion of supertime would allow us to give metaphysical meaning to the idea of time passing more or less quickly from a subjective viewpoint. Of course, this view may conflict with other intuitions that we have developed. In our ordinary experience of time, relativistic issues do not come into play, and our waking experience of how fast time passes is usually fairly stable. Given this, we tend to conceive of time as objective, and treat any variance in how we perceive its passage as a mere error in estimation. For the current purpose, I believe such intuitions are misleading. The following two examples are intended to illustrate that it is in fact quite natural to assign primary importance to the notion of supertime as defined here. In each of them, we will imagine a choice between two alternatives that result in you having different amounts of time but equal amounts of supertime left in your life. I argue that you should be (close to) indifferent between the options in both scenarios.

Example 1. It is the year 2400, and you are part of a group of people on a lifelong space voyage. The group is about to split up into two subgroups that will take separate spacecraft. It is common knowledge that the two subgroups will never communicate again, either with each other or with the people left on Earth. You get to choose in which subgroup you will be. They are indistinguishable, except the two spacecraft will move to orbits around different massive bodies, with different relativistic time dilations. If you choose to be on spacecraft 1, your life will therefore be shorter in Earth time than it would be on spacecraft 2. As a result, your first reaction may be that you would prefer to be on spacecraft 2. But, I argue, upon closer inspection there is little reason for this. This is because, to make up for the shorter amount of Earth time in your life on spacecraft 1, correspondingly more events will happen per unit of Earth time on spacecraft 1. You would experience entirely similar lives on the two spacecraft, with equally many interesting events taking place on both. If it were possible to communicate from Earth to the spacecraft, you might prefer being on spacecraft 2 because (for example) more papers, books, and movies would be produced on Earth and sent to spacecraft 2 for your consumption during your life. But we have assumed that such communication is impossible. As far as I can see, there does not seem to be any compelling reason to have a preference about on which spacecraft you continue your voyage.

Example 2. It is again the year 2400, but this time we will stay on the surface of the Earth. After a long and happy life, you have regrettably contracted an incurable disease that, if left untreated, will kill you almost immediately. Unfortunately, the only possible treatments will put you in a type of comatose state until your death. You will, however, have wonderful dreams in this state. Due to secrecy issues, your friends and family will never be made aware of your predicament. There is no chance at all that any new treatment will become available during the remainder of your life. You have a choice between medications M_1 and M_2. Compared to M_1, M_2 would keep you alive for twice as long, but would allow your brain to process at only half the rate. Your first reaction may be that you would prefer to receive M_2. But again, I argue, upon closer inspection there is little reason for this. Because of the difference in brain processing rates, you would have equally many wonderful dreams under the two medications. If your friends and family could visit you in your comatose state, you might prefer for them to have that option for a longer or shorter period of time, but we have ruled this out. If you had hopes that scientists could develop a cure, you would prefer M_2 to give the scientists more time, but we have also ruled this out. As far as I can see, there does not seem to be any compelling reason to have a preference about which medication you receive.

In summary, to the extent that the question about the rate at which the Now moves poses a problem for the βA-theory, it does not pose this problem for the αA-theory, since for the latter the answer to the question need not be tautological.

3.4 *Time Travel and Gödelian Universes*

A final criticism of the (β)A-theory is that it does not make much sense of time travel scenarios. Following Lewis (1976), it seems natural to distinguish between *external* time and the time traveler's *personal* time. But if one takes external time seriously in the metaphysical sense, as would be expected of a βA-theorist, it would appear one cannot simultaneously do the same for personal time. This, in turn, necessitates unintuitive attitudes towards time travel. The following passage by Sider (2005, 333) illustrates this perfectly:

> But if personal time bears little similarity to external time then "personal time" is merely an invented quantity, and is misleadingly named at that. That I will view a dinosaur in my personal future amounts merely to the fact that I once viewed a dinosaur, and moreover that this is caused by my entry into a time machine. Since this fact bears little resemblance to the facts that constitute a normal person's genuine future, I could not enter the time machine with anticipation and excitement at the thought of seeing a dinosaur, for it is not true that I am *about* to see a dinosaur, nor is the truth much *like* being about to see a dinosaur. If anything, I should feel fear at the thought of being annihilated by a device misleadingly called a "time machine". The device causes it to be the case that I once viewed a dinosaur, but does not make it the case in any real sense that I *will* view dinosaurs.

Perhaps there is a way out of this conclusion for the βA-theorist, but I cannot see it. Or perhaps she is willing to bite the bullet and accept the conclusion that (at least backward) time travel is to be avoided at all cost. In any case, the αA-theorist avoids this issue. For her, personal time is what is taken seriously, and she can legitimately look forward to—if this is in fact something to look forward to—her encounter with a dinosaur.[28]

Closely related to the issue of time travel is that of Gödelian universes that cannot be given a global temporal ordering. The theoretical possibility of such universes perhaps poses a problem for some versions of the βA-theory. The αA-theory, however, does not require any global temporal ordering. For versions of the αA-theory with a moving I-Now, one may yet worry if such universes do not create different problems. For example, Dieks (2006) discusses an example by Reichenbach (1958, 141–42) in which a person loops around to meet his earlier self again at a particular point in spacetime. Dieks, who argues for a B-theoretic notion of local becoming, argues that this example illustrates that even a local type of spotlight is problematic. He argues that when the spotlight shines on the region in spacetime where the younger and older versions of the

[28] Well, she may still hesitate, to the extent that it is not obvious that the presence of experience, the I-Now, will follow her through the time machine rather than go somewhere else. As an example that illustrates this ambiguity, it may be one of these unmarketable time machines that also leave behind a badly burned body, apparently alive for a few more seconds, where the traveler entered the time machine. (See Hare (2009, 58) for a similar example.) But at least her believing that it will follow her back in time (rather than transitioning to a different person at the same time, or staying with a burnt body) would not cause any inconsistency with her other beliefs.

person meet, there must in fact be two distinct spotlights, one that will travel with the younger version and one that will travel with the older version. Then, the spotlight associated with the younger version loops around as that version becomes the older version, eventually reaching the same region again. By the same reasoning as before, we will again need two spotlights at this point. But the other spotlight, the one that was initially associated with the older version, is not available for the task, being meanwhile associated with an even older version. So we will need a third spotlight, and so on ad infinitum, which seems problematic.

But it is easy to find an escape from Dieks' argument. The fact that the two versions of the person are (roughly) at the same point in spacetime does not imply that the spotlight shines on them simultaneously *in the supertime sense*. That is, the "same" spotlight might earlier (in supertime) light up the younger version only (i.e. that version's experience at that point) and later (in supertime) the older version only. Hence, there is no need to introduce additional spotlights when the meeting point is reached. This illustrates one advantage of associating the spotlight with person-stages (I-Now) rather than with small regions of spacetime (Here-Now): even though the younger and the older version are both in (roughly) the same location at the same time, they correspond to different person-stages. This requires, of course, that in this type of scenario we associate the I-Now with a person-stage (where a younger and an older version of the same person at the same time are still considered separate person-stages), rather than with a pair of a person and a time, which in this case might pick out both person-stages. This interpretation of the I-Now in any case aligns better with the other arguments presented in this paper. For example, it seems hard to imagine the (simultaneous) presence *simpliciter* of the *combination* of both person-stages. Also, the older person-stage may think, looking at the younger person-stage, "Thank goodness I am no longer that immature!" The idea that the spotlight was previously (in the supertime sense) associated with the younger person-stage and now with the older person-stage seems to capture the significance of this statement well. Finally (and more speculatively), if we imagine the brain of the older stage to have slowed down and no longer to be processing at the rate of his younger self, associating the I-Now with person-stages would allow us to say that the I-Now moves at a different rate with respect to external time when associated with each of these two person-stages.

4 Conclusion

Upon inspection, key criticisms of the A-theory are only effective as criticisms of the βA-theory, and key arguments in favor of the A-theory are much more convincing as arguments for the αA-theory. To the extent I have succeeded in showing that A-theorists are rationally compelled to be α-theorists as well, surely many will interpret this as a significant blow to the A-theory because they consider the α-theory implausible. Nevertheless, some philosophers may well be willing to adopt some version of the αA-theory (Hare being an obvious example). As I emphasized earlier, a detailed discussion of the relative merits of the αA-theory and the βB-theory is outside the scope of this paper. Such a discussion is sure to revisit many familiar arguments in the philosophy of time and modality (and mind), and is unlikely to reach a swift conclusion.[29] I do hope to have convinced the reader that the αA-theory will fare better in such a comparison than the βA-theory. The former has an internal consistency that allows it to escape some of the more damaging criticisms to which the latter has fallen prey.[*]

Vincent Conitzer
Duke University
conitzer@cs.duke.edu

References

Ashby, Neil. 2002. "Relativity and the Global Positioning System." *Physics Today* 55 (5): 41–47. doi:10.1142/9789812700988_0010.

Balashov, Yuri. 2005. "Times of Our Lives: Negotiating the Presence of Experience." *American Philosophical Quarterly* 42 (4): 295–309.

Bergmann, Michael. 1999. "(Serious) Actualism and (Serious) Presentism." *Noûs* 33 (1): 118–32. doi:10.1111/0029-4624.00145.

Butterfield, Jeremy. 1984. "Seeing the Present." *Mind* 93 (370): 161–76. doi:10.1093/mind/xciii.370.161.

Cameron, Ross P. 2015. *The Moving Spotlight: An Essay on Time and Ontology*. Oxford: Oxford University Press.

[29] This seems all the less likely given that the problem connects to other challenging problems, such as the Sleeping Beauty problem—see e.g. Conitzer (2015).

[*] I am thankful to anonymous referees who provided especially thorough and helpful comments, which significantly improved the paper.

CARUSO, Eugene M., Daniel T. GILBERT, and Timothy D. WILSON. 2008. "A Wrinkle in Time: Asymmetric Valuation of Past and Future Events." *Psychological Science* 19 (8): 796–801. doi:10.1111/j.1467-9280.2008.02159.x.

CHALMERS, David J. 2010. *The Character of Consciousness*. Oxford: Oxford University Press.

CONITZER, Vincent. 2015. "Can Rational Choice Guide Us to Correct *De Se* Beliefs?" *Synthese* 192 (12): 4107–19. doi:10.1007/s11229-015-0737-x.

———. 2019. "A Puzzle about Further Facts." *Erkenntnis* 84 (3): 727–39. doi:10.1007/s10670-018-9979-6.

DEASY, Daniel. 2017. "What Is Presentism?" *Noûs* 51 (2): 378–97. doi:10.1111/nous.12109.

DIEKS, Dennis. 2006. "Becoming, Relativity and Locality." In *The Ontology of Spacetime*, edited by Dennis Dieks, 1:157–76. Philosophy and the Foundations of Physics Series. Amsterdam: Elsevier Science Publishers B.V.

DORR, Cian, and Jeremy GOODMAN. 2020. "Diamonds Are Forever." *Noûs* 54 (3): 632–65. doi:10.1111/nous.12271.

FINE, Kit. 2005. "Tense and Reality." In *Modality and Tense: Philosophical Papers*, 261–320. Oxford: Oxford University Press.

———. 2006. "The Reality of Tense." *Synthese* 150 (3): 399–414. doi:10.1007/s11229-005-5515-8.

GREENE, Preston, and Meghan SULLIVAN. 2015. "Against Time Bias." *Ethics* 125 (4): 947–70. doi:10.1086/680910.

HARE, Caspar. 2007. "Self-Bias, Time-Bias, and the Metaphysics of Self and Time." *The Journal of Philosophy* 104 (7): 350–73. doi:10.5840/jphil2007104717.

———. 2009. *On Myself, and Other, Less Important Subjects*. Princeton, New Jersey: Princeton University Press.

———. 2010. "Realism about Tense and Perspective." *Philosophy Compass* 5 (9): 760–69. doi:10.1111/j.1747-9991.2010.00325.x.

HELLIE, Benj. 2013. "Against Egalitarianism." *Analysis* 73 (2): 304–20. doi:10.1093/analys/ans101.

HURKA, Thomas. 1993. *Perfectionism*. New York: Oxford University Press.

JOHNSTON, Mark. 2010. *Surviving Death*. Princeton, New Jersey: Princeton University Press.

LEWIS, David. 1976. "The Paradoxes of Time Travel." *American Philosophical Quarterly* 13: 145–52. doi:10.1002/9781118922590.ch26.

LIAO, Shen-yi. 2012. "What Are Centered Worlds?" *The Philosophical Quarterly* 62 (247): 294–316. doi:10.1111/j.1467-9213.2011.00042.x.

LIPMAN, Martin A. 2015. "On Fine's Fragmentalism." *Philosophical Studies* 172 (12): 3119–33. doi:10.1007/s11098-015-0460-y.

MARKOSIAN, Ned. 2004. "A Defense of Presentism." In *Oxford Studies in Metaphysics*, edited by Dean W. Zimmerman, I:47–82. Oxford: Oxford University Press.

Maudlin, Tim. 2002. “Remarks on the Passing of Time.” *Proceedings of the Aristotelian Society* 102: 259–74. doi:10.1111/j.0066-7372.2003.00053.x.

McTaggart, J. McT. Ellis. 1908. “The Unreality of Time.” *Mind* 17 (68): 457–74. doi:10.1093/mind/xvii.4.457.

Merlo, Giovanni. 2016. “Subjectivism and the Mental.” *Dialectica* 70 (3): 311–42. doi:10.1111/1746-8361.12153.

Olson, Eric T. 2009. “The Rate of Time’s Passage.” *Analysis* 69 (1): 3–9. doi:10.1093/analys/ann001.

Parfit, Derek. 1984. *Reasons and Persons*. Oxford: Oxford University Press.

Price, Huw. 2011. “The Flow of Time.” In *The Oxford Handbook of Philosophy of Time*, edited by Craig Callender, 276–311. Oxford Handbooks. Oxford: Oxford University Press.

Prior, Arthur Norman. 1959. “Thank Goodness That’s Over.” *Philosophy* 34 (128): 12–17. doi:10.1017/s0031819100029685.

Prior, Arthur Norman, and Kit Fine. 1977. *Worlds, Times and Selves*. London: Gerald Duckworth & Co. Edited by Kit Fine; based on manuscripts by Prior with a preface and a postscript by Kit Fine.

Putnam, Hilary. 1967. “Time and Physical Geometry.” *The Journal of Philosophy* 64 (8): 240–47. doi:10.2307/2024493.

Reichenbach, Hans. 1928. *Philosophie der Raum-Zeit-Lehre*. Berlin: Walter de Gruyter. English translation:Reichenbach (1958)

———. 1958. *The Philosophy of Space and Time*. Mineola, New York: Dover Publications. English translation of Reichenbach (1928) by Maria Reichenbach and John Freund.

Sider, Theodore. 2005. “Traveling in A- and B- Time.” *The Monist* 88 (3): 329–35. doi:10.5840/monist200588326.

Skow, Bradford. 2009. “Relativity and the Moving Spotlight.” *The Journal of Philosophy* 106 (12): 666–78. doi:10.5840/jphil20091061224.

———. 2011. “On the Meaning of the Question ’How Fast Does Time Pass?’.” *Philosophical Studies* 155 (3): 325–44. doi:10.1007/s11098-010-9575-3.

———. 2012. “Why Does Time Pass?” *Noûs* 46 (2): 223–42. doi:10.1111/j.1468-0068.2010.00784.x.

Smith, Nicholas J. J. 2011. “Inconsistency in the A-Theory.” *Philosophical Studies* 156 (2): 231–47. doi:10.1007/s11098-010-9591-3.

Stein, Howard. 1968. “On Einstein-Minkowski Space-Time.” *The Journal of Philosophy* 65 (1): 5–23. doi:10.2307/2024512.

Suhler, Christopher, and Craig Callender. 2012. “Thank Goodness That Argument Is Over: Explaining the Temporal Value Asymmetry.” *Philosophers’ Imprint* 12 (15).

Sullivan, Meghan. 2018. *Time Biases: A Theory of Rational Planning and Personal Persistence*. Oxford: Oxford University Press.

TURRI, John. 2013. "That's Outrageous." *Theoria* 79 (2): 167–71. doi:10.1111/theo.12005.

VALBERG, Jerome J. 2007. *Dream, Death, and the Self*. Princeton, New Jersey: Princeton University Press.

———. 2013. "The Temporal Present." *Philosophy* 88 (3): 369–86. doi:10.1017/s0031819113000016.

WILLIAMS, Donald Cary. 1951. "The Myth of Passage." *The Journal of Philosophy* 48 (15): 457–72. doi:10.2307/2021694.

ZIMMERMAN, Dean W. 2005. "The A-Theory of Time, the B-Theory of Time, and 'Taking Tense Seriously'." *Dialectica* 59 (4): 401–57. doi:10.1111/j.1746-8361.2005.01041.x.

———. 2007. "The Privileged Present: Defending an 'A-Theory' of Time." In *Contemporary Debates in Metaphysics*, edited by Theodore Sider, John Hawthorne, and Dean W. Zimmerman, 211–25. Contemporary Debates in Philosophy. Malden, Massachusetts: Blackwell-Wiley.

———. 2011. "Presentism and the Space-Time Manifold." In *The Oxford Handbook of Philosophy of Time*, edited by Craig Callender, 163–245. Oxford Handbooks. Oxford: Oxford University Press.

Determinism, "Ought" Implies "Can" and Moral Obligation

Nadine Elzein

Haji argues that determinism threatens deontic morality, not via a threat to moral responsibility, but directly, because of the principle that "ought" implies "can". Haji's argument requires not only that we embrace an "ought" implies "can" principle, but also that we adopt the principle that "ought" implies "able not to". I argue that we have little reason to adopt the latter principle, and examine whether deontic morality might be destroyed on the basis of the more commonly embraced "ought" implies "can" principle alone. I argue that despite what look like initially compelling reasons why we might suppose that this weaker conclusion is similarly destructive to deontic morality, we actually have good reason to doubt that it has any practical relevance for moral deliberation at all.

While most of the literature on morality and determinism focuses on threats to moral responsibility, determinism might be thought to threaten morality on separate grounds. Haji draws on the popular principle that "ought" implies "can", in order to show that determinism undermines deontic morality (1998, 1999, 2002, 2019). Similar arguments are presented by Lockie (2018), although Lockie, unlike Haji, does not intend to defend scepticism about obligation, but rather to show that any such scepticism is inherently self-defeating.

By "deontic morality", Haji has in mind any moral use of the terms "ought" and "ought not", as well as moral judgements of right and wrong. While he concedes that judgements of moral "good" and "bad" may still make sense within a deterministic framework, he argues that the action-demanding normative terms associated with obligations and prohibitions would be seriously undermined. Determinism precludes moral duty.

However, as Haji himself makes explicit, in order to reach this conclusion, we need not only an "ought" implies "can" principle, but also an "ought" implies "able not to" principle (2002, 28). A similar principle is found in Lockie's work (2018, 181). I will argue, firstly, that even if we accept the popular "ought" implies "can" principle, there are good reasons to reject any "ought" implies "able not to" principle. Secondly, without the "ought" implies "able not to" principle, such arguments are limited to establishing a much weaker conclusion; we cannot conclude that there are no moral duties at all, only that there are no *unfulfilled* moral duties. Thirdly, while this weaker conclusion may look similarly problematic at first sight, from a practical perspective it actually makes very little difference to morality.

1 Determinism, Ability, and "ought" Implies "can"

The principle that "ought" implies "can" has certainly seemed compelling to many,[1] although it's not uncontroversial.[2] Haji originally calls his "ought" implies "can" principle "K", and then later "Kant's Law/Obligation". But for present purposes, let us simply call this sort of principle "OIC" (so as to match the broader class of principles under discussion). Haji (2002, 14) formulates his version of OIC roughly as follows:

> OIC. As of time t, an agent S, ought morally to do something A at time t^* (where t^* may either be t or a time later than t) only if S can, as of t, do A at t^*; and, as of t, S ought not to do A at t^* only if S can, as of t, not do A at t^*.

According to this principle, an agent only ought to do something if she actually can do it, and ought only to refrain from doing something if she actually can refrain from doing it.

[1] The principle is commonly thought to originate with Kant, and was famously defended by Moore (1922). Since then it is more often taken to be a basic platitude than explicitly argued for, but there are some explicit defences of the principle: see Sapontzis (1991), Griffin (1992), Streumer (2003, 2007, 2010), and Vranas (2007). For defences of related principles, see Graham (2011) and Kühler (2013).

[2] For some critiques, see Lemmon (1962), Williams (1965), Brouwer (1969), Trigg (1971), Fraassen (1973), Brown (1977), Sinnott-Armstrong (1984, 1988), Rescher (1987, ch. 2, 26–54), Saka (2000), Fischer (2003), and Heintz (2013). Cf. Kekes (1984) and Stern (2004).

1.1 *The Analysis of "can"*

Given that there are broad variations in the way that we might interpret "can",[3] there are also variations in the way that we might interpret OIC. Haji's (2002, 23) most moderate definition is as follows:

> MODERATE OIC. Agent *S* ought to do something *A*, only if *S* has the opportunity to do *A*, is physically and psychologically able to do *A*, and *A*'s accomplishment is not "strictly out of *S*'s control".

While this is taken to be the bare minimum required for ability, Haji adds that it may also require being motivationally able, and having the right sort of "know-how" (2002, 16–24).

Physical and psychological possibility are fairly straightforward notions. Plausibly an agent is only "able" to perform actions that are consistent with their psychological characteristics and their physical abilities. The inclusion of the stipulation that the agent must be "psychologically able" may, however, seem controversial. It means that an agent with a strong aversion, say, may count as unable to do something, even if she could succeed in doing it should she choose to. One reason we might nonetheless endorse this reading, as Haji points out, is that it is natural to suppose that an agent with a serious enough phobia might be excused for her failure to do something that her phobia prevents her from doing. For instance, we would not typically consider an agent "able" to save a drowning child if a severe phobia rendered her incapable of entering the water (Haji 2002, 22).

Moreover, endorsing a relatively strong sense of "can" may prove indispensable to the argument as a whole. That is because the argument aims to establish that the ability to do otherwise is ruled out by determinism, where this involves the very *same* sense of ability for which it will be true that "ought" implies "can". Any weakening of the sense of "can" utilised in the OIC principle may risk introducing a corresponding weakening of the argument for supposing that determinism rules out the ability to do otherwise in precisely

[3] Among other points of contention, there is a long-standing dispute about whether "can" ought to be analysed conditionally (Moore 1903; Ayer 1946; Smart 1961; Schlick 1939; Lewis 1981; Berofsky 2002), non-conditionally (Campbell 1951; Chisholm 1964; Lehrer 1968; Inwagen 1983, 2000, 2004, 2008; Kane 1999; Clarke 2009; Grzankowski 2014), or dispositionally (Smith 1997, 2003; Vihvelin 2004, 2011, 2013; Fara 2008). Even within these camps there is significant scope for disagreement. For more general discussions, see also Kratzer (1977), Mele (2003), Maier (2015), and Weir (2016).

that sense. For example, Haji notes that if we supposed a merely conditional analysis of “can” would do, according to which the ability to do otherwise simply requires that the agent could do otherwise if she chose to, then this would make it dubious to suppose that determinism rules this ability out (2002, 67–68).

In fact, Haji argues that even if such conditional abilities are present, determinism robs us of the opportunity to do otherwise. If any factors, internal or external, prevent an agent from *exercising* some skill they have, then this will constitute a barrier to their having the opportunity to exercise it (2002, 22).[4]

Finally, the “control” requirement is supposed, at the very least, to rule out having the “ability” to do things that happen purely by fluke (Haji 2002, 22). In analysing such control, Haji cites Vihvelin, who states: “We make judgments about ability on the basis of evidence of a reliable causal correlation between someone’s attempts to do a certain kind of act and the success of her attempts.” (2000, 142). This sort of control neither entails nor is entailed by possession of the other senses of “ability”. Plausibly, an agent’s phobia may make her psychologically and motivationally unable to purchase a pet snake, but doing so may not be “strictly out of her control”; were she to try, she could reliably succeed. Similarly, if a golf novice hits a hole in one on her first attempt, this certainly shows that she is physically able to hit a hole in one, but if it is an unrepeatable fluke, then it will still be “strictly out of her control”.

1.2 *Determinism and Obligations*

Haji and Lockie use rather complex arguments to reach the conclusion that determinism rules out all obligations. Moreover, Lockie’s argument incorporates the additional goal of showing that any argument in favour of determinism would be self-defeating, and Haji’s argument incorporates his attempt to show that if nothing is obligatory, then nothing is right or wrong either. I am not going to address the latter part of Lockie’s argument,[5] and I am not going to consider whether Haji is right to suppose that wrongness and rightness de-

[4] I am doubtful about the idea that the very same sense of “can” that’s at issue in OIC is also the sense in which the ability to do otherwise might plausibly be ruled out by determinism. We have already noted that if we invoke weaker definitions of “able to” in our OIC principle, it will be difficult to establish that the relevant abilities are threatened by determinism. But for the purposes of this discussion, I will simply grant this point. See Haji (2002, 60–65) for his own arguments to this effect.

[5] I have examined Lockie’s transcendental argument in more detail elsewhere (Elzein and Pernu 2019).

pend on obligation. While this claim has been contested,[6] I am happy to grant it. Moreover, in what follows, it is the status of actions as obligatory (rather than right or wrong) that will be the prime focus. So for present purposes, we can work with a simplified version of the argument, which might go as follows:

1. If determinism is true, no agent is ever able to act otherwise than they do act. (basic premise)
2. If no agent is ever able act otherwise than they do act, then no agent ever has an obligation to act otherwise than they do act. (premise derivable from OIC)
3. If determinism is true, no agent ever has an obligation to act otherwise than they do act. (from 1 and 2, via hypothetical syllogism)

While 3 is an interesting conclusion, it is weaker than the the one that is ultimately defended by either Haji or Lockie. It does not entail that if determinism is true, there are no obligations, merely that that there are no *unfulfilled* obligations. It leaves open that agents sometimes both have and fulfil moral duties. In order to reach the stronger conclusion, that there are no obligations at all, Haji introduces a parallel principle, which he calls "CK" (2002, 28). Lockie (2018, 182) puts forward a similar principle. Elsewhere, Haji gives the same sort of principle different titles, such as "Kant's Law/Impermissible" (Haji 2019, 8) or "Obligation/Alternate" (Haji and Herbert 2018a, 186). Let us simply call this whole class of principles "OIANT principles" (for "ought" implies "able not to"). Haji (2002, 28) defines the relevant sort of principle, omitting the temporal indices, as follows:

> OIANT. If one ought to do *A*, then one can refrain from doing *A* (and if one ought not to do *A*, then one can do *A*).

If we grant OIANT, we can also establish that there are no obligations to do what we actually do, given our inability to do otherwise. A simplified argument of this form runs as follows:

1. If determinism is true, no agent is ever able to act otherwise than they do act. (basic premise)

[6] See Pereboom (2001, 141–47) for an objection, and Haji (2002, 51–52) for his defence.

doi: 10.48106/dial.v74.i1.03

2. If no agent is ever able act otherwise than they do act, then no agent ever has an obligation not to act otherwise than they do act. (premise derivable from OIANT)
3. If determinism is true, no agent ever has an obligation not to act otherwise than they do act. (from 1 and 2, via hypothetical syllogism)
4. If determinism is true, no agent has an obligation to act as they actually do act. (from 3, an equivalence through double negation)

The final step from 3 to 4 is valid provided we grant that "not acting otherwise" entails "acting as one actually does". For present purposes, "acting as one actually does" should be understood broadly, so as to be fulfilled if the agent does not act otherwise; hence it should include the agent's inaction, if the agent in question is not actually doing anything. Granted this broad reading, it should be uncontroversial that "not acting otherwise" directly entails "acting" as one actually does. It should be similarly obvious, granted this broad reading, that premise 2 is entailed by OIANT.

The first argument shows that, given determinism, no agent has an obligation to act otherwise than they do act. The second argument shows that, given determinism, no agent has an obligation to act as they actually do either. Between the two arguments, this rules out all moral obligations.

While the first argument appears compelling, the second argument seems considerably weaker. The principle upon which it rests, OIANT, seems more dubious than the principle invoked by the first argument, OIC. If we reject the argument from OIANT to the conclusion that if determinism were true, no one would be obligated to do what they actually do, then we are left with a weaker conclusion: that if determinism were true, no one would be morally obligated to act otherwise than they do act.

2 How Plausible is OIANT?

Haji offers various lines of argument in favour of accepting OIANT: the first is a simple appeal to symmetry between OIC and OIANT. Lockie's work also draws on the intuition that there ought to be symmetry between such principles. However, even if we doubt that there is any obvious *inherent* reason to suppose that the two principles are symmetrical, we might argue that we ought to accept such symmetry on the basis that both principles are taken to be motivated primarily by a two-way freedom requirement (this seems to be the supposed basis of the symmetry for Haji). Haji also offers a "theory-fuelled" argument, which appeals to a particular analysis of obligation. I will argue that OIANT is, at least on the face of it, inherently implausible before going on to deal with each of these arguments in turn.

2.1 *The Prima Facie Implausibility of OIANT*

It has already been noted that psychological ability is crucially included in the definition of "able to" invoked in Haji's OIC and OIANT principles. In light of this, however, "ought" implies "able not to" has some undesirable implications. Many actions that seem obviously morally prohibited are also psychological impossibilities for most psychiatrically well-adjusted individuals. For instance, my psychology is such that I could not take a chainsaw and use it to saw off the arms of a small child. To be clear, I don't mean a child that has gangrene, say, and needs those limbs removed urgently on pain of death, but a perfectly healthy child; one whose limbs I have no reason to remove. In fact, I could not do such a thing even if I *were* offered reasons, if they were of the wrong sort: e.g. I could not saw off the arms of a child for a monetary incentive (even if I were offered a very reasonable market rate). Does this entail that it is not morally obligatory for me to refrain from sawing off the arms of small children?

This conclusion seems counterintuitive. It is the fact that such an action would be morally reprehensible which may well, in this case, explain *both* my irresistible aversion to it *and* my reasons for supposing that it is morally obligatory that one refrains from such behaviour.

Unlike Haji, I think it is plausible to suppose that my inability to do such a thing entails that I cannot be held responsible for not doing it, and hence deserve no praise.[7] The moral expectation that I refrain from dismembering small children is a very easy standard for me to meet. It seems close to the bare minimum you might reasonably expect of me, so I hardly deserve a medal. But it seems one thing to say that I don't deserve praise, and quite another to say that sawing off the arms of small children would not be morally impermissible. We are usually quite happy to talk about being psychologically

[7] Haji is persuaded on the basis of Frankfurt's argument (1969) that, despite the threat to deontic morality, determinism poses no threat to moral responsibility (1998, 2002). See also Haji and McKenna (2004, 2006). Obviously, however, given the threat to deontic morality, determinism entails that there would be no right or wrong actions to actually blame or praise agents for. In contrast, I remain sceptical about whether Frankfurt-style examples really do establish that the ability to do otherwise is irrelevant to moral responsibility (Elzein 2013, 2017).

doi: 10.48106/dial.v74.i1.03

compelled to do things that we also have a duty to do. We might even suppose that it is the very *fact* that something is perceived as morally prohibited that (at least sometimes) explains an agent's psychological aversion to doing it.

The principle that "ought" entails "able not to" surely seems dubious. We ought to accept it only if we are offered very compelling arguments.

2.2 *The Defence from Apparent Symmetry*

The first argument appeals to the apparent symmetry between "ought" implies "can" principles and "ought not" implies "can" principles (along with, presumably, the latter's complement stipulation, that "ought" implies "able not to"). Haji argues "that it is difficult to see why control requirements of deontic obligatoriness would differ, in this respect from control requirements of deontic wrongness" (2002, 29). He interprets OIC as postulating an alternative possibilities condition as a control requirement for obligatory actions, and supposes that similar considerations would count in favour of accepting an alternative possibilities condition on prohibited ones.

Even "ought" implies "can" is controversial, but it has a strong history of philosophical support behind it and it seems highly intuitive. "ought" implies "able not to", in contrast, has nothing like the same standing. As Nelkin notes, the principle is not usually seen as axiomatic, and the alleged symmetry that Haji sees between these sorts of principle is hardly obvious (2011, 102).

In fact, I think there is a plausible basis for "ought" implies "can" that simply has no parallel in the case of "ought" implies "able not to". The appeal of "ought" implies "can" principles may in fact *not* rest on any control requirement that involves alternative possibilities. More plausibly, their appeal may be grounded in the simple idea that it is unreasonable to demand the impossible. We may well suppose that it is unreasonable to demand the impossible *without* supposing that this rests on a control requirement that involves alternative possibilities.

Any demand that is impossible to meet will, by an obvious logical entailment, also be a demand with respect to which the agent lacks two-way control. But there is no entailment in the other direction. There is certainly no logical entailment from the plausible idea that it is unreasonable to demand the impossible to the far less plausible claim that it is unreasonable to demand the unavoidable.[8]

If there are cases in which we are plausibly required to do something that we also cannot refrain from doing, then we have good reason to suppose that it is the unreasonableness of the demand to do the impossible that is doing all of the work in rendering principles like OIC plausible, and that two-way control is irrelevant. Of course, we have already examined such a case: the case of morally abhorrent actions that an agent is also psychologically incapable of.

Moreover, think about cases in which it is uncertain whether or not one is physically capable of committing some wrong. For example, I think that it would be morally impermissible for me to leave the house with a kitchen knife and stab to death the first person I see. However, I have absolutely no idea whether I could physically *succeed* in such an endeavour, even supposing I tried my best. It seems absurd to suppose that I should first have to be in a position to know whether I could succeed in order to work out whether stabbing an innocent bystander is morally impermissible (appeal to some theory of normative ethics ought to settle *that* question quite irrespective of my abilities).

There is also a clear a disparity here with respect to duty and prohibition. Plausibly, I can only be morally required to save the drowning child if I am capable of it. If it is uncertain whether I will be physically able to, then we might plausibly say that I have a duty to try, even if I could not have a duty to succeed in my attempt. In contrast, it barely seems coherent to assert that it would be impermissible for me to *try* to stab someone to death while asserting at the same time that it would not be impermissible for me to *actually* stab someone to death. For one thing, I could hardly *succeed* in such an attempt without first *making* the attempt, so if the latter is prohibited, it seems the former must be too. Moreover, it seems that the very reason we are prohibited from attempting certain things is precisely because it would be wrong to actually do those things, so a stand-alone prohibition against attempting would typically make very little sense unless coupled with a prohibition against actually doing what one is attempting to do.

Moreover, there are obvious reasons why we might expect such an asymmetry. In general, having a duty to do something might be thought to depend on our having strong moral reasons to do it. One would expect moral rea-

[8] Granted, the demand may be pragmatically pointless in any situation in which all parties *know* that it will be inevitably met, but this hardly renders it unreasonable.

sons to behave in ways that parallel reasons of any other sort, such as, for instance, epistemic or prudential ones. And reasons of every other sort seem to be asymmetric with respect to our abilities in precisely the way that I claim moral reasons are. Perhaps it cannot be true that an agent ought to believe something if she is incapable of believing it. But it does not seem to follow that she could not have good reason to believe something that she is incapable of doubting (if a belief is indubitable, this is typically thought to be a point in its favour). Or consider prudential reasons. If you are starving hungry (barring any conflicting considerations) you have good reason to eat. If you are incapable of eating, this would undermine those reasons. But it's not at all obvious that if you cannot resist eating, that would in any way weaken the reasons you have in favour of eating.[9]

There are clear grounds for supposing that our reasons are limited to those things that we are able to do, while not being similarly limited to what we are able to avoid. Our reasons are typically based on some sort of independent *value* that's at stake. If a reason for performing some action or believing some proposition is based on some value (e.g. good evidence or a strong moral or prudential case), then insofar as we are capable of sensitivity to that value, we will be sensitive to the reasons it generates. But there would be no point at all in possessing a parallel capacity for *in*sensitivity towards those same values. Here's another way to put the point: if we are violating some core value, we had better have a good excuse for doing so. Being incapable of respecting the value certainly *is* a good excuse. If we are instead respecting the value, we need no excuse for doing so, so no parallel ability to do otherwise is called for in order to render our behaviour intelligible. That something is impossible is, in itself, a *reason* for not bothering. In contrast, the fact that we cannot avoid choosing to do something doesn't undermine the rationale for doing it at all. In some cases, it may well be the very strength of the rationale in favour of performing some action or adopting some belief that *explains* why doing so might be irresistible to us.

[9] This is not entirely uncontroversial. Lockie (2018) argues that prudential and epistemic reasons, as well as moral ones, depend on our ability to avoid doing or believing the thing in question. I am doubtful about OIANT principles in relation to all of these classes of reasons, but I think that Lockie is right in maintaining that there could be little intelligible basis to suppose that moral reasons were unique in this respect, hence if OIANT principles are to be plausible in the moral realm, we should expect them to be defensible in the epistemic and prudential realms too. Though of course, if we accept OIANT principle across the board, including in the epistemic realm, we would then, arguably, need to embrace Lockie's further conclusion: that any argument in favour of determinism would be automatically self-refuting.

Demanding the impossible is unreasonable on the basis that an *inability* to do something may render one's otherwise bad or irrational behaviour perfectly reasonable in the circumstances. This is not dependent on any alternative possibilities requirement for control, as evidenced by the fact that a person's perfectly decent but unavoidable behaviour may well be entirely reasonable and explicable, even if they cannot resist this behaviour, on the basis that it is explained by their sensitivity to certain values. Such an explanation would more plausibly be *weakened* by introducing the additional ability to be *insensitive* to those values as opposed to being *strengthened* by it.

Moreover, whether a demand constitutes a demand for the impossible is asymmetric with respect to what the agent *must* do and what the agent *cannot* do. While it is unreasonably demanding to expect an agent to do the impossible, it is in no way similarly unreasonably demanding to expect an agent to do the inevitable. Since the requirement is so easily met, quite the opposite seems to be true; the inevitability is, if anything, evidence for the conclusion that such a requirement is *un*demanding. But in any case, there is certainly no parallel entailment of demandingness. This is precisely why psychiatrically well-adjusted individuals don't deserve medals for not dismembering small children.

We cannot support OIANT then, by a simple appeal to the alleged symmetry with OIC. Moreover, it is not all all obvious that the insistence on symmetry can be propped up with the consideration that both OIC and OIANT depend on a two-way freedom.

2.3 *The "Theory-Fuelled" Defence*

The "theory-fuelled" defence draws on Feldman's analysis of obligation in terms of the comparative value of the possible worlds accessible to agents (1986). More recently, Haji calls this the "doing the best we can" model (DBWC) (2019; see also Haji and Herbert 2018a).

In short, the analysis contends that we are morally obligated to actualise the best world that we can actualise of all of those "accessible" to us, where "best" is understood in terms of a ranking of the "deontic" or "intrinsic" value of worlds, according to whichever theory of normative ethics is endorsed (e.g. for a utilitarian it may be the world with the greatest sum of utility, for a Kantian it may the world in which we act in accordance with universalisable maxims, whereas for a virtue ethicist it may be the world in which we best act in accordance with the virtues).

doi: 10.48106/dial.v74.i1.03

There needn't be a *unique* best world; perhaps various worlds are tied for first place. But we are obligated to actualise *a* best world. However, some facts may be "unalterable"; there are certain states of affairs that would occur in every possible world accessible to us (e.g. the sun will rise tomorrow, various statements about the past will be true, etc.) If those states of affairs occur in all of the worlds that are accessible to us, then it is trivially true that they will also occur in all of the *best* worlds accessible to us. But now we have a problem: it appears that anything unalterable will automatically be obligatory. We will automatically be obligated to actualise any world that we cannot avoid actualising. Yet this is counterintuitive; it seems intuitively wrong to say that I have a moral duty to actualise a world in which the sun rises tomorrow or to actualise a world in which certain statements about the past are true.

Haji's solution is to appeal to an OIANT principle. That is, we assume that further to supposing that we can only be obligated to bring about states of affairs that are accessible to us, we must *also* suppose that we can only be obligated to bring about any particular state of affairs on the explicit condition we are *also* able to actualise a world in which those states of affairs do *not* obtain.

Perhaps this is one way to maintain a DBWC theory consistent with ensuring that the unalterable should not automatically be obligatory. But it is not the only way, and it's hardly obvious that it is the most plausible way. For instance, instead of endorsing OIANT, we could instead add the (far more compelling) stipulation that we can only be obligated to bring about any outcome insofar as that outcome is causally dependent on our *intentions*.[10]

In fact, Haji's claim that the relevant sort of ability for duty requires that actions not be "strictly out of one's control" commits to precisely this. More recently, Haji and Herbert have defended the claim that the sort of ability relevant to duty ought to be robust, in the sense that requires, among other things, that it is strongly agentive, where this involves being brought about by an agent *intentionally* (2018a, 2018b). However, if having a duty requires that we are able to fulfil that duty in precisely this robust sense, this already rules out having the duty to bring about *some* unalterable states of affairs; it rules out precisely having the obligation to bring about states of affairs that

[10] To be clear, I do not mean to suppose that the outcome must be caused by a *prior* intention. Rather, we should include any outcome that could be brought about through the agent's own deliberate efforts. This means, at least, that the agent's *intention in acting* is causally relevant to the outcome.

will occur independently of our intentions, and hence rules out having such obligations as seeing to it that the sun rises tomorrow.

Moreover, this plausibly explains why it seems intuitively obvious that we *are* obligated to refrain from dismembering small children, even if not refraining from such behaviour is a psychological impossibility, consistent with the fact that it does *not* seem plausible that we are obligated to see to it that the sun rises tomorrow. Since the very point of moral duties is to guide our intentions, we should expect those duties to be limited in scope to those outcomes that are dependent on our intentional behaviour.

Short of having some independent reason to favour a solution that requires us to invoke OIANT over the principle that duties are limited to intention-dependent states of affairs, it seems we ought to favour the latter. While OIANT principles seem inherently problematic, the principle that one cannot be obligated to bring about a state of affairs that will happen independently of one's intentions seems like a basic truism. Given the ready availability of this solution, a state of affairs being unalterable need not make it automatically obligatory (even if we explicitly reject OIANT). Importantly, however, the fact that some state of affairs is unalterable doesn't *rule out* our having an obligation to bring it about either.

Haji and Herbert further note that if we explicitly *presume* that if something is unalterable, then it cannot be obligatory, this would also provide a basis from which to argue in favour of OIANT principles (2018a, 188). But I am arguing precisely that we have no good independent reason to accept such a presumption. The fact that I am not robustly capable of committing certain morally heinous acts may well establish that my avoidance of such acts is unalterable. But the point is precisely that we have no good reason to suppose that this is inconsistent with it being obligatory that I refrain from committing those acts. So while the presumption that unalterability rules out obligatoriness could certainly provide a basis for accepting an OIANT principle (via a fairly obvious entailment), such a presumption is itself no more plausible than the OIANT principles it is invoked to establish and is no less in need of independent justification.

In sum then, it seems that we have no reason to accept OIANT. Recall, however, that OIANT was a crucial component of the argument to the conclusion that determinism entails that nobody ought morally to do anything. Without it, we are entitled only to the weaker claim that, given determinism, no one ought morally to act otherwise than they do. We must now assess whether, from a practical perspective, this weaker conclusion turns out to be just as destructive.[11]

The following section assesses the implications of embracing just the weaker conclusion entailed by determinism and OIC, given a rejection of OIANT. In particular, the aim is to question whether this weaker conclusion *alone* should be regarded as destructive to deontic morality, even if we follow Haji in supposing that no one has a duty to do otherwise.[12]

3 The Lack of Obligation to Act Otherwise

The conclusion that nobody is obligated to act otherwise than they actually do may seem problematic enough. Let us call this claim "Unfulfilled Obligation Scepticism" (UOS):

> UOS. If an agent *S*, as of a time *t*, actualises a world in which state of affairs *p* occurs, this entails that *S* had no moral obligation, as of *t*, to actualise a world in which state of affairs *p* does not occur.

This means that only our *actual* choices and actions could possibly count as obligatory. We may sometimes both have and fulfil moral obligations, but we can never have a moral obligation that we contravene. Perhaps this alone undermines deontic morality. UOS may seem to threaten moral deliberation, obligation, or motivation, rendering them practically unintelligible. Let's examine these potential threats in turn.

3.1 *UOS and Moral Deliberation*

Firstly, it might be argued that UOS renders moral deliberation practically impossible. By "moral deliberation", I mean reasoning about what to do in advance of deciding, rather than reasoning about how to appraise an action that has already occurred.

There are several reasons why UOS might look problematic. We always know in advance that there is no way that our actions will possibly count

[11] For illuminating explorations of arguments to this more modest effect, see Nelkin (2011, 100–103) and Jeppsson (2016).

[12] Since the following section is premised explicitly on assessing the implications of rejecting OIANT and embracing OIC *alone*, any readers who are unpersuaded by the arguments so far, aimed at establishing that we can embrace the latter without the former, can essentially stop reading here.

as "forbidden" at the time that we perform them. Moreover, whether we are obligated to perform any action seems closely dependent on whether we choose to, so we might suppose that UOS robs us of any intelligible way to give rational weight to our purported duties prior to actually making a choice.

Suppose that Ada is a highly rational moral agent, who has recently become convinced of the truth of UOS. She believes that she can only be morally obligated to do something if she does in fact do it. She now faces the following situation: Ada's uncle has arranged in his will for her to receive all of his fortune should he die. However, he is planning to change his will when he visits the solicitor's office later today. Her uncle has two small children and had previously supposed that his wealthy wife's ample income would stand them in good stead should he suddenly die, so he had planned to leave his fortune to Ada, his favourite niece. However, his wife has just died in a freak accident (leaving her fortune to her husband). If he should suddenly die too, his children would now be left orphaned and destitute, while Ada would receive all of his wealth, including that of his late wife. In contrast, Ada has a decent job and a reasonably high income of her own. She will be fine without a substantial inheritance. He is therefore planning to change his will, leaving the bulk of his fortune to his children and a much more modest sum for Ada. She can appreciate the reasonableness of her uncle's decision.

However, while she is alone visiting him, he collapses unconscious, and appears to be dying of a heart attack. No one else knows that Ada is visiting. She could easily walk away without calling an ambulance. She would then be rich enough to buy that Ferrari she always wanted. As a rational moral agent, Ada certainly would have supposed that she had a moral obligation to call an ambulance *prior* to being persuaded of the truth of UOS. But she must now work out what bearing this principle has. Should it change the way that she morally deliberates?

I endorse the idea that we ought to do the best we can, where this involves being obligated to bring about the best of the intention-dependent states of affairs accessible to us. So Ada ought to actualise the best intention-dependent state of affairs she can. This only seems to require two abilities: firstly, she must be able to compare the deontic value of the worlds that would result from various rival intentions, and secondly, she needs to suppose that she can actualise the best of them. We ought to ask whether UOS poses any obstacle to her doing either of these things.

Firstly, let's think about her ability to assess the value of the intention-dependent states of affairs between which she is deliberating. On virtually any theory of normative ethics, the world in which she calls an ambulance will look superior to the world in which she does not call an ambulance. If she doesn't call an ambulance, she will perform no morally admirable actions, and her greed and cruelty will result in an innocent man dying, and his children being left orphaned and destitute. If she does call an ambulance, she will have done a good deed, and through her fairness and kindness, she would ensure that he survives to care for his children. For deontologists, virtue ethicists and consequentialists alike then, the world in which she calls an ambulance will be ranked morally superior to those in which she refrains.

Does she need assurance of her duty in advance? It seems not. On any plausible DBWC analysis, moral duties are not going to be stand-alone considerations that exert their moral pressure on us independently of the other facts about the situation. A world w that we might actualise does not count as morally superior to some other world w' on the *basis* that we are morally obligated to actualise w instead of w' (that supposition would render the DBWC account entirely vacuous). The explanation is always the other way around: we are morally obligated to favour actualising w over w' precisely *because* we have some independent basis to suppose that w is superior to w'. The obligation arises because one of these worlds has a higher "intrinsic value". Values are conceptually prior to obligations: duties are the conceptual outputs of values.

But the point needn't rest on accepting a DBWC analysis either. Quite independently of whether one accepts that analysis, it is a mistake to think that duty is conceptually prior to moral value. Consider Kant. There can be few theorists who afford duty a more fundamental status. Yet even for Kant, duties are not independent additional substantive reasons for acting; they are derived from considerations about the rational wills of other agents, which confer on them a status as ends in themselves. While Kant encourages us to act "from duty", as opposed to merely "in conformity with duty" (1998, 10–11), he certainly doesn't suppose that duties exist and exert pressure independently of the values that give rise to them; respecting duty is simply the same thing as respecting other rational beings. It's hard to imagine any plausible system of ethics according to which duties are not derived from some prior moral value.

Perhaps it will be accepted that Ada (as a rational agent with some theory of normative ethics up her sleeve) knows that the world in which she calls an ambulance for her uncle is better than the world in which she refrains from calling an ambulance (i.e. she knows that there are substantive moral considerations in favour of calling an ambulance). Granted that she knows this, she must also know she is obligated to call an ambulance insofar as she can. But given determinism, we may worry that she has no reason to think that she can.

This concern is misguided. Firstly, we must dispense with any idea that if her intention is determined, then her actions are fixed no matter what she intends. To reason like this would be to commit the "fatalist's fallacy": even if her action is predetermined, this does not entail that it isn't conditional upon her intentions. If she is determined to call an ambulance, this will be *because* it is determined that her deliberative process culminates in her forming an intention to call an ambulance, and this brings it about that she calls an ambulance. Determinism does not make our attempts to act causally ineffective.

Secondly, she has no reason to suspect, in advance of making up her mind, that she cannot call an ambulance. While it is possible that determinism robs her of the ability to call an ambulance, it might just as easily rob her of the ability to refrain. She has no reason to favour the presumption that her calling an ambulance is impossible over the presumption that it is inevitable. The only way that she can find out which of these she is determined to do is by reaching a *decision*.

From an epistemic perspective, both decisions remain open. As Pereboom (2001, 147–48), Fischer (2006), and Jeppsson (2016) have all argued, such epistemic openness is all we need in order for it to be rational to make a value-driven choice. As Fischer puts the point, if one were asked to choose which of two doors to walk through, and told that behind one them is a million dollars while behind the other there is a den of rattlesnakes, it would be ludicrous to suppose that the truth of determinism might weaken the rational case in favour of choosing the door with the money, or that one would be forced to just "wait and see what happens" instead of making a value-driven choice (2006, 329).

Moreover, suppose we grant that determinism introduces a doubt about whether Ada can call an ambulance (we should not grant this, given the deliberatively irrelevant nature of the "doubt", when both options remain epistemically open, but suppose we grant it anyway). Doubts about whether we can do things do not usually weaken our rationale for *trying* when there is something morally significant at stake. Obviously, sometimes failure comes with other off-putting risks; you may be reluctant to dive into the river to save the drowning child, but it is usually the risk to your own life rather than the possibility of failing in your attempt that causes such reluctance. There is always *some* risk of failure, even with the simplest actions, regardless of determinism. One is "always at the mercy of the world", as O'Shaughnessy famously notes (1973, 370). But it would be very strange for anyone to suppose that this should stop us from even attempting to bring about better outcomes.

Suppose that Sofia is in a hospital when the main power supply fails. Luckily, there is a short-term emergency power supply that will keep the electricity going for five minutes, during which time the back-up generator can be activated, saving the lives of hundreds of patients whose life-support machines will otherwise fail. Now suppose that Sofia is the only person with access to the button that activates the back-up generator. There would be something seriously wrong with Sofia if she reasoned as follows: "I only ought to activate the back-up generator if I can. But there is no guarantee that this button works, so I don't know that I can. I therefore see no reason to bother pressing it". Ordinarily, we do not need a guarantee that we can do something before we attempt to do it when there are morally significant outcomes at stake.

There seems to be no reason to suppose that UOS poses any serious obstacle to moral deliberation. Nonetheless, something emerges from this picture that might seem troubling. Essentially, we can escape being duty-bound to do things simply by choosing not to do them. If Ada does not call an ambulance, it will turn out, once her choice has been made, that she has done nothing wrong. Her choosing not to call an ambulance conveniently establishes that she had no moral obligation to call one. Moral obligations become easily escapable.

On the one hand, it may be argued that there is something conceptually amiss about the idea of a moral obligation that could easily be escaped; we might think that inescapability is an essential condition of moral duty. Hence we would still have a serious threat to deontic morality if it turned out that all of our purported "duties" were easily escapable. On the other hand, the worry may be about motivation; perhaps it will be accepted that we could have duties that were easily escapable, but we might wonder why anybody would comply with them.

3.2 *UOS and Moral Obligation*

The problem of easy escapability arises because we seem to have some power over whether we do certain things: even if causal determinism entails that we are unable to do otherwise, it does not entail that our actions are "strictly out of our control"; there is often a reliable causal correlation between our attempts to do things and the success of those attempts. UOS thus seems to give us a further power that might seem unpalatable; the power to escape being duty-bound to do something merely by choosing *not* to do it.[13]

We may well be aware of the fact that in forming the intention to act as we do, we will also be conjuring up proof that we lack any ability to do otherwise, and will therefore be actualising a situation in which we have no duty to do otherwise. This may appear to leave our moral duties precariously at the mercy of our wills. I see two reasons why this implication might look problematic; the first appeals to a Kantian notion of obligation, and the second rests on a broader conceptual concern about the inescapability of duty.

Firstly, philosophers influenced by Kant may suppose that moral duties are necessarily "categorical imperatives". Kant distinguished hypothetical imperatives, which depend on our contingent aims and desires, from categorical ones, which apply to us necessarily regardless of our contingent aims and desires (1998, 25). When one is morally obligated to do something, the obligation is inescapable in the sense that one ought to do it (insofar as one can) regardless of whether one wants to do it.

Kant's claim that moral duties are categorical imperatives is controversial. While this claim is plausibly at the core of any objectivist analysis of metaethics, many philosophers favour subjectivism. If moral duties are grounded in our subjective aims and desires, they will not be "inescapable" in this Kantian sense.[14] But I am inclined to side with Kant here, so I will not pursue this line of argument. I doubt that anything without the character of a categorical imperative could seriously count as a "moral obligation".

[13] In fact, whether such a power will count as making our duties "easily escapable" may depend on one's view of deterministic agency. Some incompatibilists will suppose that even if an agent can escape a duty merely by intending to do so, this doesn't make duties "escapable" in any significant sense because agents lack control over which intentions they form in the first place. For someone who takes this view, the problem of easy escapability doesn't seem to arise at all. But even some incompatibilists will be concerned about the idea that intending not to fulfil a duty suffices to establish that the agent was never subject to a duty in the first place. This may be worrying irrespective of whether we suppose that the intention itself is freely formed.

[14] Contemporary subjectivism has its roots in the work of early modern sentimentalists, such as Hutcheson, Hume and Smith, and finds more recent expression in that of 20th century non-cognitivists, such as Ayer (1936), Stevenson (1937, 1944), Hare (1952), and Gibbard (1990). But even those who advocate gentler forms of mind-dependence of morality, like Williams (1979) will struggle to accept that moral duties could be categorical imperatives. See also Foot (1972) and McDowell (1978).

UOS is, however, perfectly consistent with the claim that moral duties are categorical imperatives. The DBWC notion of moral obligation certainly does not entail that moral duties depend on an agent's subjective aims and desires (with the possible exception of certain duties towards oneself, if there are any). The reason why we are morally obligated to actualise certain possible worlds is because they are the most valuable of the ones that we are able to actualise, according to our favoured theory of normative ethics. And the reason why determinism, given Haji's argument, entails that we are never obligated to actualise alternative worlds is not because we do not *want* to actualise those worlds, but because we *cannot* actualise them.

Ada should call an ambulance if she can. This has nothing to do with whether she wants to call an ambulance, and everything to do with the fact that the world in which she calls an ambulance is more valuable than the world in which she does not. It is not more valuable because her own subjective aims and desires deem it to be (perhaps she prefers the world in which she inherits a fortune and buys a Ferrari). It is more valuable because of the comparatively high worth of her character, her actions, and/or the likely outcome of those actions. More generally, whatever your favoured analysis of obligation, I maintain that it is these sorts of substantive moral considerations that ground Ada's duties, and these need not leave her duties precariously contingent on her subjective aims and desires.

While it may be an essential feature of moral obligations that they are categorical imperatives then, this is not inconsistent with UOS. There is, however, a stronger sense in which it might be claimed that moral duties are necessarily inescapable. We might suppose that there is something wrong with the idea that there could be duties that are opted into; duties that we could have escaped being subject to in the first place. This sense does seem plausibly to be threatened by UOS, but it's doubtful that this really is an essential feature of duty at all.

Promise-making is a prime example of a duty that has to be opted into. We typically suppose that we are duty-bound to keep our promises, even if we could have escaped taking on such a duty in the first place. The important point is that we did *not* escape taking on this duty. Consider another example: it is obligatory to feed one's children as opposed to letting them starve. Nonetheless, many of us are under no such obligation, because we have chosen not to have children. While the same means of contraception were presumably available to many of those who chose to have children, citing this fact would hardly get them off the hook for letting their children starve. Again, the fact that they could, in theory, have escaped the obligation does not usually imply that they cannot have a genuine obligation if they did not *actually* escape it.

There seems to be no sense of inescapability such that it both plausibly qualifies as an essential feature of moral obligation and is plausibly ruled out by UOS.

3.3 *UOS and Moral Motivation*

Perhaps it is not moral obligation that is threatened by UOS, but moral *motivation*. While we may intelligibly have duties that are escapable in the sense specified by UOS, the worry may be that this would threaten any basis that we might have for complying with them.

Return to Ada: suppose we accept that her ability to easily escape being duty-bound does not undermine her duty, so long as she doesn't in fact escape it. We might now worry about what sort of motivational basis Ada could have to incur the duty: by merely not bothering to call an ambulance, she can ensure that she had no obligation to call one in the first place. She only has a duty insofar as she willingly opts into it. Given that she stands to gain so much from not opting into it, we might wonder what incentive she could have for opting in.

We have already noted that duties do not, however, provide extra reasons for action that exert pressure on us independently of the moral considerations that give rise to them (see 3.1). I maintain that a competent moral agent acts out of duty not merely because it *is* her duty, but because she cares about the substantive moral considerations which underpin the duty (in terms of any DBWC analysis, these considerations determine the relative values of the rival intention-dependent worlds that she might choose to actualise). It is only if we accept the dubious assumption that the desire not to contravene a duty is the *sole* basis of moral motivation (and that the desire to fulfil duties is always curiously absent) that UOS seems to seriously undermine moral motivation.

Putting aside the possibility of determinism and UOS, let's think about ordinary cases that parallel the sort of escapability of duty that we are contemplating. Suppose that Aisha believes that she ought to give blood so long as she is eligible to. She also knows that she has a blood donation appointment in one month's time. Now suppose that she is planning to go on holiday before the appointment, and she is trying to decide where to go. She suddenly remembers that if she opts for the destination in sub-Saharan Africa instead of the destination in Europe, this will stop her from being eligible to give blood for at least a year. If it stops her from being eligible to give blood, it will also remove any moral duty that she has to give blood. Should we expect this to motivate her to opt for sub-Saharan Africa instead of Europe? Insofar as Aisha counts as a competent moral agent, I very much doubt that we should expect this. She may even regard it as a reason *not* to opt for the destination in sub-Saharan Africa.

Competent moral agents typically care about their duties because they care about the moral pressures that give rise to them. The reason why Aisha may be willing to incur the duty, even though she has been given an easy way of escaping it is because she cares about people who need blood transfusions. It is because of those people, after all, that she even takes herself to *have* a duty to give blood if she is eligible to; she thinks that the world in which she contributes to the supplies of blood banks is better than the world in which she does not. All those car crash victims and children with leukaemia are not going to just *go away* because she is not personally duty-bound to help them. If she didn't care about these people, she might just as easily contravene the moral duty as escape it.

This brings us to the crux of the issue: the very *same* considerations that count in favour of fulfilling the duty, should you have it, count just as strongly in favour of opting into the duty, if you need to do so in order to fulfil it. And the very *same* considerations that count in favour of opting out of the duty, if you can, count just as strongly in favour of contravening the duty, if you cannot. In no case then, does the fact that the duty can only be fulfilled if opted into (i.e. UOS) change the agent's reasons for deciding either way. Just like Aisha, the reasons that Ada has for fulfilling her duty to call an ambulance for her uncle (should she have such a duty) also count in favour of incurring the duty if she needs to incur it in order to fulfil it. And the same reasons she has to opt out of incurring the duty would count in favour of contravening the duty if its existence did not depend on her opting into it. In no case does it appear rational for her to arrive at a different decision, given UOS, than she would have arrived at without it.

It's unclear why anyone would be keenly motivated not to contravene a duty, while at the same time caring so little about fulfilling one. Such a mindset seems to be directly inconsistent with the sort of sensitivity to value that characterises competent moral deliberation. What exactly is the imagined psychology of an agent who is highly motivated by an aversion to contravening duties while also trying to avoid fulfilling them? Such an agent, despite her thorough commitment to not contravening duties, would be completely indifferent to the moral pressures that actually give rise to duties, as well as being positively *averse* to fulfilling a duty if it's possible to escape it. Even if it were possible for an agent to have this bizarre attitude towards moral pressures, this certainly does not capture the way most of us morally deliberate.

A competent moral agent typically reasons *from* considerations about the respective values of the courses of action between which she is deliberating *to* conclusions about what she ought morally to do. The moral landscape for anyone who reasons in this way seems to be largely untouched by UOS. So it does not appear to pose a serious threat to moral motivation.

4 Conclusion

Haji argues, similarly to Lockie, that there could be no moral obligations at all if determinism were true. In order to establish this conclusion, however, we must invoke both an "ought" implies "can" principle and an "ought" implies "able not to" principle. In section 1, I argued that without OIANT, we could establish only the weaker conclusion that there are no *unfulfilled* moral duties. In section 2, I argued that we ought to reject OIANT, and hence that only the weaker conclusion has been plausibly established. Finally, in section 3, I argued that while this weaker conclusion may initially look just as damaging, it actually has surprisingly little practical importance for morality. While I believe (contra Haji, and in agreement with Lockie) that determinism plausibly threatens moral responsibility, I deny that it poses any serious independent threat to deontic morality.

I admit that aspects of this thesis seem paradoxical. It seems odd to suppose that if determinism is true, this entails that nobody ever violates a moral duty. The air of paradox arises, I think, from two sources. Firstly, from the fact that we do not know in advance what we are capable of doing, since we do not know in advance which actions are impossible and which are inevitable. This means that acting otherwise remains an epistemically and pragmatically live option when we contemplate our potential moral duties in advance. Secondly, it may well be that the sense of "can" typically used in relation to principles like OIC is actually distinct from the sense of "can" according to which determinism robs us of the opportunity to do otherwise.

I have granted for the sake of argument that OIC is true and that we can use a single sense of "can" both in formulating OIC and in defence of the incompatibilist claim that nobody can do otherwise if determinism is true. This has the upshot that nobody can be obligated to act otherwise if determinism is true, and hence that there are no unfulfilled duties in a deterministic world. If that conclusion seems too counterintuitive to accept, then an alternative strategy would be to question whether we should accept all of the following three theses: (1) that OIC is true, (2) that determinism may well be true, and (3) that no one can do otherwise if determinism is true in precisely the same senses of "can" according to which "ought" implies "can". My goal has been to argue that even if we *accept* all three, the threat to morality might not be as all-encompassing as it seems. Whether we should accept all three is another question entirely.[15]

Nadine Elzein
University of Warwick
nadine.elzein@warwick.ac.uk

References

AYER, Alfred Jules. 1936. *Language, Truth and Logic*. 1st ed. London: Victor Gollancz.

———. 1946. "Freedom and Necessity." *Polemic* 5: 36–44.

BEROFSKY, Bernard. 2002. "Ifs, Cans, and Free Will: The Issues." In *The Oxford Handbook of Free Will*, edited by Robert H. Kane, 181–201. Oxford Handbooks. New York: Oxford University Press.

BROUWER, Frederick E. 1969. "A Difficulty with 'Ought Implies Can'." *The Southern Journal of Philosophy* 7 (1): 45–50. doi:10.1111/j.2041-6962.1969.tb02040.x.

BROWN, James M. 1977. "Moral Theory and the Ought-Can Principle." *Mind* 86 (342): 206–23. doi:10.1093/mind/lxxxvi.342.206.

CAMPBELL, Charles Arthur. 1951. "Is 'Free Will' a Pseudoproblem?" *Mind* 60 (240): 441–65. doi:10.1093/mind/LX.240.441.

CHISHOLM, Roderick M. 1964. *Human Freedom and the Self*. University of Kansas: Department of Philosophy.

CLARKE, Randolph. 2009. "Dispositions, Abilities to Act, and Free Will: The New Dispositionalism." *Mind* 118 (470): 323–51. doi:10.1093/mind/fzp034.

ELZEIN, Nadine. 2013. "Pereboom's Frankfurt Case and Derivative Culpability." *Philosophical Studies* 166 (3): 553–73. doi:10.1007/s11098-012-0061-y.

———. 2017. "Frankfurt-Style Counterexamples and the Importance of Alternative Possibilities." *Acta Analytica* 32 (2): 169–91. doi:10.1007/s12136-016-0305-0.

[15] I am very grateful to Yuhan Fu, Ivo Evans Storrie, Dr Anneli Jefferson, and Dr Tuomas K. Pernu, for valuable comments on an earlier draft of this paper. I am also very grateful to Professor Ishtiyaque Haji and to anonymous reviewers for their incredibly helpful feedback.

Elzein, Nadine, and Tuomas K. Pernu. 2019. "To Be Able to, or to Be Able Not to? That Is the Question: A Problem for the Transcendental Argument for Freedom." *European Journal of Analytic Philosophy* 15 (2): 13–32. doi:10.31820/ejap.15.2.1.

Fara, Michael. 2008. "Masked Abilities and Compatibilism." *Mind* 117 (468): 843–65. doi:10.1093/mind/fzn078.

Feldman, Fred. 1986. *Doing the Best We Can: An Essay in Informal Deontic Logic*. Philosophical Studies Series 35. Dordrecht: D. Reidel Publishing Co.

Fischer, John Martin. 2003. "'Ought-Implies-Can,' Causal Determinism and Moral Responsibility." *Analysis* 63 (1): 244–50. doi:10.1093/analys/63.3.244.

———. 2006. "Free Will and Moral Responsibility." In *The Oxford Handbook of Ethical Theory*, edited by David Copp, 321–55. Oxford Handbooks. Oxford: Oxford University Press.

Foot, Phillipa R. 1972. "Morality as a System of Hypothetical Imperatives." *The Philosophical Review* 81 (3): 305–16. doi:10.2307/2184328.

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

Frankfurt, Harry G. 1969. "Alternate Possibilities and Moral Responsibility." *The Journal of Philosophy* 66 (23): 828–39. doi:10.4324/9781315248660-2.

Gibbard, Allan F. 1990. *Wise Choices, Apt Feelings: A Theory of Normative Judgement*. Cambridge, Massachusetts: Harvard University Press.

Graham, Peter A. 2011. "'Ought' and Ability." *The Philosophical Review* 120 (3): 337–82. doi:10.1215/00318108-1263674.

Griffin, James. 1992. "The Human Good and the Ambitions of Consequentialism." *Social Philosophy and Policy* 9 (2): 118–32. doi:10.1017/s0265052500001436.

Grzankowski, Alex. 2014. "'Can' and the Consequence Argument." *Ratio* 27 (2): 173–89. doi:10.1111/rati.12033.

Haji, Ishtiyaque. 1998. *Moral Appraisability: Puzzles, Proposals, and Perplexities*. New York: Oxford University Press.

———. 1999. "Moral Anchors and Control." *Canadian Journal of Philosophy* 29 (2): 175–203. doi:10.1080/00455091.1999.10717510.

———. 2002. *Deontic Morality and Control*. Cambridge: Cambridge University Press.

———. 2019. *The Obligation Dilemma*. Oxford: Oxford University Press.

Haji, Ishtiyaque, and Ryan Herbert. 2018a. "Ability, Frankfurt Examples, and Obligation." *The Journal of Ethics* 22 (2): 163–90. doi:10.1007/s10892-018-9267-3.

———. 2018b. "Indeterministic Choice and Ability." *The Journal of Ethics* 22 (2): 191–203. doi:10.1007/s10892-018-9268-2.

Haji, Ishtiyaque, and Michael McKenna. 2004. "Dialectical Difficulties in the Debate about Freedom and Alternative Possibilities." *The Journal of Philosophy* 101 (6): 299–314.

———. 2006. "Defending Frankfurt's Argument in Deterministic Contexts: A Reply to Palmer." *The Journal of Philosophy* 103 (7): 363–72. doi:10.5840/jphil2006103715.

HARE, Richard M. 1952. *The Language of Morals*. Oxford: Oxford University Press.

HEINTZ, Lawrence L. 2013. "Excuses and 'Ought' Implies 'Can'." *Canadian Journal of Philosophy* 5 (3): 449–62. doi:10.1080/00455091.1975.10716123.

INWAGEN, Peter van. 1983. *An Essay on Free Will*. Oxford: Oxford University Press.

———. 2000. "Free Will Remains a Mystery." In *Philosophical Perspectives 14: Action and Freedom*, edited by James E. Tomberlin, 1–19. Oxford: Basil Blackwell Publishers.

———. 2004. "Freedom to Break the Laws." In *Midwest Studies in Philosophy 28: The American Philosophers*, edited by Peter A. French and Howard K. Wettstein, 334–50. Boston, Massachusetts: Basil Blackwell Publishers.

———. 2008. "How to Think about the Problem of Free Will." *The Journal of Ethics* 12 (3–4): 327–41. doi:10.1007/s10892-008-9038-7.

JEPPSSON, Sofia. 2016. "Reasons, Determinism and the Ability to Do Otherwise." *Ethical Theory and Moral Practice* 19 (5): 1225–40. doi:10.1007/s10677-016-9721-x.

KANE, Robert H. 1999. "Responsibility, Luck, and Chance: Reflections on Free Will and Indeterminism." *The Journal of Philosophy* 96 (5): 217–40. doi:10.2307/2564666.

KANT, Immanuel. 1785. *Grundlegung zur Metaphysik der Sitten*. Riga: Johann Friedrich Hartknoch.

———. 1998. *Groundwork of the Metaphysics of Morals*. Cambridge Texts in the History of Philosophy. Cambridge: Cambridge University Press. Translation ofKant (1785).

KEKES, John. 1984. "'Ought Implies Can' and Two Kinds of Morality." *The Philosophical Quarterly* 34 (137): 459–67. doi:10.2307/2219064.

KRATZER, Angelika. 1977. "What 'Must' and 'Can' Must and Can Mean." *Linguistics and Philosophy* 1 (3): 337–56. doi:10.1007/bf00353453.

KÜHLER, Michael. 2013. "Who Am i to Uphold Unrealizable Normative Claims?" In *Autonomy and the Self*, edited by Michael Kühler and Nadja Jelinek, 191–212. Philosophical Studies Series 119. Dordrecht: Springer Verlag.

LEHRER, Keith. 1968. "Cans Without Ifs." *Analysis* 29 (1): 29–32. doi:10.1093/analys/29.1.29.

LEMMON, Edward John. 1962. "Moral Dilemmas." *The Philosophical Review* 71 (2): 139–58. doi:10.2307/2182983.

LEWIS, David. 1981. "Are We Free to Break the Laws?" *Theoria* 47 (3): 113–21. doi:10.1111/j.1755-2567.1981.tb00473.x.

LOCKIE, Robert. 2018. *Free Will and Epistemology. A Defence of the Transcendental Argument for Freedom*. London: Bloomsbury Academic.

MAIER, John. 2015. "The Agentive Modalities." *Philosophy and Phenomenological Research* 90 (1): 113–34. doi:10.1111/phpr.12038.

MCDOWELL, John Henry. 1978. "Are Moral Requirements Hypothetical Imperatives?" *Proceedings of the Aristotelian Society, Supplementary Volume* 52: 13–29. doi:10.1093/aristoteliansupp/52.1.13.

Mele, Alfred R. 2003. “Agents’ Abilities.” *Noûs* 37 (3): 447–70. doi:10.1111/1468-0068.00446.

Moore, George Edward. 1903. *Principia Ethica*. Cambridge: Cambridge University Press.

———. 1922. “The Nature of Moral Philosophy.” In *Philosophical Studies*, 310–39. London: Routledge & Kegan Paul.

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

O’Shaughnessy, Brian. 1973. “Trying (as the Mental ’Pineal Gland’).” *The Journal of Philosophy* 70 (13): 365–86. doi:10.2307/2024676.

Pereboom, Derk. 2001. *Living Without Free Will*. Cambridge: Cambridge University Press.

Rescher, Nicholas. 1987. *Ethical Idealism. An Inquiry into the Nature and Function of Ideals*. Berkeley, California: University of California Press.

Saka, Paul. 2000. “Ought Does Not Imply Can.” *American Philosophical Quarterly* 37 (2): 93–105.

Sapontzis, Steve F. 1991. “‘Ought’ Does Imply ‘Can’.” *The Southern Journal of Philosophy* 29 (3): 382–93.

Schlick, Moritz, ed. 1930. *Fragen der Ethik*. Schriften zur wissenschaftlichen Weltauffassung 4. Wien: Springer Verlag.

———. 1939. “When Is a Man Responsible?” In *Problems of Ethics*, 143–58. New York: Prentice-Hall, Inc. Authorized translation of Schlick (1930) by David Rynin.

Sinnott-Armstrong, Walter. 1984. “‘Ought’ Conversationally Implies ‘Can’.” *The Philosophical Review* 93 (2): 249–61. doi:10.2307/2184585.

———. 1988. *Moral Dilemmas*. Oxford: Basil Blackwell Publishers.

Smart, Jamieson John Carswell. 1961. “Free-Will, Praise and Blame.” *Mind* 70 (279): 291–306. doi:10.1093/mind/lxx.279.291.

Smith, Michael A. 1997. “A Theory of Freedom and Responsibility.” In *Ethics and Practical Reason*, edited by Garrett Cullity and Berys Gaut, 293–320. Oxford: Oxford University Press.

———. 2003. “Rational Capacities, or: How to Distinguish Recklessness, Weakness, and Compulsion.” In *Weakness of Will and Practical Irrationality*, edited by Sarah Stroud and Christine Tappolet, 17–38. Oxford: Oxford University Press.

Stern, Robert. 2004. “Does ’Ought’ Imply ’Can’? And Did Kant Think It Does?” *Utilitas* 16 (1): 42–61. doi:10.1017/s0953820803001055.

Stevenson, Charles Leslie. 1937. “The Emotive Meaning of Ethical Terms.” *Mind* 46 (181): 14–31. doi:10.1093/mind/xlvi.181.14.

———. 1944. *Ethics and Language*. New Haven, Connecticut: Yale University Press.

Streumer, Bart. 2003. “Does ’Ought’ Conversationally Implicate ’Can’?” *European Journal of Philosophy* 11 (2): 219–28. doi:10.1111/1468-0378.00184.

———. 2007. "Reasons and Impossibility." *Philosophical Studies* 136 (3): 351–84. doi:10.1007/s11098-005-4282-1.

———. 2010. "Reasons, Impossibility and Efficient Steps: Reply to Heuer." *Philosophical Studies* 151 (1): 79–86. doi:10.1007/s11098-009-9422-6.

Trigg, Roger. 1971. "Moral Conflict." *Mind* 80 (317): 41–55. doi:10.1093/mind/lxxx.317.41.

Vihvelin, Kadri. 2000. "Libertarian Compatibilism." In *Philosophical Perspectives 14: Action and Freedom*, edited by James E. Tomberlin, 139–66. Oxford: Basil Blackwell Publishers.

———. 2004. "Free Will Demystified: A Dispositional Account." *Philosophical Topics* 32 (1–2): 427–50. doi:10.5840/philtopics2004321/211.

———. 2011. "How to Think about the Free Will / Determinism Problem." In *Carving Nature at Its Joints. Natural Kinds in Metaphysics and Science*, edited by Joseph Keim Campbell, Michael O'Rourke, and Matthew H. Slater, 313–40. Topics in Contemporary Philosophy 7. Cambridge, Massachusetts: The MIT Press.

———. 2013. *Causes, Laws, and Free Will. Why Determinism Doesn't Matter*. Oxford: Oxford University Press.

Vranas, Peter B. M. 2007. "I Ought, Therefore i Can." *Philosophical Studies* 136 (2): 167–216. doi:10.1007/s11098-007-9071-6.

Weir, Ralph Stefan. 2016. "Relative Modality and the Ability to Do Otherwise." *European Journal of Analytic Philosophy* 12 (1): 47–62.

Williams, Bernard Arthur Owen. 1965. "Ethical Consistency." *Proceedings of the Aristotelian Society, Supplementary Volume* 39: 103–23. doi:10.1093/aristotelian-supp/39.1.103.

———. 1979. "Internal and External Reasons." In *Rational Action. Studies in Philosophy and Social Science*, edited by Ross Harrison, 17–28. Cambridge: Cambridge University Press.

Consciousness, Revelation, and Confusion

Are Constitutive Panpsychists Hoist by their Own Petard?

Luke Roelofs

Critics have charged constitutive panpsychism with inconsistency. Panpsychists reject physicalism for its seeming inability to explain consciousness. In making this argument, they commit themselves to the idea of "revelation": that we know, in some especially direct way, the nature of consciousness. Yet they then attribute properties to our consciousness—like being constituted out of trillions of simpler experiential parts—that conflict with how it seems introspectively. This seems to pose a dilemma: either revelation is false, and physicalism remains intact, or revelation is true, and constitutive panpsychists are hoist by their own petard. But this is too simplistic. Constitutive panpsychists can say that our minds contain innumerable phenomenal states that are "confused" with one another: immediately present to introspection only en masse, not individually. Accepting revelation does not require ignoring the attentional, conceptual, and interpretive limitations of introspection, and these familiar limitations remove the tension between panpsychism and relevation.

What is the relationship between being conscious and knowing about consciousness? In answering this question, constitutive panpsychists face a delicate balancing act: their own case against physicalism requires that being conscious reveals something of the metaphysics of consciousness, but the stronger they make this claim of revelation, the stronger becomes an objection to their own view sometimes called "the revelation problem". In this paper I argue that this balancing act, though delicate, is not impossible: there is a plausible, well-motivated "medium-strength" sort of revelation, strong enough to bring down physicalism but weak enough to leave constitutive panpsychism standing.

In section 1, I lay out the background to the panpsychism-physicalism debate; in section 2, I distinguish six "revelation theses"; in section 3 I analyse the structure and varieties of the revelation problem; and in section 4 and section 5 I outline how to address this problem while retaining as much as possible of the theses discussed in section 2.

1 Are Panpsychists Hoist by their Own Petard?

Panpsychists think all the fundamental physical things are phenomenally conscious, where "fundamental physical things" is a placeholder for whatever fundamental entities feature in the true physical theory (particles, fields, strings, spacetime, etc.). The "constitutive" part of "constitutive panpsychism" describes the relationship between macroexperiences (the experiences of humans and other animals) and the postulated microexperiences of the fundamental physical entities.[1] This relationship should be something like the relationship between the physical features of human bodies (macrophysics) and the physical features of the fundamental entities (microphysics). That relationship (which we might call being constituted, being grounded, or being nothing over and above) generates no "explanatory gap": even when the details currently elude us, it seems clear that macrophysics is fully accounted for by microphysics. When you have the right particles, arranged in the right pattern, exerting the right forces on one another, and the right laws governing them, there is no further problem about how to get hands, chairs, planets, etc.: those "come for free" when the microphysical foundations are there.

The failure of consciousness to fit into this neat picture is the objection to physicalism that motivates most contemporary panpsychists. Whereas the distribution of and relations among subatomic particles seems to explain everything about my body, it leaves unexplained why there is anything it feels like to be me, and why it feels the particular way it does. In particular, even

[1] Some panpsychists would not link "macro" and "micro" (terms conveying size) with "human-like" and "fundamental" in this way. In particular, "cosmopsychists" think that the fundamental physical entity is the cosmos as a whole, which is (obviously) bigger than a human being, not smaller (see Gaudry 2008; Jaskolla and Buck 2012; Shani 2015; Nagasawa and Wager 2017; Goff 2017). Though I am sympathetic to cosmopsychism, I do not believe that it changes the essential contours of the revelation problem, though it requires some re-formulating, as noted in footnotes 11 and 14. For now I will, for convenience, speak as though the fundamental physical entities are very small.

knowing the full story about the particles seems to be compatible with not knowing what the experiences are like (this is the "knowledge argument," cf. Jackson 1982; Nemirow 1990; Ball 2009), and it seems that a world might have been physically identical and yet differed from ours in respect of consciousness (the "conceivability argument," cf. Kripke 1980; Chalmers 2009). There is a vast literature on whether these are good reason to reject physicalism (see, e.g, Chalmers 1996; Dennett 2007; Stoljar 2006; Díaz-León 2011), but here I will assume that they are. What comes next? In particular, is constitutive panpsychism, often offered as an attractive non-physicalist alternative, defensible?

Constitutive panpsychism treats consciousness as a fundamental ingredient of nature, but tries to treat it the same as other fundamental ingredients (mass, charge, spin, force, location, etc.). Just as those other fundamentals are widespread in nature, with human beings as simply one particular arrangement of them, so is consciousness: human experience is not metaphysically special, just a complicated combination of widespread components. Constitutive panpsychism thus retains the monistic spirit of physicalism despite recognising consciousness as fundamental. Importantly, non-constitutive versions of panpsychism, on which human consciousness somehow "emerges from" or is "caused by" microconsciousness but not literally "made up of" it, do not secure this advantage. The macrophysical properties of the brain seem to be wholly constituted by the microphysical properties of its parts, so if its macroscopic consciousness is not similarly constituted by microconsciousness, the hoped-for reconciliation of mind and matter falls apart.

This imposes an explanatory burden: constitutive explanations of human consciousness in terms of microconsciousness have to do better than physicalist explanations. And one major line of criticism has been that they do not: there is just as much difficulty in explaining how many simple minds combine into complex minds as in explaining how mindless things generate minds. This broad objection is often called "the combination problem" (Seager 1995, 280; Chalmers 2017; Roelofs 2019), and has received much discussion from both defenders and critics of panpsychism.

One specific strand of the combination problem is "the revelation problem": macroexperiences do not *seem* introspectively to be built up out of microexperiences. And constitutive panpsychists can't just say: "Well they *are*, sometimes things aren't what they seem." That would license physicalists to likewise say: "Exactly! Consciousness *seems* distinct from purely physical facts, but it's actually not." If being conscious doesn't reveal the true nature of

doi: 10.48106/dial.v74.i1.04

consciousness, the case against physicalism is weakened; if it does, then the truth of constitutive panpsychism should be introspectively obvious, which it is not.

This talk of "seeming" and "obviousness" is not the most precise way of presenting things. Authors articulating the sense that there is a problem here say things like:

> [...] it is hard to see how smooth, structured macroscopic phenomenology could be derived [from microexperiences isomorphic to microphysics]; we might expect some sort of "jagged," unstructured phenomenal collection instead. (Chalmers 1996, 306)

> It is hard to see how [microexperiences] could somehow add up to the phenomenal properties with which we are familiar—properties with the specific, homogeneous character with which we are all acquainted [...]. (Alter and Nagasawa 2012, 90–91)

> [Revelation is] inconsistent [...] with my conscious experience turning out to be, in and of itself, quite different from how it appears to be in introspection: i.e. turning out to be constituted of the experiential being of billions of micro subjects of experience [...]. (Goff 2006, 57; cf. Lee 2019, 290–98)

Similar remarks were made by certain non-reductive mind-brain identity theorists in the last century, writing about a perceived "grain problem":

> [Any experience's] physiological substrate, presumably, is a highly structured, not to say messy, concatenation of changes in electrical potential within billions of neurons in the auditory cortex [...]. How do all these microstructural discontinuities and inhomogeneities come to be *glossed over* [...]? (Lockwood 1993, 274)

> How is it that the occurrence of a smooth, continuous expanse of red in our visual experience can [...] involve particulate, discontinuous affairs such as transfers of or interactions among large numbers of electrons, ions, or the like? (Maxwell 1978, 398)

Indeed, Lewis makes a very similar argument, though he rejects the idea that experience reveals its nature and so presents the argument as a *reductio* of this idea:

> If we know exactly what the qualia of our experiences are, they can have no essential hidden structure - no "grain" - of which we remain ignorant. If we didn't know whether their hidden "grain" ran this way or that, we wouldn't know exactly what they were. [...] if nothing essential about the qualia is hidden, then if they seem simple, they are simple. (Lewis 1995, 142, fn. 14)

Although I think all the above quotations express a similar sort of concern, they do so with different emphasis and framing, and the exact nature of the problem involved is far from clear. In section 3 I try to identify the problems more precisely, and in section 4 and section 5, I resolve them.

2 The Revelation Problem and the Revelation Thesis

Before examining the revelation problem for panpsychism, we need to examine the background idea of a "revelation thesis" connecting consciousness to knowledge of consciousness. There are actually several different ideas under the broad heading of "revelation": I will distinguish a total of six distinct revelation theses, resulting from a two-fold distinction permuted with a three-fold distinction.

The two-fold distinction concerns whether the claim says (a) that the full truth about consciousness will always be manifest (a "reality→appearance" direction of implication), or (b) that what is manifest about consciousness is always true (an "appearance→reality" direction of implication).[2] Claims of the first sort rule out any aspect of consciousness being "hidden" from us, while claims of the second sort rule out any sort of "illusion" about consciousness.

The three-fold distinction is about the topic of a revelation thesis - what kind of reality it connects with what kind of appearance. Putting things for now in reality→appearance terms, we can distinguish the claims:

[2] Byrne and Hilbert (2007, 77), draw this distinction for colour properties: they "treat Revelation as equivalent to the conjunction of two theses [...] SELF-INTIMATION [and] INFALLIBILITY", with the former being reality→appearance and the latter appearance→reality.

doi: 10.48106/dial.v74.i1.04

1. That someone having an experience[3] can know that they are presently having that token experience;
2. That someone having an experience can gain a special kind of understanding of that phenomenal property;
3. That this understanding reveals "the complete nature" of a certain type of experience.

The first thesis is sometimes called "self-presentation" or "luminosity", as distinguished from "revelation" (Stoljar 2006, 223). But in other discussions it is presented as an integral part of a broader idea called "revelation." (e.g. Goff 2017, 109–10). The second thesis is sometimes put in terms of forming concepts, sometimes of special sorts (e.g. Chalmers 2003b; Goff 2017, 109–10) and sometimes just in terms of "understanding" (e.g. Stoljar 2006, 229). The third thesis is sometimes put in terms of knowing a phenomenal property's "essence" or "nature", or knowing all the essential or necessary truths about it.[4] Sometimes the term "revelation" or "revelation thesis" is used specifically for one of these theses, or for the set of them together, or for the conjunction of the second and third. But they are worth distinguishing because, as I will show, they support quite distinct revelation arguments against constitutive panpsychism, which need to be addressed in quite different ways.

Moreover, we can distinguish reality→appearance and appearance→reality directions of each of the three, yielding a total of six revelation theses (RT1–RT6), as follows:

Topic	Reality → Appearance direction	Appearance → Reality direction

[3] Differents authors speak variously of qualia, experiences, types of experience, and types of conscious state: for clarity I will in what follows speak of *phenomenal properties* as the things which phenomenal concepts capture, and whose natures they reveal, and of *experiences* as instantiations of phenomenal properties. To have an experience is to instantiate a phenomenal property, i.e. to be conscious.

[4] Some example formulations: the special understanding of an experience type we gain from undergoing it "reveals the essence of Q [the experience type]: a property of Q such that, necessarily, Q has it and nothing else does" (Lewis 1995, 141–42); "for every essential truth T about E, [the subject] knows, or is in a position to know, T" (Stoljar 2006, 228); "the complete nature of the type to which [the experience] belongs is apparent to the concept user" (Goff 2017, 110). Cf. also colour-revelation theses: "If it is in the nature of the colors that p, then after careful reflection on color experience it seems to be in the nature of the colors that p" (Byrne and Hilbert 2007, 77); "The intrinsic nature of canary yellow is fully revealed" (Johnston 1992, 223). Cf. Lee (2019, 291–93), Liu (2019, 2020).

Instantiation	**Revelation Thesis 1**: If someone instantiates a phenomenal property, it will introspectively seem to them that they are instantiating that property. (Call this the "luminosity" thesis.)	**Revelation Thesis 2**: If it introspectively seems to someone that they are instantiating a phenomenal property, then they really are instantiating that property. (Call this the "no illusions" thesis.)
Understanding	**Revelation Thesis 3:** If someone instantiates a phenomenal property, they will be in a position to form a pure phenomenal concept of it. (Call this the "understanding from experiencing" thesis.)	**Revelation Thesis 4:** If someone is in a position to form a pure phenomenal concept of a phenomenal property, they must be instantiating that property. (Call this the "no understanding without experiencing" thesis.)
Knowledge of nature	**Revelation Thesis 5:** If someone has a pure phenomenal concept, reflection upon it can reveal the whole nature of the corresponding phenomenal property. (Call this the "self-intimation" thesis)	**Revelation Thesis 6:** If someone's reflection upon a pure phenomenal concept presents some feature as pertaining to the nature of the corresponding phenomenal property, that feature really does pertain to the nature of that property. (Call this the "infallibility" thesis)

I think these six theses, though logically independent, form a fairly natural package together, and I will refer to this package (i.e. the conjunction RT1–RT6) as "the revelation approach".[5] This package is particularly important for undergirding modal arguments against physicalism, a role which it is held to have both by its defenders and its critics (e.g. Stoljar 2009, 2013; Damnjanovic 2012; Liu 2019, 2020). Lewis, for instance, attributes RT5 and RT6 to Kripke, as a presupposition of the latter's inference from the conceivability of pain without any associated brain state to their separate possibility

[5] The component theses are often connected by the idea that subjects stand in a certain special relation of "acquaintance" to their experiences (see Chalmers 2003b; Goff 2015): being directly acquainted with our experiences is what lets us know of their occurrence, and understand their properties in a way that fully reveals their nature. Acquaintance is often taken to be one species of a broader category of relations, called "awareness", which likewise enable knowledge of various kinds, but which include more mediated forms of awareness like visual awareness, auditory awareness, etc. I am very happy to accept these claims about acquaintance and awareness, but they will not be distinctively important in the discussion that follows.

(Lewis 1995, 328, fn. 3). Goff (2017, 74–76, 96–106) likewise argues that the conceivability and knowledge arguments require that phenomenal concepts be "transparent", effectively meaning that RT5 and RT6 must be true.[6] And Chalmers' version of the conceivability and knowledge arguments relies on the premise that the primary and secondary intensions of phenomenal concepts are equivalent (Chalmers 2003a, 2009), which implies RT5 and RT6.[7]

Although RT5 and RT6 have the clearest role, the falsity of the other revelation theses would also leave the anti-physicalist arguments on a shaky footing. For instance, if RT3 were false, we could worry whether we possessed the pure phenomenal concepts whose "transparency" drove the arguments; if RT2 were false, we could worry that the properties these concepts expressed were not even instantiated (as argued by, e.g. Pereboom 2016, 2019); and RT4 is essential to the knowledge argument, which relies on the premise that someone who has never experienced colour cannot know what seeing colour is like.[8]

3 What is the Revelation Problem, Exactly?

So what exactly is the supposed problem for panpsychists? How is it distinct from other aspects of the combination problem? Fundamentally, it concerns a perceived incompatibility between three things:

- the way human consciousness appears in introspection;

[6] The arguments might not require going all the way to RT5 and RT6. Stoljar (2006, 229–30) suggests that all that is strictly required is that we have a form of access to the natures of phenomenal properties that allows us to know at least something, if not everything, about these natures. Goff argues against such an intermediate position, saying that for any property whose nature we grasp only part of, we can "split" the property into two components, one with an unknown nature and one with a known nature. The arguments against physicalism can then be run just with respect to "that aspect of phenomenal properties whose nature we know", and for that sub-property RT5 and RT6 will be true. In this paper I will suppose that Goff is right, and seek to defend RT5 and RT6 in their "whole nature" form.

[7] A concept's primary intension is available to reflection, while its secondary intension is the nature of the property that concept expresses, so the coincidence of these two intensions implies that the natures of the properties expressed by pure phenomenal concepts are available to reflection by those who possess the concepts.

[8] The revelation approach also comes up in other places. RT1, the "luminosity" thesis, is sometimes appealed to as a distinguishing feature of consciousness (Rosenthal 1993, 359; Kriegel 2009; Strawson 2015, 9). Other philosophers draw on RT1 and RT2 to develop an epistemology of introspection (Chalmers 1996, 218–19; 2003b; Smithies 2019).

- the way human consciousness would be, if constitutive panpsychism were true;
- revelation: the idea that introspection gives special insight into the reality of consciousness.

The third element makes any discrepancy between the first and second seem fatal. Yet that third element is also something panpsychists cannot readily give up.

How should we spell out these core elements? I think there are actually three slightly different arguments to be made here, and then a fourth argument which engages with the debate on a different combination problem, the "palette problem". Let us consider the pure revelation arguments first, which differ primarily in whether they rely on the appearance→reality or reality→appearance direction of implication: the first argument says, "Consciousness appears to be X, but panpsychism implies it is not really X," while the second and third say, "Consciousness fails to appear to be X, but panpsychism implies it really is X." The first focuses on some positive introspective appearance, and accuses constitutive panpsychists of treating that appearance as an "illusion". The others focus simply on the *absence* of a certain appearance.

We can call the first argument the "no illusions" argument, since its third premise is RT2, the "no illusions" thesis:

1. If constitutive panpsychism is true, then human consciousness is always "particulate".
2. Human consciousness (often) appears introspectively to be "smooth".
3. Consciousness can't appear a way that it's not. (RT2)
4. Being "smooth" and being "particulate" are incompatible.
5. Human consciousness is (often) smooth. (from 2 and 3)
6. Human consciousness is (often) not particulate. (from 4 and 5)
7. Constitutive panpsychism is false. (from 1 and 6)

Obviously much turns on the meaning of the terms "particulate" and "smooth", but despite the frequency with which they (and similar terms like "continuous" and "fragmented") appear in statements of the problem, it is unclear how to define them, and consequently unclear how plausible premises 1, 2, and 4 are. This definitional question will be central to my discussion in the next section.

The second and third arguments (involving a "reality→appearance" implication) are both suggested in Chalmers' formulation of what he calls "the revelation argument" (2017, 190). Chalmers notes that although constitutive panpsychism holds consciousness to be “constituted by a vast array of microexperiences”, this vast array is not revealed to us in introspection. This poses a problem if we think both that introspection reveals the nature of consciousness, and that “whatever constitutes consciousness is part of its nature”.

I distinguish two arguments here because I think talk of “introspection” upon “consciousness” can be taken in two quite different ways. One is that introspection focused *on macroexperiences* doesn’t reveal *that* they are constituted by microexperiences. The other is that introspection focused *on microexperiences* isn’t even possible. The former appears to violate what I above called RT5, the “self-intimation” thesis: reflection upon a pure phenomenal concept reveals the whole nature of a phenomenal property. The latter appears to violate both what I above called RT3, the “understanding-from-experience” thesis, and RT1, the “self-presentation” thesis: having an experience should allow knowledge of its occurrence and a pure phenomenal concept of it.

Focusing on either macroexperiences or microexperiences yields the following two arguments, which I will call the “macroexperience-focused” and “microexperience-focused” argument. The first runs thus, with RT5 as third premise:

1. If constitutive panpsychism is true, each human experience (“macroexperience”) is constituted by a vast array of microexperiences.
2. A vast array of microexperiences is not revealed by reflection on macrophenomenal concepts (i.e. phenomenal concepts based on macroexperiences).
3. The nature of a phenomenal property is revealed by reflection on phenomenal concepts based on experiences of it. (RT5)
4. Whatever constitutes something is part of its nature.
5. The natures of macroexperiences do not involve vast arrays of microexperiences. (from 2 and 3)
6. Macroexperiences are not constituted by vast arrays of microexperiences. (from 4 and 5)
7. Constitutive panpsychism is false. (from 1 and 6)

Clearly, the soundness of this argument depends crucially on what is meant by talk of a property’s “nature”, since that will affect the meaning of premises 3 and 4; this question will be at the heart of my discussion in the next section.

The third (“microexperience-focused”) revelation argument runs thus, with a conjunction of RT1 and RT3 as its third premise:

1. If constitutive panpsychism is true, consciousness is constituted by a vast array of microexperiences.
2. We cannot know introspectively about microexperiences, nor form microphenomenal concepts (i.e. phenomenal concepts based on microexperiences).
3. If a subject is having an experience, they can know introspectively that they are, and form phenomenal concepts based on it. (RT1 and 3)
4. If experiences constitute a subject’s consciousness, that subject undergoes them.
5. We are not undergoing a vast array of microexperiences. (from 2 and 3)
6. Human consciousness is not constituted by a vast array of microexperiences. (from 4 and 5)
7. Constitutive panpsychism is false. (from 1 and 6)

Finally, there is an interaction between a revelation thesis, specifically RT5, and another aspect of the combination problem, the “palette problem”. How do the huge range of phenomenal qualities that humans experience arise from a fundamental base which appears to involve only a quite small number of fundamental properties? One solution is the “small palette hypothesis”: there are only a few basic phenomenal qualities, corresponding to the fundamental physical properties, which are somehow “blended” to generate a plethora of different qualities for different macroscopic creatures (see Roelofs 2014; Coleman 2015, 2017; Chalmers 2017, 204–6), whose pattern of similarities and differences are explained by their differing proportions of the basic ingredients. Some critics of the small palette hypothesis object that some of our phenomenal qualities are too heterogeneous to be blended out of a small set of common elements, because they are *completely* dissimilar, with nothing phenomenally in common. Goff (2017, 195), for instance, claims that, “Minty phenomenology and red phenomenology have nothing in common” (cf. a similar argument in McGinn 2006, 96). This line of criticism relies on RT5 to rule out these qualities being similar in a way that we cannot recognise (Goff 2017, 195–97). Call this the “small-palette revelation argument”, the full structure of which is very similar to that of the macroexperience-focused revelation argument:

doi: 10.48106/dial.v74.i1.04

1. If the small palette hypothesis is true, then any two phenomenal qualities experienced by humans have something phenomenal in common.
2. Reflection on some pairs of human experiences (e.g. red and minty) does not reveal them to have anything phenomenal in common.
3. The nature of a phenomenal quality is revealed by reflection on phenomenal concepts based on experiences of it. (RT5)
4. The natures of two things determine whether they have anything phenomenal in common.
5. If a pair of phenomenal qualities has something phenomenal in common, reflection on phenomenal concepts based on experiences of them will reveal this. (from 3 and 4)
6. Some pairs of human experiences have nothing phenomenal in common. (from 2 and 5)
7. The small palette hypothesis is false. (from 1 and 6)

All four arguments have a similar four-premise form: first, a supposed implication of constitutive panpsychism (or small-palette forms of it); second, an introspective datum; third, an epistemological thesis about introspection; and fourth, a metaphysical claim, given which the other three premises entail the falsity of constitutive panpsychism (or small-palette forms of it). But despite their common form, I will argue that the arguments go wrong in quite different ways.

4 Ways of Responding to the Revelation Arguments

The challenge for constitutive panpsychists is to rebut the above four arguments without abandoning the revelation approach, components of which underpin all of them. I will show how to rebut each argument in turn, while keeping the relevant revelation theses as strong as I can.

4.1 *The No-Illusions Revelation Argument*

Consider first the "no illusions" argument, which had the following four premises:

1. If constitutive panpsychism is true, then human consciousness is always "particulate".
2. Human consciousness (often) appears introspectively to be "smooth".

3. Consciousness can't appear a way that it's not.
4. Being "smooth" and being "particulate" are incompatible.

One option for constitutive panpsychists is to deny premise 1, based on defining "particulate" in such a way that a field-based ontology, or a substance-monist ontology, or some other account of physical reality, renders it false that the material world, and any consciousness isomorphic to it, is particulate (see in particular Nagasawa and Wager 2017, 120–21). If the other three premises (and constitutive panpsychism) are accepted, this implies that the kind of consciousness we enjoy is incompatible with some physical theories (those which make matter "particulate") and that we know introspectively that our world is not any of those ways.

However, I think this approach is a mistake. Even if particles are not ultimately real, Lockwood's point still holds: even the simplest experience involves billions of neurones, ions, and neurotransmitters. Even if the space containing two sodium ions is ultimately just a set of derivative aspects of the one substance, there is still a striking difference in the electrical properties of different regions of that space. To dismiss the problem because particles are not in the fundamental ontology would be too easy. Consequently, I suggest the following definition of "particulate":

> X is *particulate* iff X comprises a very large but finite number of parts which differ significantly (in some properties) and discontinuously (on some dimension).

This definition makes the physical brain particulate whatever the fundamental physics turns out to be. Of course this definition will only be as precise as "very large" and "differ significantly and discontinuously". The vagueness of such terms does not stop us from taking "a trillion or more" as a clear case of "very large", and "the mass and charges differences between a water molecule, a potassium ion, and a region of empty space between them" as a clear case of "differ significantly and discontinuously".[9]

[9] Note also that the definition requires only that the properties of the parts vary discontinuously in *some* dimension, i.e. on some natural way of ordering them, not on all: intuitively, the salient facts about brain parts like potassium ions are things like the abrupt drop in mass from inside the ion's nucleus to outside it, but this abrupt drop might vanish if we instead consider all parts of the brain in a list ordered by mass. But if we want to define "particulate" in a way that does justice to the no-illusions argument, the possibility of finding some dimension on which all variation is continuous should not disqualify the brain from being particulate.

doi: 10.48106/dial.v74.i1.04

That leaves three remaining options: deny premise 2 (i.e. contradict the supposed introspective observation), deny premise 3 (i.e. reject this particular revelation thesis), or deny premise 4 (i.e. deny that smoothness and particulateness are incompatible). But everything depends on what "smooth" means. What is the feature of experience that is being reported by those who feel the pull of this argument?

One option is to define "smooth" by ostension: consider some experiences without discernible internal structure, what Lockwood (1993, 274) calls a "phenomenally flawless" experience, and stipulate that "smooth" means the noteworthy feature of those experiences. That would ensure the truth of premise 2, but would make it hard to adjudicate the truth of premise 4. My preference is to define "smooth" in such a way as to ensure the truth of premise 4, e.g:

> X is *smooth* iff it is not particulate.

There are then a few different ways for something to be smooth: since being particular requires parts, for instance, simple things would count as smooth by default. Alternatively, something might be smooth if its parts do not differ significantly in any respect, or do not differ discontinuously along any dimension. The panpsychist must then deny either premise 2 or premise 3: either say that experience does not appear smooth, or say that it does but isn't.[10] At first glance, both options look difficult: premise 3 is, after all, part of the Revelation Approach (RT2), and if premise 2 is false, why did anyone ever advance the argument in the first place?

The way out lies in scrutinising the word "appears", and drawing a distinction between illusions, strictly so-called, and easy misinterpretations. Consider some non-mental examples: at first an act appears noble, an argument compelling, a speech beautiful, and yet then I find that upon giving the matter more thought, this appearance vanishes, and I come to think I was mistaken. The act now appears fanatical, the argument sophistical, the speech saccharine; I think myself foolish for being gullible enough for the act, argument, or speech to ever appear otherwise to me. I might say I was subject to an "illusion", but all this mean is that the act, argument, and speech were such that they could be very readily misjudged.

[10] Using the ostensive definition would just translate denial of premise 2 into denial of premise 4: either way, the claim is that there is no property incompatible with particulateness that consciousness introspectively seems to have.

Contrast this with a white object seen under pure red light, or a straight stick seen half in water, or an ambitious Scottish nobleman hallucinating a dagger. The object appears red but isn’t, the stick appears bent but isn’t, and there appears to be a dagger, but there isn’t. Here no reflection on the appearances will change them, and the subject cannot hold themselves rationally accountable for being subject to them (perhaps for forming beliefs based on them, but not for the appearances themselves). Here we have a stronger sense of “illusion”: it is not that these perceptions are easy to misjudge, it is that their very content is false. Call this the “quasi-perceptual” sense of “appears”, contrasting with the “ready-interpretation” sense (cf. Stoljar 2013; Kammerer 2018).

Premise 3 (RT2) is most plausible if read with the “quasi-perceptual” sense of “appears”. Plausibly it makes no sense to think that my impression of my own experience is an “illusion” in this stronger sense: surely it would be the “impression” that deserves to be called my experience, since this is what I am immediately aware of. To think that consciousness might appear falsely in this way seems to involve forgetting that consciousness *is* how things appear to me (cf. Liu 2020). Or at least, this thought has some appeal, and panpsychists need not disagree with it.

But premise 3 is less plausible if understood in terms of the “ready-interpretation” sense of “appears”, saying that if consciousness is readily interpreted as having some property, it must actually have that property. After all, which interpretations come readily depends on the subject’s expectations, background assumptions, interpretive style, etc. An absolute principle, that no false interpretation could come readily to *anyone*, would be very close to saying, implausibly, that consciousness was never misinterpreted.

So we should read premise 3 as saying that consciousness cannot appear a way it’s not, in the quasi-perceptual sense of “appear”. For the argument to remain valid, premise 2 must also be read in terms of the quasi-perceptual sense of “appear”, not the “ready-interpretation” sense. But now premise 2 is much more deniable. We can deny premise 2, in this strong sense, by taking the appearance of smoothness to be a matter of what interpretations come readily, and not of how things quasi-perceptually appear.

This is my preferred response to the “no illusions” argument: our consciousness really is particulate, not smooth, but it is readily misinterpreted as smooth. But this misinterpretation demands an explanation - what is it about the way consciousness *does* appear, which makes us judge it “smooth”?

One answer appeals to the difference between represented structure and structured representations: that is, experience represents things as being smooth, rather than itself being smooth (versions of this proposal appear in: Clark 1989; Stoljar 2001). Critics have worried that experience itself really does seem to display the relevant sort of smoothness (e.g. Alter and Nagasawa 2012, 91), and that representing a smooth expanse may be insufficient for introspectively seeming, even in the weak sense, to be smooth (consider the sentence "space is infinitely divisible"). Another answer is to say that many experiences quasi-perceptually appear to have, and thus (by RT2) actually have, some property similar to, but not identical to, "smoothness". In section 5 I flesh out this approach.

4.2 *The Macroexperience-Focused Revelation Argument*

Next, consider the macroexperience-focused argument, whose premises are:

1. If constitutive panpsychism is true, each human experience ("macroexperience") is constituted by a vast array of microexperiences.
2. A vast array of microexperiences is not revealed by reflection on macrophenomenal concepts (i.e. phenomenal concepts based on macroexperiences).
3. The nature of a phenomenal property is revealed by reflection on phenomenal concepts based on experiences of it.
4. Whatever constitutes something is part of its nature.

I see little prospect for denying premises 1 and 2,[11] and premise 3 is one of the revelation theses I want to preserve. Chalmers, when he lays out the argument of which this is a variant, advises panpsychists to attack premise 4: to drive a wedge between something's nature and what constitutes it. I agree that this is the right tack, but everything turns on what kind of "nature" is in question, which in turn depends on how we read premise 3, the self-intimation thesis. I think there is a plausible and well-motivated sense of "knowing a nature"

[11] It might look like cosmopsychists can wriggle out of premise 1. But this is illusory: the only way cosmopsychists can deny premise 1 is to commit to an analogous premise that supports a *harder* revelation argument. If they deny that the brain is constituted by neurones, ions, etc., they must instead accept a replacement premise 1*: "If constitutive panpsychism is true, each human experience ('macroexperience') constitutes a vast array of microexperiences." We then run the same argument, with premise 4 replaced by 4*: "Whatever something constitutes is part of its nature." And I think premise 4* is noticeably *more* plausible than premise 4.

which explains why premise 4 is false, without undermining anti-physicalist arguments.[12]

First consider this common gloss: knowing the nature of a property means being in a position to know a priori every necessary truth about that property.[13] If I know the nature of squareness, I am in a position to know a priori every necessary truth about squareness (like what squares' internal angles sum to, or what kinds of triangles they can be divided into), though not to know contingent truths about it (like whether it is my sister's favourite shape). Likewise if I know the nature of being water, I can know every necessary truth about being water (like that water is a chemical compound, or its molecular mass), though not every contingent truth about it (like whether it is instantiated on Earth). This suggests that we know the natures of mathematical properties, but do not automatically know the natures of chemical properties, though perhaps we do now, given scientific progress. And those results seem plausible.

But this gloss is inadequate. Consider someone who knew the nature of squareness but not the nature of triangularity (if that were possible). They would not be in a position to know a priori that every square can be divided into four right-angled triangles. This suggests a refinement: knowing the nature of some property means being in a position to know *a priori* all the necessary truths about that property which involve only other properties whose natures you also know. To put it another way, to know a priori a necessary truth involving two properties, you need to know the natures of both: just knowing the nature of one is not enough.[14] This implies, in particular, that knowing the nature of a constituted property is not sufficient to know about its constitution relationships to other properties, without also knowing the natures of those other properties.

[12] The argument discussed in Lee (2019) combines premises 3 and 4 into a single claim, "Structure Luminosity: If a subject introspects an experience, then that subject is in a position to know the phenomenal realizers of that experience" (2019, 292). Lee argues (in my view plausibly) that this is false, but does not clearly identify which elements of it remain true, and whether they are enough for anti-physicalist arguments.

[13] I am abstracting away from difficulties of memory, attention, and general cognitive skills: in practice, many necessary truths might be just too complicated or subtle for a human mind to entertain, but that should not stop us from saying that someone is in a position to know them if all they would need to do so is an enhancement of their general cognitive skills.

[14] This is not a retreat from the idea that the phenomenal property's "whole nature" is revealed. There is no part of its nature that is hidden: there are only hidden connections between its nature and other natures, and those connections are hidden for the simple reason that those other natures are hidden.

I think this provides a plausible reading of "knowing a property's nature", and thereby of RT5, which does precisely what constitutive panpsychists need it to do: substantiate their arguments against physicalism, without substantiating the revelation argument against their own view. For on this reading of "knowing a nature", that we know the natures of macrophenomenal properties implies that for any other set of properties whose natures we know, we are in a position to tell a priori whether those properties are sufficient to constitute macrophenomenal properties. And the case against physicalism is that physical properties do not seem a priori to constitute macrophenomenal properties. Of course, this attack only works if we know the natures of physical properties (e.g. if we think of them as exhausted by what physics says about them, as what Stoljar (2001) calls the "t-physical" properties, and what Strawson (2006) calls "physicsal" properties). It will not work if we think of physical properties as whatever properties physical things have which in fact account for their satisfying the descriptions given by physics (what Stoljar (2001) calls the "o-physical" properties). But that way out is no use to standard physicalism, which needs physical properties to be well-understood: to say that the reason the conceivability argument fails is that there is some mysterious hidden nature of the physical, which plays some crucial role in accounting for consciousness, is to embrace the kind of "non-standard physicalism" (cf. Stoljar 2006) that is no longer incompatible with panpsychism.

But why doesn't knowing the natures of macrophenomenal properties substantiate a parallel argument against constitutive panpsychism? Because panpsychists do not claim that we know the natures of microphenomenal properties, because we are not the microsubjects who instantiate those properties (though see the next subsection for some complications of this claim). Without knowledge of the candidate constituting properties, we cannot determine a priori their suitability to constitute macrophenomenal properties. All the constitutive panpsychist is committed to is a conditional claim: *if* we were able to grasp the natures of microphenomenal properties, then we could, in principle, see a priori that, when properly arranged, they constitute macrophenomenal properties.

4.3 *The Microexperience-Focused Revelation Argument*

Thirdly, consider the microexperience-focused revelation argument: why can't we introspect microexperiences like we can macroexperiences? The premises of this argument are:

1. If constitutive panpsychism is true, consciousness is constituted by a vast array of microexperiences.
2. We cannot know introspectively about microexperiences, nor form microphenomenal concepts.
3. If a subject is having an experience, they can know introspectively that they are, and form phenomenal concepts based on it.
4. If experiences constitute a subject's consciousness, that subject undergoes them.

Again, I see little hope in denying premises 1 or 2,[15] which leaves three options: deny premise 3 ("we *are* undergoing microexperiences, but cannot introspect them"), deny premise 4 ("microexperiences constitute our consciousness, but we do not undergo them"), or show the argument to be invalid.

Goff's approach in his (2017, 189ff.) is to deny premise 4, to "loosen" the relation between microexperiences and macroexperiences, so that although microexperiences in some sense constitute (or "ground", "compose", or "form") macroexperiences, the phenomenal character of the latter contains nothing of the former. The cost of this is that the constitution relation between microexperiences and macroexperiences is thereby made more mysterious. If this relation were one in which both constituted and constituter were undergone by the same subject, it could be akin to familiar relations among macroexperiences. For instance, the relation between my total phenomenal field right now and the component experiences that it subsumes (sounds I'm hearing, colours I'm seeing, twinges of physical discomfort, etc.) is plausibly something like constitution. It would be nice if panpsychists could assimilate the microexperience-macroexperience relation to familiar relations like this, where a single subject undergoes all the experiences involved; without that link it is hard to see why microexperiences should really be said to "constitute" a macroexperience, as opposed to somehow giving rise to it as a distinct product.

I think the best approach is to say the argument is invalid *when premise 3 is qualified* in certain ways that are independently necessary to make it plausible. An unqualified form of premise 3 faces easy counterexamples: ferrets

[15] Again, though one might think cosmopsychists can deny premise 1, there is no advantage to be gained thereby: the replacement premise 1* - "If constitutive panpsychism is true, human consciousness constitutes a vast array of microexperiences" - will support a revised version of the argument, when paired with 4* - "If experiences are constituted by a subject's consciousness, that subject undergoes them." And again, 4* seems to me even more plausible than 4.

undergo many experiences, but cannot form phenomenal concepts, or know that they are having experiences. But plausibly this is not a counter-example to what premise 3 was intended to say! The problem is not that ferrets' experiences are somehow hidden from them, but just that they lack the conceptual competence to recognise their experiences as such. A qualified version of premise 3 would allow for this: it would say that certain kinds of knowledge and concept-formation are possible whenever a subject undergoes an experience *and* meets various other conditions. Another plausible requirement is attention: one must focus on an experience in order to introspect it, and if one is unable to direct one's attention, introspection will be impossible.[16]

So let us consider a qualified reading of premise 3, that includes these conditions: introspective knowledge is possible whenever a subject undergoes an experience, *and* is capable of conceptualising it, *and* focuses their attention on it. The argument has now become invalid: line 5 ("we are not undergoing a vast array of microexperiences") no longer follows from 2 and 3. There are two reasons why we might be phenomenally undergoing microexperiences but be unable to know them introspectively, compatibly with this weaker reading of premise 3: if humans cannot conceive of experiences as such, or if they are unable to attend to microexperiences. While the first of these options is clearly false, the second is, I think, the best option for the constitutive panpsychist in rebutting the microexperience-focused argument.

This implies that while microexperiences are phenomenally conscious for us, they are not access-conscious for us. That is, microexperiences are presented to us, "right there", characterising the phenomenal character of our consciousness, but they are not presented in such a way that we can cognitively select, access, and identify them. Our relationship to them is rather like our relationship to elements of our experience that are very faint, which require a lot of effort to focus on and distinguish from their surroundings, and which it is correspondingly easier to distract us from. If something in my peripheral vision is roughly the same colour as its surroundings, it would be hard for me to notice it, and if I were distracted, exhausted, or inebriated I might find attending to it all but impossible. Yet it is still part of my phenomenology, not somehow hidden from me. The constitutive panpsychist, I am suggesting, should claim that this near-impossibility of attending to peripheral vision while distracted is intensified to a real practical impossibility

[16] Goff's statement of revelation (2017, 109–10) mentions attention explicitly, and Chalmers appeals to inattention as a primary reason for thinking that his principles of "detectability" and "reliability" can only hold for the most part, not absolutely (Chalmers 1995, 326; 1996, 218–19).

with microexperiences. In section 5 I situate this impossibility claim within a broader picture of how the mind is constituted by microexperiences, which will help to motivate this response to the microexperience-focused argument.

4.4 *The Small-Palette Revelation Argument*

Finally, consider the small-palette revelation argument, whose premises are:

1. If the small palette hypothesis is true, then any two phenomenal qualities experienced by humans have something phenomenal in common.
2. Reflection on some pairs of human experiences does not reveal them to have anything phenomenal in common.
3. The nature of a phenomenal quality is revealed by reflection on phenomenal concepts based on experiences of it. (RT5)
4. The natures of two things determine whether they have anything phenomenal in common.

Since this is not an argument against constitutive panpsychism per se, there are technically five options for constitutive panpsychists in responding to it: deny one of the premises, or accept the conclusion. Accepting the conclusion would mean accepting a "large palette" version of constitutive panpsychism, with all human and animal qualities present in the base even though that is more than there are distinct physical roles to play (see, e.g. Lewtas 2013). The downside is that this sacrifices the appealing parsimony, and isomorphism with physics, that had seemed to set constitutive panpsychism apart from traditional sorts of dualism. Denying premise 3 is also unattractive, since it undermines the case for panpsychism over physicalism.

Denying premise 4 here (as Lee does, 2019, 300–301)is harder than denying premise 4 of the macroexperience-focused argument, that "what constitutes something is part of its nature". I denied the latter because knowing a property's nature is not enough to know necessary truths about it which involve the nature of another property; we would have to know that other property's nature as well. But when it comes to comparing two qualities that we do experience distinctly, it seems to follow that we should be able, in principle, to discern every necessary truth about how those qualities relate, and that should include their resemblance or common constituents.[17]

[17] Could we find a more carefully qualified version of RT5, on which knowing the natures of two properties enables us to know whether one suffices to constitute the other, but not whether and

doi: 10.48106/dial.v74.i1.04

We might deny premise 4 in the same way we might deny premise 4 of the microexperience-focused argument, by saying that although the basic qualities constitute the macroqualities, they do not characterise them - the "blending" leaves no trace of the ingredients at all. But this has the same downsides discussed in the last subsection: if microqualities in no way characterise the macroqualities, the form of constitution involved seems mysterious.

That leaves denying premise 1 or premise 2. Premise 1 might seem undeniable, due to the "interchangeability" of different neurons: experiences of redness and of mintiness involve neurones made of all the same sorts of subatomic particles, so how can one contain any ingredient missing from the other? Any ingredient of the redness experience comes from electrons, quarks, photons, etc., and those same things are all present in the physical basis of a mintiness experience, so how could they not show up in the latter? But this falsely assumes that each macroexperience should contain every ingredient present in its neural basis, as though each one were the independent product of one discrete subset of neurones. It might instead be that several macroexperiences are all grounded in the activity of the same neurones, being just different aspects of the complex, differentiated experience produced by those neurones.

Consider a bar magnet, whose macroscopic behaviour displays a "north pole" and "south pole". The north pole does not arise from one half of the magnet, and the south pole from the other half: both macroscopic features arise from very same microscopic physical things, because those things are themselves internally differentiated and their different aspects add up to what looks, from a macroscopic perspective, like two different things. It would be a mistake to say "since all the particles generating the magnet's north pole also have south poles, why don't their south poles show up in the magnet's north pole?" Perhaps mintiness and redness are likewise different aspects of the same complex experience, itself arising from the combination of a great many internally differentiated microexperiences, combining in different ways depending on such things as firing rates and degrees of neural synchrony. Then they might have nothing phenomenal in common, despite being constituted by the same things.

However, there are limitations to this response. It might allow for a few completely dissimilar pairs to be compatible with the SPH, but not that many -

how they resemble each other? Maybe, but this feels ad hoc to me; I see no plausible way to motivate it.

if there are a hundred completely dissimilar qualities experienced by humans, saying that they arise from the way that internally differentiated aspects of microexperiences are combined starts to load microexperiences with too much structure for us to retain the SPH. To keep the palette small, there shouldn't be too many completely dissimilar pairs of qualities, which is why this response to the argument works best when combined with another: denying premise 2.

Denying premise 2 means denying that redness and mintiness have absolutely nothing at all in common. After all, our ability to recognise two things as akin to one another is usually enhanced by our ability to recognise and attend to the features they share, and if we never experience their shared features in isolation, we may take them to be entirely unlike even if they are not. Sometimes, of course, two qualities seem inarticulately alike even without an identifiable shared feature; this is why we routinely describe qualities of one modality using terms drawn from another (warm, harsh, sweet, soft, loud, etc.). The SPH and RT5 can both be retained as long as idealised scrutiny of these inchoate likenesses would reveal a system of qualitative connections over our entire experiential range. This view is defended by Coleman:

> [...] just as it's possible to move across the colour spectrum in tiny, almost undetectable steps, it must be possible to move from tastes to sounds, sounds to colors, and so on, via equally tiny steps. Tiptoeing between modalities already seems *conceivable* in certain cases, perhaps even actual. We know that what we experience as "taste" is really some kind of fusion of qualia sourced from the nose and from the tongue [...]. To address qualitative incommensurability we must stretch to conceiving of such continuities as the rule rather than the exception. (Coleman 2017, 264, emphasis in original; cf. Coleman 2015; Hartshorne 1934, 35ff.)

This claim does not seem to me obviously false, but it is at least dubitable. Consequently, the revelation approach may be most threatening to constitutive panpsychists not through any of the three pure revelation arguments, but through intensifying the palette problem. Accepting revelation pushes constitutive panpsychists towards either a large-palette solution like Lewtas's, or towards Coleman's very bold and ambitious form of the small-palette hypothesis.

5 Confusion and Revelation

Identifying a premise of an argument that might be false is often not, by itself, an effective way to persuade critics. For all that I have said so far, this "medium-strength" version of revelation, interpreted and qualified so as to undermine arguments against panpsychism while substantiating arguments against physicalism, might be technically consistent but ad hoc and unmotivated, a dingy corner of logical space which panpsychists can awkwardly retreat to. But in fact, these qualified revelation theses flow from a reasonable picture of the limits of human self-knowledge, on which the introspective ignorance that constitutive panpsychism implies differs only in degree from familiar forms of introspective ignorance.

It is commonplace to say that when two experiences become phenomenally unified, they form a composite experience which subsumes them: they still exist, and are still undergone by the subject, but they are now "undergone together". We easily recognise this when we can discern introspectively not just the composite experience but also its components: but what if the discernibility of the component experiences is not an automatic consequence of the composite experience being composite? We might consider the idea that it depends instead on having the right structure of informational relations among the components.[18] Perhaps if these relations make the subject's overall dynamics differentially sensitive to multiple distinct features of the experience, the composite experience will be characterised by contrast among those features: they will stand out as distinct things. If not, those features will be present in the composite experience in an undifferentiated way, as a single element whose phenomenal quality is a seamless blend of its components. In short: the component experiences all go in together, but the way they are present in the composite experience depends on how they are organised.

What explains why experiences should compose in this way is a further question, which I cannot here address (though see Roelofs 2016; 2019, 123–25, 166–70). But suppose some conditional like this were true: when distinct experiences are unified, they can be distinguished by the subject only if they have the right informational structure. Although the human brain is an exquisitely structured processor of information, it has limits. The overall dynamics of the brain can perhaps be sensitive to whether a neurone fires, but not (as far as we know) to which ions in that neurone played which roles in its firing. Since

[18] This is a long-standing idea among panpsychists, though spelling it out in detail is not simple. See Chalmers (1996), 284–292; Chalmers (2017), 209–210; Gabora (2002); Roelofs (2019), 171–176.

individual events at the microscopic level are informationally inaccessible, they will be experienced by the whole in a blended way. They each make a minute difference to the quality of some element of the whole's experience, but they do not stand out as distinct elements of it. To use a term made famous by Leibniz, they are "confused" with one another, the way that the sounds of each bit of water striking the shore are "confused" in the roar of the sea.[19]

I have elsewhere elaborated more fully on the idea of confusion as I understand it (2019, 126–29), but the essential idea is captured in the following definition:

> Two experiences are *confused* with each other, relative to a subject, iff that subject cannot distinguish them by attending to one without simultaneously attending to the other.[20]

It is important to emphasise that confusion is not a matter of a subject "perceiving" things outside themselves so poorly that they cannot distinguish the parts of that outside thing. Confusion is a matter of how the subject's own states are related, not a relation between them and something external. For example, someone viewing a pointillist painting, for whom the many dots of paint "blur together", is not thereby subject to confusion, if they simply have a single experience that is the product of many external objects. A better example would be someone with an untrained palate, who drinks coffee and experiences (let us stipulate) the same diversity of taste and flavour experiences as a practiced connoisseur but experiences them together as a single blended flavour, without being able to pick out the bitterness from the aroma, etc.

Confusion may depend on circumstances. When we are tired, distracted, or drunk we often cannot distinguish things which we could under better conditions. Then our experiences are confused only relative to those circumstances. Confusion can also depend on a subject's conceptual repertoire: sometimes we cannot distinguish two things using their present concepts, but would

[19] This idea of the mind as comprising a vast number of "little perceptions", most of which cannot be distinguished from one another by the subject, is arguably present in several early modern writers as well as Leibniz, in particular Spinoza, Wolff, and Kant. For discussion see Wilson (1980), Thiel (2011), Liang (2017), and Indregard (2018). To use a more modern phrasing from Andrew Lee (2019), they make up the non-introspectible "microstructure of experience".

[20] In the primary instance confusion is defined over tokens, but we can easily define a secondary sense in which two types are confused for a subject when any token of those types onto which a given subject could direct a given operation would be confused with a token of the other type.

be able to if we learnt new ones. Call confusion which can be removed by adjusting the subject's bodily surroundings or condition, or improving their conceptual repertoire, or in some similarly mild way, "shallow confusion", and call confusion which persists even into ideal conditions, "robust confusion".

In between shallow and robust is confusion which persists until the subject becomes distinctly acquainted with a token of the same type as the confused elements. For example, suppose the sensory component of pain is robustly confused with the unpleasant affect pain involves, except for subjects who have experienced "pain asymbolia", the rare condition of feeling pain without finding it at all unpleasant (cf. Grahek 2007; Klein 2015). If they regain normal pain experiences, they might find themselves newly able to attend to its sensory element in isolation. If this were to happen, we might say that their original confusion was "nearly-robust": removable only by somehow acquainting them with (a token of the same type as) one of the confused elements on its own.[21]

When confusion is shallow, we have an easy way to tell that we suffer from it: we remove it and contrast the resulting distinction with the earlier confusion. With sufficiently robust confusion, we would not have such means of recognising it; we could not tell that we were confused. And if we suffered from confusion that was "nearly-robust", it would be undetectable, except by means of independent acquaintance with elements of the same type as the confused ones. We could, that is, be subject to a lot of confusion without being able to tell, introspectively. And if constitutive panpsychism is true - in particular, if micro-experiences corresponding to all the physical details of our brains were somehow present in our consciousness - then we should expect just that: all the experiences of our microparts would be confused relative to us. Call this the Radical Confusion Hypothesis.

Confusion is defined functionally, but that does not imply that confusion is a purely functional fact that makes no phenomenal difference. My suggestion is that undergoing two confused experiences feels different to undergoing two distinguishable experiences, even if those experiences are the same in all intrinsic respects. When the components of an experience are distinguishable by the subject, they are phenomenally present as discernible, separate, parts - there is an experience of phenomenal contrast, of things standing out against

[21] In other work (2019, 128–29), I also distinguish between "strong" and "weak", and "symmetrical" and "asymmetrical" confusion, but this does not substantially affect the argument so I omit it here for simplicity.

other things. But when they are confused, they are present qualitatively, as contributions to the total quality of the experience they blend into.

How would the Radical Confusion Hypothesis help with the four revelation arguments? Recall that in response to the "no illusions" argument, I denied premise 2: that human consciousness positively appears introspectively to be "smooth" (there defined as "not particulate"). I maintained that this is false if "appears introspectively" is read in a strong, quasi-perceptual sense; it is true only if "appears introspectively" is read in a weaker sense, as meaning "it is easy and natural to interpret experience this way".

Now I can say *why* this misinterpretation is easy and natural: because many human experiences display something close to "smoothness", namely, all their component experiences are nearly-robustly confused with each other, distinguishable only by a subject who already knows what to look for. A subject who lacks any distinct acquaintance with the ingredients will be unable to distinguish them or discern their internal structure. We might say that experiences all of whose components are confused with one another are "pseudo-smooth", and it is true (and introspectively obvious!) that many of our experiences are pseudo-smooth. But to infer genuine smoothness from pseudo-smoothness is a metaphysical over-interpretation which goes beyond the introspective deliverances: it is inferring absence of structure from the failure of structure to be manifest in a certain way (it is thus very similar to the "headless woman illusion" discussed by Armstrong (1968), where not seeing someone's head gives us the vivid but false impression that they have no head). The noticeable quality that some experiences have, which prompted the "no illusions" argument, is just what radical confusion feels like.

Second, in response to the macroexperience-focused argument I denied premise 4, that whatever constitutes something is part of the "nature" that is revealed to us by pure phenomenal concepts. I suggested that a priori reflection tells us only those necessary truths that involve *only* properties whose nature we know - such as whether one could constitute the other. But just knowing the nature of one property does not tell all the things that could constitute it, nor what constitutes a particular instance of it.

I can now elaborate on this distancing of constitution from "nature". Macroexperiences are composite experiences composed of many microexperiences confused with one another. Their phenomenal character is determined by combining the phenomenal characters of those component experiences, which they subsume in fundamentally the same way that a person's total experience at any one time subsumes the partial experiences they are having

doi: 10.48106/dial.v74.i1.04

at that time. But just as two composites might end up sharing certain properties despite being constituted by different sets of parts, and despite their properties being mere combinations of the properties of their parts, two composite experiences might have the same phenomenal character, despite being constituted by different sets of microexperiences. The particular parts might be essential to the particular macroexperience, but not to the property that it is an instance of.

I also said, in response to the small-palette revelation argument, that distinct macroexperiences might arise from the same neural basis: we need not assume that each distinguishable element of our consciousness contains the entire phenomenal nature of one discrete subset of physical entities. The radical confusion hypothesis reinforces this point: it says that which experiences phenomenally contrast or phenomenally blend with one another in human experience reflects the informational structure of the brain, so a single macroexperience might not correspond to any discrete section of the underlying physical substrate. Instead, it will correspond to a set of features of the substrate such that information about them collectively is extracted and used by the brain, but information about them individually is not. Thus different macroexperiences based in the same brain area might have different, even non-overlapping sets of phenomenal ingredients, because they reflect different features of the same microexperiences.

Finally, in response to the microexperience-focused argument I suggested that our ignorance of microexperiences is compatible with our undergoing them, if we cannot attend to them. Now I can add that our inability to attend to microexperiences is part-and-parcel of their being confused for us. Their radical confusion is explained by the limitations discussed above on how much information about microscopic brain events can be extracted by the rest of the brain.[22] Because radically confused experiences cannot be distinctly attended to, we cannot know them or their natures, even though the experiences "present themselves" in the sense that if their subject could attend to them they could know them and their natures by introspection.

An opponent might object that even though attending to particular experiences can be harder or easier, depending on, e.g. architectural facts about the brain, it cannot be *strictly impossible* for me to attend to an experience, if it is really is an experience I am undergoing. I reply that distinctly attending

[22] This allows for a limited sense in which microexperiences *are* accessible: namely that they can be accessed only by acts which are also accessing many other microexperiences at the same time. They cannot be *individually* accessed, but they can be accessed *collectively*.

to microexperiences is *not* strictly impossible, just impossible in practice (as discussed in Lee 2019, 296–97). They are manifest in our consciousness, but incredibly difficult to pick out. After all, it is very difficult for the large-scale dynamics of our brain to be sensitive to changes in a single particle, but there is no in-principle impossibility in there being such sensitivity, perhaps using advanced technology or strange altered states of consciousness.[23]

6 Conclusions

The idea of "revelation", that having an experience provides a special insight into its nature, is a key weapon in the armoury of anti-physicalists. But for constitutive panpsychists there is a risk it will blow up in their faces. I have argued, however, that a suitably-qualified form of the revelation approach can bring down physicalism while leaving panpsychism standing: a form which reconciles the profound fallibility of the human mind's self-knowledge with the perfect transparency of its access to its itself. Although nothing does or could "conceal" our own experiences from us, we are nevertheless limited in our ability to attend to their elements, prone to misinterpret them, and consequently unable to tell introspectively just how composite they might really be.[*]

Luke Roelofs
New York University
luke.mf.roelofs@gmail.com

References

ALTER, Torin, and Yujin NAGASAWA. 2012. "What Is Russellian Monism?" *Journal of Consciousness Studies* 19 (9–10): 67–95.

[23] Note that there need not be any sharp boundary between "the simplest experiential element that we can distinguish" and "the most complex experiential element that is radically confused." For different people, under different conditions, different distinctions among one's internal states and processes may be possible. Radically confused experiences are not a qualitatively distinct sort of experience from distinguishable ones, any more than "places I can walk to in ten minutes" are a sharply separate set of places from those I can walk to in ten minutes; my walking ability, like my introspective discernment, waxes and wanes as I change and as conditions change.

[*] This paper expands on ideas presented over pages 132–137 of Roelofs (2019). Their further development owes a great deal to audiences at the Australian National University and the CEU's workshop "Russellian Monism: Time for the Details" in Budapest.

doi: 10.48106/dial.v74.i1.04

Armstrong, David M. 1968. "The Headless Woman Illusion and the Defence of Materialism." *Analysis* 29 (2): 48–49. doi:10.1093/analys/29.2.48.

Ball, Derek. 2009. "There Are No Phenomenal Concepts." *Mind* 118 (472): 935–62. doi:10.1093/mind/fzp134.

Braddon-Mitchell, David, and Robert Nola, eds. 2009. *Conceptual Analysis and Philosophical Naturalism*. Cambridge, Massachusetts: The MIT Press.

Byrne, Alex, and David R. Hilbert. 2007. "Color Primitivism." *Erkenntnis* 66 (1–2): 73–105. doi:10.1007/s10670-006-9028-8.

Chalmers, David J. 1995. "Absent Qualia, Fading Qualia, Dancing Qualia." In *Conscious Experience*, edited by Thomas Metzinger, 309–29. Paderborn: Ferdinand Schöningh.

———. 1996. *The Conscious Mind: In Search of a Fundamental Theory*. Oxford: Oxford University Press.

———. 2003a. "Consciousness and Its Place in Nature." In *The Blackwell Guide to the Philosophy of Mind*, edited by Stephen P. Stich and Ted A. Warfield, 102–42. Blackwell Philosophy Guides. Oxford: Basil Blackwell Publishers.

———. 2003b. "The Content and Epistemology of Phenomenal Belief." In *Consciousness: New Philosophical Perspectives*, edited by Aleksandar Jokić and Quentin Smith, 220–72. Oxford: Oxford University Press.

———. 2009. "The Two-Dimensional Argument Against Materialism." In *The Oxford Handbook of Philosophy of Mind*, edited by Brian P. McLaughlin, Ansgar Beckermann, and Sven Walter, 313–37. Oxford Handbooks. Oxford: Oxford University Press.

———. 2017. "The Combination Problem for Panpsychism." In *Panpsychism. Contemporary Perspectives*, edited by Godehard Brüntrup and Ludwig Jaskolla, 179–214. Oxford: Oxford University Press.

Clark, Andy. 1989. "The Particulate Instantiation of Homogeneous Pink." *Synthese* 80 (2): 277–304. doi:10.1007/bf00869488.

Coleman, Sam. 2015. "Neuro-Cosmology." In *Phenomenal Qualities. Sense, Perception, and Consciousness*, edited by Paul Coates and Sam Coleman, 66–102. Oxford: Oxford University Press.

———. 2017. "Panpsychism and Neutral Monism: How to Make up One's Mind." In *Panpsychism. Contemporary Perspectives*, edited by Godehard Brüntrup and Ludwig Jaskolla, 249–82. Oxford: Oxford University Press.

Damnjanovic, Nic. 2012. "Revelation and Physicalism." *Dialectica* 66 (1): 69–91. doi:10.1111/j.1746-8361.2012.01290.x.

Dennett, Daniel C. 2007. "What RoboMary Knows." In *Phenomenal Concepts and Phenomenal Knowledge: New Essays on Consciousness and Physicalism*, edited by Torin Alter and Sven Walter, 15–31. Oxford: Oxford University Press.

Díaz-León, Esa. 2011. "Reductive Explanation, Concepts, and a Priori Entailment." *Philosophical Studies* 155 (1): 99–116. doi:10.1007/s11098-010-9560-x.

Gabora, Liane. 2002. "Amplifying Phenomenal Information: Toward a Fundamental Theory of Consciousness." *Journal of Consciousness Studies* 9 (8): 3–29.

Gaudry, Justin. 2008. "Does Physicalism Entail Cosmopsychism?" https://panexperientialism.blogspot.com/2008/05/does-physicalism-entail-cosmopsychism.html.

Goff, Philip. 2006. "Experiences Don't Sum." *Journal of Consciousness Studies* 13 (10–11): 53–61.

———. 2015. "Real Acquaintance and Physicalism." In *Phenomenal Qualities. Sense, Perception, and Consciousness*, edited by Paul Coates and Sam Coleman, 121–45. Oxford: Oxford University Press.

———. 2017. *Consciousness and Fundamental Reality*. Oxford: Oxford University Press.

Grahek, Nikola. 2007. *Feeling Pain and Being in Pain*. 2nd ed. Cambridge, Massachusetts: The MIT Press.

Hartshorne, Charles. 1934. *The Philosophy and Psychology of Sensation*. Chicago, Illinois: University of Chicago Press.

Indregard, Jonas Jervell. 2018. "Consciousness as Inner Sensation: Crusius and Kant." *Ergo* 5 (7): 173–201. doi:10.3998/ergo.12405314.0005.007.

Jackson, Frank. 1982. "Epiphenomenal Qualia." *The Philosophical Quarterly* 32 (127): 127–36. doi:10.2307/2960077.

Jaskolla, Ludwig, and Alexander J. Buck. 2012. "Does Panexperiential Holism Solve the Combination Problem?" *Journal of Consciousness Studies* 19 (9–10): 190–99.

Johnston, Mark. 1992. "How to Speak of the Colors." *Philosophical Studies* 68 (3): 221–63. doi:10.1007/bf00694847.

Kammerer, François. 2018. "Can You Believe It? Illusionism and the Illusion Meta-Problem." *Philosophical Psychology* 31 (1): 44–67. doi:10.1080/09515089.2017.1388361.

Klein, Colin. 2015. "What Pain Asymbolia Really Shows." *Mind* 124 (494): 493–516. doi:10.1093/mind/fzu185.

Kriegel, Uriah. 2009. *Subjective Consciousness: A Self-Representational Theory*. New York: Oxford University Press.

Kripke, Saul A. 1980. *Naming and Necessity*. Oxford: Basil Blackwell Publishers.

Lee, Andrew. 2019. "The Microstructure of Experience." *Journal of the American Philosophical Association* 5 (3): 286–305. doi:10.1017/apa.2019.4.

Lewis, David. 1995. "Should a Materialist Believe in Qualia?" *Australasian Journal of Philosophy* 73 (1): 140–44. doi:10.1080/00048409512346451.

Lewtas, Patrick Kuehner. 2013. "What Is It Like to Be a Quark?" *Journal of Consciousness Studies* 20 (9–10): 39–64.

Liang, Yibin. 2017. "Kant on Consciousness, Obscure Representations and Cognitive Availability." *The Philosophical Forum* 48 (7): 345–68. doi:10.1111/phil.12169.

Liu, Michelle. 2019. "Phenomenal Experience and the Thesis of Revelation." In *Perception, Cognition and Aesthetics*, edited by Dena Shottenkirk, Manuel Curado, and Steven S. Gouveia, 227–51. Routledge Studies in Contemporary Philosophy 119. New York: Taylor & Francis.

———. 2020. "Explaining the Intuition of Revelation." *Journal of Consciousness Studies* 27 (5–6): 99–107.

LOCKWOOD, Michael. 1993. "The Grain Problem." In *Objections to Physicalism*, edited by Howard Robinson, 271–92. Oxford: Oxford University Press.

MAXWELL, Grover. 1978. "Rigid Designators and Mind-Brain Identity." In *Minnesota Studies in the Philosophy of Science, Volume IX: Perception and Cognition: Issues in the Foundations of Psychology*, edited by C. Wade Savage, 365–404. Minneapolis, Minnesota: University of Minnesota Press.

MCGINN, Colin. 2006. "Hard Questions: Comments on Galen Strawson." *Journal of Consciousness Studies* 13 (10–11): 90–99.

NAGASAWA, Yujin, and Kai WAGER. 2017. "Panpsychism and Priority Cosmopsychism." In *Panpsychism. Contemporary Perspectives*, edited by Godehard Brüntrup and Ludwig Jaskolla, 113–29. Oxford: Oxford University Press.

NEMIROW, Lawrence. 1990. "Physicalism and the Cognitive Role of Acquaintance." In *Mind and Cognition*, edited by William G. Lycan, 490–99. Oxford: Basil Blackwell Publishers.

PEREBOOM, Derk. 2016. "Illusionism and Anti-Functionalism about Phenomenal Consciousness." *Journal of Consciousness Studies* 23 (11–12): 172–85.

———. 2019. "Russellian Monism, Introspective Inaccuracy, and the Illusion Meta-Problem of Consciousness." *Journal of Consciousness Studies* 26 (9–10): 182–93.

ROELOFS, Luke. 2014. "Phenomenal Blending and the Palette Problem." *Thought* 3 (1): 59–70. doi:10.1002/tht3.113.

———. 2016. "The Unity of Consciousness, Within Subjects and Between Subjects." *Philosophical Studies* 173 (12): 3199–3221. doi:10.1007/s11098-016-0658-7.

———. 2019. *Combining Minds: How to Think about Composite Subjectivity*. Oxford: Oxford University Press.

ROSENTHAL, David M. 1993. "State Consciousness and Transitive Consciousness." *Consciousness and Cognition* 2: 355–63. doi:10.1006/ccog.1993.1029.

SEAGER, William E. 1995. "Consciousness, Information and Panpsychism." *Journal of Consciousness Studies* 2 (3): 272–88.

SHANI, Itay. 2015. "Cosmopsychism: A Holistic Approach to the Metaphysics of Experience." *Philosophical Papers* 44 (3): 389–437. doi:10.1080/05568641.2015.1106709.

SHOTTENKIRK, Dena, Manuel CURADO, and Steven S. GOUVEIA, eds. 2019. *Perception, Cognition and Aesthetics*. Routledge Studies in Contemporary Philosophy 119. New York: Taylor & Francis.

SMITHIES, Declan. 2019. *The Epistemic Role of Consciousness*. Oxford: Oxford University Press.

STOLJAR, Daniel. 2001. "Two Conceptions of the Physical." *Philosophy and Phenomenological Research* 62 (2): 253–81. doi:10.1111/j.1933-1592.2001.tb00056.x.

———. 2006. *Ignorance and Imagination*. Oxford: Oxford University Press.

———. 2009. "The Argument from Revelation." In *Conceptual Analysis and Philosophical Naturalism*, edited by David Braddon-Mitchell and Robert Nola, 113–38. Cambridge, Massachusetts: The MIT Press.

———. 2013. "Qualitative Inaccuracy and Unconceived Alternatives." *Philosophy and Phenomenological Research* 86 (3): 745–52. doi:10.1111/phpr.12030.

Strawson, Galen. 2006. "Realistic Monism: Why Physicalism Entails Panpsychism." *Journal of Consciousness Studies* 13 (10–11): 3–31. doi:10.1093/acprof:oso/9780199267422.003.0003.

———. 2015. "Self-Intimation." *Phenomenology and the Cognitive Sciences* 14: 1–31. doi:10.1007/s11097-013-9339-6.

Thiel, Udo. 2011. *The Early Modern Subject. Self-Consciousness and Personal Identity from Descartes to Hume*. Oxford: Oxford University Press.

Wilson, Margaret Dauler. 1980. "Objects, Ideas, and 'Minds': Comments on Spinoza's Theory of Mind." In *The Philosophy of Baruch Spinoza*, edited by Richard Kennington, 103–20. Studies in Philosophy and the History of Philosophy 7. Washington, District of Columbia: The Catholic University of America Press.

doi: 10.48106/dial.v74.i1.04

In Defence of Facts
Grounding, Essential Properties and the Unity Problem

Donnchadh O'Conaill

A common conception of facts is as worldly entities, complexes made up of non-factual constituents such as properties, relations and property-bearers. Understood in this way facts face the unity problem, the problem of explaining why various constituents are combined to form a fact. In many cases the constituents could have existed without being unified in the fact—so in virtue of what are they so unified? I shall present a new approach to the unity problem. First, facts which are grounded are unified by the obtaining of their grounds. Second, many ungrounded facts are such that they must obtain if their non-factual constituents exist (e.g. if the property *F*ness is essential to a particular, *a*, then if *a* exists the fact that *a* is *F* must obtain). In this way the obtaining of these facts is explained by the essence of some of their constituents. I also address the possibility of facts which are brutely unified (i.e. neither grounded nor essentially unified), and compare the account I offer with some of the main alternatives.

It is common for facts to be understood as worldly entities, complexes made up of non-factual constituents such as properties, relations and property-bearers. A number of authors have presented a problem for facts understood in this way, the problem of unity. This is the problem of explaining the difference between the existence of all the constituents of a fact and the obtaining of that fact. For instance, a particular entity *a* might exist and a property *F*ness might be instantiated, but it does not follow that the fact that *a* is *F* obtains—so in virtue of what does this fact obtain?

In this paper I offer a new line of defence against the unity problem. After outlining the compositional conception of facts in section 1, I shall state the unity problem in section 2 and possible responses to it in section 3. In section 4 I outline the first part of my defence, which appeals to the notion of grounding:

if a fact is grounded, its unity is explained by the obtaining of its grounds. This raises the issue of whether there are facts which are not grounded, and if so how the unity of these facts can be explained. In section 5 I consider how the unity problem might be addressed if every fact were grounded. In section 6 I propose that many facts which are not grounded are plausibly such that the properties they involve are essential to their property-bearers. Because of this, the constituents of these facts are essentially unified. In section 7 I address the possibility that there could be ungrounded facts which are not essentially unified, facts whose unity is not explained in either of the two ways I propose. In section 8, I briefly compare my account of the unity of facts with alternative views proposed by Arianna Betti and William Vallicella. While the account I offer has certain drawbacks compared to these alternatives, it also has important advantages, and should be taken as seriously as any other account of the unity of facts.

1 The Compositional Conception of Facts

On the *compositional conception*, a fact is a complex entity made up of non-factual constituents (hereafter "constituents"). In this section I shall present some key aspects of facts thus understood.[1]

First, facts are *non-representational entities*: they do not have truth- or accuracy-conditions, nor do they refer to or designate anything, and they are not about anything in the sense in which intentional states are about their objects. Some facts will include representational entities among their constituents (e.g. the fact that the sentence "Tom is wet" is true). But such facts do not themselves represent anything. Furthermore, facts are not metaphysically posterior to propositions which state them; e.g. the identity of the fact that *a* is *F* is not metaphysically determined by the proposition "*a* is *F*" (in contrast with what Kit Fine terms the propositional conception of facts—see his 1982, 51–52).

Second, I take facts to be composed of property-bearers and properties (for the purposes of this paper I include relations among the properties). Both properties and property-bearers are relatively coarse-grained entities: for instance, the property *being water* is identical with the property *being composed of H_2O molecules*, whereas the concepts "being water" and "being

[1] A more thorough statement of this conception is offered by Betti (2015, 7, 18–30; see also Vallicella 2016a, 115). In what follows I shall ignore questions concerning states of affairs as distinct from facts.

composed of H_2O molecules" are distinct. Correspondingly, the fact composed of this property and a certain mass of material (e.g. the fact that this body of liquid is water) is more coarse-grained than the proposition "this body of liquid is composed of water." The properties which help to make up facts are universals. Property-bearers can be either particulars or universals.[2]

Third, because facts are composed of entities which exist and help to make up the world, I take it that facts themselves help to make up what exists; in this sense, they are *worldly* entities (see Betti 2015, 22–24). This conception of facts thus closely corresponds to one rejected by P.F. Strawson, according to whom a fact "is not something in the world. It is not an object; not even (as some have supposed) a complex object consisting of one or more particular elements (constituents, parts) and a universal element (constituent, part)" (1950, 135).

Fourth, the way in which a fact's constituents are combined to make up that fact is *non-mereological*. In the present context this can be understood as follows: for a fact to obtain is not the same as for its constituents to exist; rather, it is for its constituents to exist *and* be combined or arranged in some specific way (Betti 2015, 65).[3] For instance, the fact that *a* is *F* (hereafter, "*Fa*") obtains only if *a instantiates* the property *F*ness; the fact that *a* is larger than *b* obtains only if *a* and *b* stand (in a particular order) in the relation *larger than*. It is helpful to have a term which allows us to contrast the existence of all the constituents of a fact with the obtaining of this fact. When all the constituents exist, I shall refer to them as forming an *aggregate*, where for an aggregate of entities to exist just is for each entity in the aggregate to exist. One might then say that whereas an aggregate is a mereological sum, a fact is a non-mereological complex (Armstrong 1997, 119–22; Meinertsen 2008, 3).[4]

[2] I assume a sparse view of properties, on which it is not the case that each predicate corresponds to a distinct property or relation. For the most part this will not matter in what follows, but it is worth noting that I do not assume that formal ontological predicates such as "instantiates" correspond to distinct properties. Therefore, I do not accept that facts have so-called "secondary" constituents, e.g. a relation of instantiation or a non-relational tie which binds *a* and *F*ness. I shall mostly use examples of facts containing one property-bearer and one property; however, the compositional conception is not itself committed to this restriction.

[3] Alternatively, this form of composition can be understood as involving non-extensional mereology (Bennett 2013, 101–2). The notion of non-mereological composition has been challenged by David Lewis (e.g. 1986), but it is accepted by all proponents of the compositional conception of facts. As I understand it, the problem of unity is based on accepting this conception and raising a challenge concerning facts understood in this way.

[4] I do not intend talk of aggregates to be ontologically committing—I use it for convenience and if necessary it could be replaced by plural quantification over the constituents (Betti 2015, 53).

doi: 10.48106/dial.v74.i1.05

It is frequently claimed that because facts exhibit non-mereological unity, the existence of the constituents of a fact does not itself suffice for the obtaining of that fact (Vallicella 2000, 246; Betti 2015, 54). I shall question this claim later, but for the time being I accept it. It is also often claimed that a fact is something *over and above* its constituents. This claim is sometimes supported by the contention that the constituents can exist without the fact obtaining (Vallicella 2000, 238). It is also sometimes supported by appealing to the non-mereological composition of facts: “philosophers who do accept facts say that when Hargle is sad, alongside these two things (Hargle and sadness) there is also a third thing in the world: a special ‘being together’ of these two things in a real unity over and above the two things” (Betti 2015, 30). I shall return to these claims in section 6.

Fifth, I accept what Gonzalo Rodriguez-Pereyra terms the *structuralist criterion* of fact identity: “facts are identical if and only if they have the same constituents combined in the same way” (1998, 520). That is, fact *A* is identical with fact *B* iff (i) the constituents of *A* are all identical with the constituents of *B*, and vice-versa; (ii) the mode of combination of the constituents in *A* is identical with the mode of combination of the constituents in *B*; (iii) *A* obtains at exactly the same time as *B*. By “mode of combination”, I mean the specific kind of non-mereological composition which characterises each fact. This could be that a particular instantiates a property, or that two entities stand in a certain relation. In the case of asymmetric relations, it would also include entities standing in a certain order in that relation, so that, e.g. *aRb* would involve a different mode of combination than *bRa*.

This criterion suggests the following asymmetry: while the identities of the constituents of a fact help to determine its identity-conditions, the reverse does not hold (Vallicella 2016a, 117). Therefore, on the compositional conception “facts are built up out of ontologically more basic materials” (2016a, 115). This view of facts can be contrasted with one in which the constituents of facts are abstractions from them, such that the identity of the constituents is determined by the identity of the facts to which they belong.

Finally, I assume that whenever the constituents of a fact A are arranged in the mode of combination characteristic of A, A thereby obtains (e.g. if a property-bearer *a* instantiates a property *F*ness, the fact *Fa* obtains).[5] This

For ease of presentation, I shall write as though a property must be instantiated in order to exist—however, the discussion can easily be adapted to accommodate a Platonist conception of properties.

[5] This formulation sets aside issues to do with the time at which the constituents are arranged.

assumption can be challenged.[6] For instance, in E.J. Lowe's four-category ontology when a universal property is had by a property-bearer we do not need to posit a fact; rather, the property-bearer is characterised by a particular property or mode (2006). But while one could object to facts in this way, this seems to be a different issue to the problem of unity.[7]

2 The Problem of Unity

Given the above conception of facts, the unity problem is relatively easy to outline. Consider Fa. For this fact to obtain, its constituents (a and Fness) must be combined in a specific, non-mereological manner; only in this way will they achieve the kind of unity characteristic of a fact. This kind of unity between a and Fness would be absent if, for instance, a existed and Fness was instantiated, but a did not instantiate Fness (e.g. if some entity b, wholly distinct from a, instantiated Fness). In that scenario, a would exist and Fness would be instantiated, but they would not exist together in the way characteristic of Fa.

The unity problem is simply the problem of explaining *why*, given that specific non-factual entities (e.g. a and Fness) each exist, they are united to form the fact Fa. More generally, it is the problem of explaining for any fact why, given that its constituents each exist, they are unified in the specific mode of combination characteristic of that fact. A solution to this problem for a specific fact, A, would be a metaphysical explanation of why, given that the constituents of A each exist, they are arranged in the mode of combination characteristic of A. As Betti puts it, the problem is "how to account for the unity of relations with their relata and for the unity of properties with their bearers" (2015, 42). Elsewhere she glosses the problem as the search for "something *in virtue of which* those constituents form a unity" (2015, 45; see also Vallicella 2000, 242; Orilia 2006, 214; Meinertsen 2008, 3).[8]

[6] Thanks to Jani Hakkarainen for raising this point.

[7] The unity problem could be reframed as a problem concerning the unity of specific property-bearers and properties, without mentioning facts. With regard to Lowe's ontology, the problem would be that of explaining why a specific property-bearer has the modes which characterise it.

[8] Metaphysical explanations, the kind of explanations expressed by "in virtue of" claims, encompass grounding explanations but also other forms of explanation (see fn. 36 below). Katarina Perovic suggests that the problem of unity consists of "a cluster of problems that are frequently run together" (2016, 145). I have some sympathy with this view, but I suggest that the problem outlined in the main text does not conflate different issues. In terms of the various problems Perovic distinguishes, the problem of unity corresponds both to what she terms the explanatory

The unity problem thus characterized is relatively straightforward to grasp, but there are potentially complicating factors which must be addressed. The first is that the problem is often described in such a way that it seems to presuppose the possibility that the constituents of a fact might exist and the fact not obtain. For instance, Vallicella asks: "What makes it the case that a number of constituents of the right kinds—constituents which are connectable so as to form a fact but *need not be connected to exist*—are actually connected so as to form an actual or existing fact?" (Vallicella 2000, 242, italics added). Here the italicized phrase expresses the assumption that the existence of the constituents need not entail the obtaining of the fact (see also Dodd 1999, 159; Wieland and Betti 2008, 510; Betti 2015, 54; Perovic 2016, 144). In what follows I shall not make this assumption, though I postpone discussion until section 6.

Second, the unity problem as I have characterised it should be distinguished from different problems with which it might be confused. For instance, at one point in their discussion Betti and Jan Wieland ask, "What grounds the difference between mereological and unmereological composition?" (2008, 513). Wieland and Betti present this as a restatement of the original unity problem,[9] but I think it is a different problem. The unity problem is the question, for any specific fact, of what it is in virtue of which its constituents are unified. An answer to this problem may in principle apply to any fact, but it will not itself explain the difference between mereological and non-mereological composition. It may be that the difference between mereological and non-mereological composition cannot be explained, but it would not follow that the unity problem cannot be solved (Vallicella makes a similar point in his 2000, 242). Similarly, Julian Dodd glosses the supposed obscurity of the unity of facts as the "problem of the nature of instantiation" (1999, 156). But we need to distinguish between an account of *what* instantiation is (an answer to the problem concerning its nature) and an account of *why* specific entities instantiate specific universals (an answer to the unity problem).[10]

problem and the Mereological Problem of Unity (2016, 146–49). I suggest that these are really the same problem. What Perovic calls the Mereological Problem concerns the ontological ground of the difference between a fact and the aggregate of its constituents. In this context, the ontological ground is whatever explains the unity of the constituents in the fact; it is that in virtue of which they together form a fact.

[9] They introduce the quoted passage by saying, "We can restate the problem immediately" (2008, 513). The context makes it clear that by "the problem" they mean the unity problem.

[10] We also need to distinguish each of these accounts from an account of *how it is possible* for distinct entities such as a and Fness to be unified (Dodd 1999, 151; Vallicella 2002, 26; Maurin

More generally, there is a difference between giving an account of *what it is for a* and *F*ness to form a fact, and explaining *why a* and *F*ness are so combined. It is perfectly legitimate to answer the first question by citing the characteristic unity of the fact. For instance, one might be contrasting the unity of a fact with the unity of parts in a mereological sum, or members in a set, in which case it makes sense to refer to *a* instantiating *F*ness. But it is not legitimate to answer the second question by citing this very unity: it is no use explaining why *a* and *F*ness are unified in *Fa* by saying that *a* instantiates *F*ness. This would be to simply re-describe what one was asked to explain.[11]

It is crucial to distinguish the unity problem from these other problems (concerning the nature of instantiation, or the difference between mereological and non-mereological composition). These questions concern the very coherence of a theory of facts; they arise insofar as the very idea of facts is considered obscure. The unity problem, in contrast, is based on the assumption that the idea of a fact is coherent. It is only given a coherent notion of facts that the distinction between a fact and the aggregate of its constituents can clearly be drawn; and it is only given this distinction that the unity problem can be posed. Therefore, in addressing the unity problem we can set these other questions aside.

3 Possible Solutions

The unity problem has been developed into an argument against facts by a number of different writers (Dodd 1999, 152; Wieland and Betti 2008, 509; Betti 2015, 51). Though details differ, each version of the argument works in roughly the same way: postulating facts gives rise to the unity problem; there are a determinate number of possible solutions to this problem available; none of these solutions succeed; therefore, we should not postulate facts.[12]

To structure the discussion I shall refer to the range of possible solutions outlined by Betti (2015, 51). The unity of a fact could be explained by:

(A) the constituents of the fact, e.g. *a* and/or *F*ness;

2015, 212–13), or an account of how non-mereological composition is possible (see Eklund's characterisation of the problem of unity for facts in his 2019, 1236–37).

[11] For further discussion of the difference between questions concerning what something is and questions concerning why it is (as it is), see Audi (2015).

[12] One possible response to this argument would be to claim that facts can obtain without their unity being explained at all. I shall consider this possibility in section 7.

doi: 10.48106/dial.v74.i1.05

(B) one or more additional constituents of the fact, i.e. a constituent which is identical to neither *a* nor *F*ness;[13]
(C) something external to the fact, i.e. something numerically distinct from either the fact or any of its constituents; or
(D) the fact itself.

Bo Meinertsen outlines a version of (B). Vallicella and Francesco Orilia argue for different versions of (C).[14] Betti interprets David Armstrong as in effect putting forward a version of (D). Dodd, Wieland and Betti argue that none of these options can work, and that the unity problem cannot be solved.

Wieland and Betti offer a dissolution of the problem which in effect rejects the assumption that a fact is something other than the aggregate of its constituents. This argument appeals to the notion of *bearer-specific properties*. A property is bearer-specific iff it is in its nature to be had by a specific bearer or bearers (Betti 2015, 90).[15] So if *F*ness is a property specific to *a*, *F*ness is such that necessarily if it exists, it is instantiated by *a*. All tropes are bearer-specific properties, but Wieland and Betti deny that all bearer-specific properties are tropes, since it can be in the nature of some bearer-specific properties to be had by many specific entities (2008, 519).[16] If properties are bearer-specific, then the unity problem would be dissolved: *a* would instantiate *F*ness as soon as *F*ness exists, and therefore there would be no difference between the fact and the aggregate of its constituents (Betti 2015, 92).

The response I shall offer to the unity problem does not fall neatly into any of Betti's options—or rather, different parts of the response fall into different options. I shall begin by outlining a version of option (C), though distinct from those offered by Vallicella or Orilia. Whether or not this version of (C) solves the unity problem for all facts depends on further assumptions. If there are facts to which it does not apply, then some other response to the problem must be offered. I shall offer a further response which can be read as a version of Betti's option (A), or as dissolving the problem in a manner similar to her appeal to bearer-specific properties. It is also important to note that, on my

[13] For instance, one might treat the relation of instantiation as a further constituent of the fact.

[14] Dixon (2018) can be understood as proposing a version of (C), though he does not specifically discuss the problem of unity.

[15] Betti prefers the phrase "relata-specific relations" (2015, 89–90). Since I am treating relations as among the properties, this difference is not important.

[16] The notion of bearer-specific properties is criticised by Vallicella (2016b, 237–40). It is defended by Wieland and Betti (2008, 521–22), and by Betti (2015, 93–96). I discuss it in section 8 below.

approach, there may be facts to which none of options (A)-(D) applies; in section 7 I shall defend the possibility of such facts.

4 Grounding and Unity

In this section I shall outline a specific conception of grounding and argue that it can help explain the unity of grounded facts.

The terminology of ground is frequently used to in order to set out the unity problem (Vallicella 2000, 243; Wieland and Betti 2008, 510–11; Betti 2015, 55). It may therefore seem odd to appeal to a notion of grounding in order to solve this problem. But the appearance of oddity here is easily explained. The notion of "ground" used to state the unity problem simply indicates whatever could solve it (i.e. whatever it is in virtue of which a fact is unified, or whatever explains its unity). Therefore, theorists writing about the unity problem have not needed to say a great deal about this notion. For instance, neither Betti nor Vallicella systematically characterize this notion or attempt to relate their use of it to the recent literature on metaphysical grounding. In contrast, the conception of grounding which I shall outline is in large part drawn from this literature. And because I am planning to put this conception to constructive use in solving the problem, I will need to say more about it.

Grounding is a form of metaphysical determination often linked to certain non-causal "in virtue of" explanations (e.g. mental facts obtain in virtue of the obtaining of certain physical facts; entities possess dispositional properties in virtue of possessing categorial properties; actions have moral properties in virtue of certain of their non-moral properties). I take grounding to be a worldly relation which underwrites some of these explanations, in much the same way that causation is typically thought of as a worldly relation which underwrites causal explanations (Audi 2012, 691; Schaffer 2016a, 84). I assume that grounding is irreflexive, asymmetric, transitive, non-monotonic, and that the full grounds of an entity necessitate the existence of that entity.[17] Many of these assumptions have been questioned in the literature, but together they form a familiar and recognisably orthodox conception of grounding.[18]

To this conception I shall need to add more detail about the relata of grounding and the specific way or ways in which they are related. I assume that

[17] On the distinction between full and partial grounds, see Fine (2012), 50. I assume that if a fact is partially grounded, it must be fully grounded.

[18] For example, Rodríguez Pereyra (2015) challenges irreflexivity, transitivity and asymmetry; Raven (2013) defends all three features.

grounding holds between facts understood on the compositional characterisation. This is not part of the orthodox view: it is common for grounding theorists to speak of facts being grounded, but they are usually non-committal as to the nature of these facts. However, the idea that worldly facts can be related by grounding is at least a familiar one (Audi 2012, 687; Raven 2012, 689; Trogdon 2018, 1289). If there are worldly facts then they look like good candidates to be grounded, assuming anything can be grounded at all.

What is it for a worldly fact to be grounded? There is probably no non-circular definition or analysis of grounding,[19] but we can still say something informative about it. Examples of informative but circular accounts are found elsewhere in philosophy. For instance, it is possible to think that knowledge cannot be analysed in a non-circular fashion, and also that it is informative to learn that knowledge must satisfy a safety condition; this can be informative even if the account is circular, i.e. if the relevant notion of safety is itself understood in terms of knowledge (Watzl 2017, 66). More generally, "Someone who accepts that there is an informative but non-reductive account of some *F* thus normally will say something about either the internal structure of *F*s or how being an *F* is related to some other phenomena" (Watzl 2017, 66–67). In the case of grounding worldly facts, I think we can do both of these things.

If a worldly fact Fa is fully grounded in other facts, then Fa obtains because the other facts obtain (at a specific time, in a specific world).[20] For a fact to obtain just is for its constituents to exist and to be unified in a certain way. Therefore, for the grounds of Fa to explain the obtaining of that fact is at the very least a good reason to accept that those grounds explain the unity of a and *F*ness. So this conception of grounding suggests a straightforward answer to the unity problem, at least as it concerns grounded facts: the constituents of grounded facts are unified by their grounds.[21]

While I think this is the correct explanation of the unity of grounded facts, it is reasonable to ask for more detail: in particular, *how* do the grounds of Fa unify its constituents? Again, it may not be possible to provide a non-circular answer to this question. But we can add more detail by considering that,

[19] Though for a recent proposal see Correia and Skiles (2019).

[20] This is a non-causal sense of "because" that tracks grounding relations. The indexing to worlds and times is adapted from Skiles (2015), 719. If Fa is on this occasion grounded by, e.g. Gb and Hc, it is possible that in other circumstances it could have been grounded by different facts. In what follows I shall omit this indexing.

[21] Strictly speaking, each grounded fact would be unified by its immediate grounds (on the distinction between mediate and immediate grounding see Fine 2012, 50–51). In what follows I shall omit this qualification.

plausibly, worldly facts can be grounded in different ways, depending on their constituents and the constituents of their grounds. In what follows I adopt the following hypothesis: for each instance of grounding, one or more of the constituents of the grounded fact stand in some specific ontological relation or relations to one or more of the constituents of each of the grounds. Which ontological relations obtain will depend on the constituents of each of the facts. For instance, the properties which help to make up the grounds may be determinates of a determinable property helping to make up the grounded fact (e.g. the fact that *a* is red is grounded in the fact that *a* is scarlet). Or the grounded fact may involve a property-bearer which is composed of the property-bearers in its grounds. Examples of this include many facts about functions (e.g. the fact that a computer is running a specific programme is grounded in facts about different sub-systems of the computer).[22]

The ontological relations holding between these constituents will determine how exactly the grounds unify the constituents of the grounded fact. For example, suppose that the fact that *a* is red is grounded in the fact that *a* is scarlet. In this case, each constituent of the grounded fact stands in a specific ontological relation to a constituent of the ground: *a* is identical with itself, and the property *being scarlet* is a determinate of the property *being red*. The determinable-determinate relation is such that, necessarily, any entity instantiating a determinate property instantiates its determinable.[23] Therefore, if the fact that *a* is scarlet obtains, this will automatically unify *a* and the

[22] These specific ontological relations are very similar to what Kelly Trogdon terms *grounding mechanisms*, "determination relations of a certain sort holding between constituents of grounding facts and constituents of the facts they ground" (2018, 1290). They are also similar to what Tobias Wilsch terms *linking principles*, principles which, roughly speaking, determine which objects and properties combine to form facts (2015, 3302–4) (thanks to an anonymous referee for suggesting these sources). For more detailed discussion of the different kinds of ontological relations which can hold between the constituents of different facts, see (O'Conaill ms). I do not regard these specific kinds of ontological relation (determinate-determinable, composition, etc.) as themselves kinds of grounding, i.e. "small-g" grounding relations (Wilson 2014, 540). Nor do I accept Wilson's criticisms of "big-G" grounding, though I cannot discuss the matter here; but see e.g. Cameron (2016); Schaffer (2016b). I should also add that other conceptions of how grounding works are available (e.g. Schaffer 2016a). I am not claiming that the account I shall sketch in the main text is how grounding *should* be understood; I am merely claiming that grounding *can* be understood in this way, and that doing so allows us to see how the grounds of a fact unify it.

[23] As Paul Audi puts it, "Anything maroon is red, and indeed, anything maroon is red *in virtue of* being maroon. So it seems that it is the natures of these properties that are responsible for the grounding relation's obtaining" (2012, 693).

property *being red* in the fact that *a* is red. In this way, the unity of *a* and the property *being red* is explained by the obtaining of the fact that *a* is scarlet.

Let us consider a slightly more complicated example: suppose there is a tower, *a*, which is exactly one metre tall and which consists of ten bricks piled on top of each other. Suppose also that the fact that *a* is one metre tall is grounded in facts about the height of each of the bricks which compose it and facts about how these bricks are arranged (i.e. they stand on top of each other). Here *F*ness is the property *being exactly one metre tall*, *G*ness is the property *being exactly ten centimetres tall*, and *H* is the relation *standing on top of each other*.

Again each constituent of the grounded fact stands in a specific ontological relation to a constituent of each of its grounds. First, the bricks together *compose a*.[24] Second, the height of the bricks, when the bricks stand in *H*, will *sum* to one metre. Here, the ontological relation holds between the property *being exactly ten centimetres tall* and the property *being exactly one metre tall* (one might say that instances of the first property, i.e. instances of *G*ness, are apt to sum together to form an instance of *F*ness when a certain number of bearers of the instances of *G*ness are suitably arranged).

So the full grounds of *Fa* (the fact that *a* is exactly one metre tall) will include ten facts of the form "*Gb*", "*Gc*", etc. (i.e. each brick is ten centimetres in height), plus a collective fact of the form "*b*, *c*, etc. are together *H*" (i.e. the bricks are stacked on top of each other). When these facts all obtain together, a fact of the form *Fa* will obtain (something which is composed of the ten bricks will be exactly one metre in height). That is, the property *F*ness and *a* (the tower composed of these ten bricks) will each exist and will be unified in the fact *Fa*.[25] Again, it should be clear how the ontological relations holding between *a* and the bricks, and between *F*ness and *G*ness and *H*ness, help to explain how the grounds of *Fa* can unify its constituents.

Each grounded fact is thus unified by something external, i.e. something identical neither with the fact itself nor with any of its constituents. This is a version of Betti's option (C). The precise details of how the constituents of

[24] Of course, it is a difficult question as to when one entity is composed by other entities, but I set this issue aside here. What matters for present purposes is that *a* is composed of the ten bricks.

[25] If "a" is a singular term then it may be objected that the obtaining of all the grounds does not suffice to ground the specific fact *F*a, because of the possibility of Ship of Theseus-style examples (Skiles 2015, 721–23). This is an important issue but I shall not address it here: for present purposes what matters is that the obtaining of the grounds explains why *F*ness is unified with whatever it is which the bricks together compose.

different grounded facts are unified remain to be worked out, but the outline of the approach is clear: examine the constituents of the grounded fact and the constituents of its grounds, and work out which ontological relations hold between them.[26]

It might be objected that this proposal begs the question: it can only work if a and Fness are already unified. Suppose that Fa is grounded in the fact that b is G (Gb). It seems clear that a's being F is a logical precondition for Fa to stand in any relation. Therefore, in order for Gb to ground Fa, a must already be F (i.e. a and Fness must be unified). Far from unifying a and Fness, any grounding relation in which Fa stands requires that it already be unified.[27]

An initial worry with this objection is that it threatens to prove too much. For with very little modification, it can be deployed against any proposed explanation of the existence of any entity whatsoever (where "existence" includes e.g. a fact's obtaining, an event's occurring, etc.). Suppose we want to explain the existence of some entity x, and we appeal to a different entity y; we say that y's existing, or something else about y, explains x's existing

[26] Could there be instances of grounding which do not feature specific ontological relations holding between constituents of the grounded fact and its grounds? These would be instances of what Trogdon terms *bare grounding*, "grounding relations that aren't instances of grounding mechanisms" (2018, 1295). Trogdon mentions as possible examples cases of logical or conceptual grounding, e.g. the fact that a is F and the fact that b is G together ground the conjunctive fact that a is F and b is G. I shall not address these examples in detail, but the following strategy is worth noting. On the compositional conception, a fact is composed of non-factual entities (properties, relations and property-bearers). Take the (proposed) conjunctive fact that a is F and b is G. What are the entities from which it is composed? It might be thought that this fact includes a conjunctive property, e.g. the property *x being F* $\wedge$ *y being G*. I am sceptical that there is such a property, but if it exists then it is plausible that the following is essentially true of it: it is instantiated iff some entity x is F and some entity y is G. In that case, the grounding of the conjunctive fact is not bare, since a specific ontological relation holds between a constituent of the grounded fact (the conjunctive property) and constituents of its grounds (i.e. Fness and Gness). On the other hand, it might be denied that the conjunctive fact includes a conjunctive property: on this view, the conjunctive fact is composed by a, b, Fness and Gness, arranged in a specific way. In that case, this proposed conjunctive fact seems to be an aggregate of the two facts which supposedly ground it. Given the compositional conception, it is not at all clear that any mere aggregate of facts should itself be counted as a *fact*. Rather,it is a collection of distinct facts. Any collection of facts can itself be treated as a fact, but this would be to use a different conception of facts, on which facts are logical or conceptual rather than worldly entities. I am not suggesting that facts understood in this way should not be posited,just that they are not the kind of facts which the problem of unity concerns.

[27] This objection was suggested to me by certain passages of Vallicella's (2000, 243, 254). However, it is not clear that he is putting forward this exact argument in these passages. Thanks also to an anonymous referee for pressing me to develop my response to this objection.

doi: 10.48106/dial.v74.i1.05

(e.g. y might be an event which causes x). For this explanation to be correct, it is necessary that x exists (if x did not exist, then its existence would not be explained). So the proposed explanation works only if x already exists. In this way, it turns out that any proposed explanation of the existence of any entity will be circular. But this is surely not so.[28]

One issue here is with the word "already": it might be objected that if x and y are events, and if effects occur after their causes, x could not already have occurred for its occurring to be explained by y. But I take it that in the objection to my proposal, the word "already" does not indicate temporal priority, but a logical precondition: for Gb to ground Fa logically requires that Fa obtains. When the term "already" is understood in this sense, y's causing x logically requires that x occurs, just as Gb's grounding Fa logically requires that Fa obtains.

This gives us reason to think *that* this objection has gone wrong. As to *how* it goes wrong, the answer lies in distinguishing *explanatory* considerations from *modal* considerations (which include what is logically or metaphysically necessary for something to exist). The basic point is this: for x to modally depend on y (so that x cannot exist unless y exists) does not preclude that x can itself explain (or help to explain) y's existence.

A couple of examples from the literature can help to clarify this point. On a widely accepted view of sets, the singleton set containing Socrates exists iff Socrates exists. But the existence of the set is widely thought to be explained (at least in part) by the existence of Socrates (see e.g. Schaffer 2016a, 53).

A second example is the Euthyphro dilemma (which has been discussed in the grounding literature—see e.g. Raven 2012, 692–93). Whichever way one responds to the dilemma, the "because" statement is true iff both the gods will that p and p is good. But this modal dependence does not rule out either explanation one might offer (e.g. that p is good because the gods will that p, or that the gods will that p because p is good).

How does this apply to my proposed explanation? It is true that Gb's grounding Fa modally depends on (logically requires) that Fa obtains. But this, I suggest, is not an explanatory dependence; we are not obliged to say that Gb's grounding Fa is explained, even in part, by Fa's obtaining. And, as per the examples outlined above, the modal dependence of a proposed explanation

[28] Vallicella argues that event causation cannot be causation of existence, precisely since both the cause and its effect must occur for a causal relation to hold between them (2002, 27). But this seems false, at least for instances of what Ned Hall terms productive causation, when an event "helps to *generate* or *bring about* or *produce* another event" (2004, 225).

on its explanans is not by itself sufficient to generate an explanatory circle. Indeed, the modal dependence of the explanation of *Fa*'s obtaining simply reflects the sufficiency of the proposed explanation: if the obtaining of *Fa*'s grounds are sufficient to explain the unity of *Fa*, then *Fa* must obtain for the explanation to be correct.

5 The Vicious Regress Argument

I have argued that grounding can account for the unity of facts which have grounds. This invites the questions of whether there are ungrounded facts, and if so what could account for their unity. There are two options to consider here:

(1) There are no ungrounded facts, and every fact is unified by its grounds;
(2) There are ungrounded facts, which are not unified in the way in which grounded facts are unified.[29]

I shall consider (1) in this section, and (2) in section 6.

In scenario (1), there are no facts such that they are not fully grounded in some other facts. Facts can form chains of grounding (a collection of facts where any two members of this collection stand in grounding relations to each other). In scenario (1), each fact will stand in an infinite descending chain or chains of grounds.

Whether such chains are possible and whether they could solve the unity problem are contentious issues. The main reason for thinking that such chains are impossible is that they seem to give rise to a vicious regress.[30] The criteria for deciding when regresses are vicious have been subject to extensive debate (Clark 1988; Nolan 2001; Maurin 2007; Wieland 2013). The criterion which seems most relevant in the present context is what Wieland terms the Failure Schema (2013, 99). This schema is summarised by Simon Blackburn: "A strategy gives rise to a vicious regress if whatever problem it was designed to solve remains as much in need of the same treatment after its use as before" (2005, 313). Examples of such strategies include the homunculus regress and the tower of turtles. In each case, a certain problem must be solved in order that something can be the case; an entity is posited in order to solve this

[29] As we shall see, there is a third option: there are no ungrounded facts, but chains of grounding terminate in entities which are not themselves facts. I shall briefly consider this option in section 6.

[30] This objection is raised by Betti concerning a version of option (B), the idea that *a* and *F*ness are unified by the presence of a further constituent such as a relation of instantiation (2015, 56–57).

doi: 10.48106/dial.v74.i1.05

problem; but the positing of this entity creates a problem of exactly the same kind as the problem the entity was posited to solve.

This suggests the following objection to scenario (1): the initial problem was how to explain the unity of some fact; in order to explain the unity of this fact, we posited grounds; but these grounds are facts each of which raises an explanatory demand of exactly the same kind as that which we initially faced. To respond to this further explanatory demand by positing further grounds would simply be to generate further problems of the same kind, and so on. What makes this regress vicious is that it makes no progress on the original question (or any progress it makes at any step in the regress is immediately cancelled out). This is arguably what goes wrong in the homonculous and turtle cases.

While this is a problem for (1), it is not clear that it is decisive. The original question was how to explain the unity of some specific fact, Fa. By appealing to the grounds of Fa, this question is answered. Granted, the answer generates a problem of the same kind: but ex hypothesi, for each new fact introduced, we will be able to appeal to *its* grounds to explain its unity. Given the scenario outlined in (1), there will never be a fact which lacks grounds, so the problem of unity can be answered for every fact posited.[31]

It may be objected that this strategy is vicious insofar as it explains the unity of facts by assuming the very possibility of any fact being unified. This objection, or something like it, crops up occasionally in the literature:

> Even if, assuming there can be facts, facts may depend on each other in never-ending chains of dependence, postulating such chains of dependence does not help when it comes to the very possibility of there being facts to begin with. (Eklund 2019, 1228)

But in the context of discussing the problem of unity, this objection seems to change the subject.[32] We began by asking a local question (what explains the unity of some specific fact or facts); now we are considering a global question,

[31] This is an important difference between the regress of facts to which the truth of (1) would commit us and what Eklund terms the *constitution regress* (2019, 1227–29). The constitution regress very plausibly is vicious, because no step in this regress explains the fact with which the regress started. So rather than a different problem of the same type occurring at each step, as is the case with the regress generated by accepting (1), in the constitution regress the very problem we started with is never solved.

[32] This is not to suggest that Eklund himself is guilty of this. In the section where the passage I quoted appears, he is discussing the constitution regress, which is different to the regress which the truth of (1) would set up (see fn. 31 above).

what is required for the possibility of any fact whatsoever.[33] But the chain of grounds was introduced to answer a series of local questions, e.g. why each specific fact is unified: "To claim that an infinite regress is vicious because it doesn't allow us to answer the global question is to have accused it of having failed to carry out a task it was not designed to complete" (Bliss 2013, 408).[34]

There is more to be said on these specific points and on other ways of characterizing regresses as vicious, but I shall not explore these issues here. My provisional conclusion is that while the regress argument is a problem for the proponent of (1), it is not clearly decisive: that is, it is not obvious that the regress argument renders (1) untenable. That said, it is worth asking how the unity problem might be solved for ungrounded facts.

6 Essentially Unified Facts

I propose that at least some ungrounded facts are *essentially unified*.[35] These facts are such that the properties which make them up are essential to their property-bearers, and so the property bearer could not exist without instantiating that property.[36] For instance, suppose that the property *being negatively*

[33] Vallicella makes a similar point: the problem of unity "does not concern the nature of fact-unity in general, but the existence of fact-unity in particular cases" (2000, 242).

[34] Orilia offers a solution to the unity problem which also appeals to an infinity of facts, and he responds to the threat of a vicious regress in a similar way (2006, 233). I shall not discuss Orilia's position in detail, but it is worth mentioning two differences between it and my own. The first is that the facts to which Orilia appeals, facts which contain instances of an exemplification relation (what I have termed "instantiation") are ad hoc; they are posited solely in order to solve the unity problem, without any independent reason to accept them. In contrast, the conception of grounding I have outlined can limit itself to facts which are relatively uncontroversial. Of course there are controversies surrounding grounding claims, but in general such claims are introduced as a way of ordering facts which we have independent reasons to accept. Second, Orilia's view commits one to a necessarily infinite regress and a necessary infinity of facts given the obtaining of any fact (2006, 230). Even if such a regress is metaphysically possible, considerations of parsimony would favour not positing an infinity of facts if it can be avoided. As we shall see, the grounding response to the unity problem does not by itself commit one to positing an infinity of facts.

[35] In the next section I shall consider ungrounded facts which are not essentially unified. It is worth noting that there may be grounded facts which are essentially unified. The unity of these facts would be over-determined. But it is plausible that the vast majority of grounded facts are not essentially unified. Indeed, as noted in section 2, it is often assumed in the literature on the problem of unity that the constituents of a fact could all exist without together composing that fact.

[36] The relevant notion of "essence" is the non-modal conception made familiar by Kit Fine. In particular, I have in mind Fine's notion of *constitutive essence* (1995, 276). Note also that I am not suggesting that essentially unified facts are grounded in essential facts about their constituents.

doi: 10.48106/dial.v74.i1.05

charged is essential to any electron. In that case, a specific election *e* could not exist without instantiating this property, i.e. without the fact that *e* is negatively charged obtaining. More generally, the thesis that some facts are essentially unified entails rejecting the following assumption: "Even if *a* and *F*ness cannot exist except in some state of affairs or other, there is nothing in the nature of *a* and nothing in the nature of *F*ness to require that they combine with each other to form *a's being F*" (Vallicella 2000, 238).

Essential unity can be usefully compared with bearer-specific properties (see section 3). A bearer-specific property is such that if instantiated, it is necessarily instantiated by some specific entity or entities. In essential unity, it is the property bearer which is such that if it exists, it necessarily instantiates a certain property. In each case, one of the constituents of a fact is such that its existence (or instantiation) requires that it combine with the other constituent or constituents.

There are two ways in which essential unity might be said to explain the unity of some ungrounded facts. One way to understand facts which are essentially unified is that there is no ontological difference between them and the aggregate of their constituents. Since there is no difference, there is no need for any explanation of this difference, and the unity problem dissolves. This reasoning mirrors Betti's explanation for why bearer-specific properties dissolve the problem: "If *R* is relata-specific, and thus it is in the nature of *R* to relate *a* and *b*, then *aRb* exists as soon as *R* exists. So, there is simply no difference between $a + R + b$ and *aRb*" (2015, 92).

This way of dissolving the unity problem might be thought to face the following objection: it removes any motivation to think of essentially unified facts as genuinely *facts*, as entities over and above their constituents. This is Betti's own conclusion: bearer-specific properties not only dissolve the problem of unity, but also remove the need for the ontological category of compositional facts (2015, 106).[37]

Even if this point is correct, it is compatible with a way of solving (rather than dissolving) the unity problem. Consider the aggregate of entities which we wrongly took to form an essentially unified fact. Let us term this aggregate a *quasi-fact*. Each quasi-fact will include a number of property-bearers and properties or relations such that each property or relation is essential to the property-bearers. We can then adjust the notion of grounding as follows: a fact

Rather, the proposed explanation of the unity of essentially unified facts is an *essentialist explanation* (Glazier 2017, 2872).

[37] This conclusion is questioned by Vallicella (2016a, 236).

can be grounded by another fact, or by a quasi-fact, or by some combination of facts and quasi-facts. Every grounded fact will be unified by its grounds (either facts or quasi-facts); and since there is no difference between a quasi-facts and the aggregate of its components, the problem of unity will not arise for quasi-facts.

That said, I am drawn towards the other way in which essential unity can solve the problem. First, I think there *is* an ontological difference between essentially unified facts and the aggregate of their constituents, even though the existence of the constituents suffices for these facts to obtain. The aggregate of *e* and *being negatively charged* just is *e* and this property considered together. It involves nothing other than these two entities; there is nothing more to the aggregate's existence than the existence of these entities. In contrast, the fact that *e* is negatively charged involves the instantiation by *e* of this property; that is, the fact involves these entities being arranged in a specific way. As it happens, these entities are such that when they exist they are automatically arranged in this way. But this does not entail that there is no ontological difference here. The fact still involves a way of being unified which the aggregate does not.[38]

The problem of unity for an essentially unified fact is solved by some of its own constituents. The problem of unity, recall, is the problem of explaining why, given that each of a fact's constituents exist, they are combined in the way characteristic of this fact. In an essentially unified fact, some of its constituents are such that necessarily, if they exist they must instantiate certain properties or stand in certain relations. Therefore, given that each of the fact's constituents exist, it is necessary that they are unified so as to form this fact. For instance, it is in virtue of the essence of *e* that it is unified with the property *being negatively charged*.

This account is in effect a version of Betti's option (A): the explanation of why the constituents are unified lies in the essence of one of the constituents itself. Betti herself rejects this option. Since the unity problem presupposes that

[38] It might be objected that the difference I am positing between essentially unified facts and the aggregates of their constituents appeals to non-mereological composition, and so begs the question in favour of facts. But it is important to be clear on what is at issue here. As was argued in section 2, we can distinguish between explaining *what it is* for constituents to form a fact (e.g. clarifying the distinctive way in which the constituents must be unified so as to form a fact), and explaining *why* a fact obtains given that its constituents exist. The discussion in this paragraph of the main text concerns the first of these issues, not the second. And as mentioned earlier, in addressing the first of these issues it is legitimate to appeal to non-mereological composition, e.g. to instantiation.

doi: 10.48106/dial.v74.i1.05

the constituents in the aggregate are numerically identical to the constituents of the fact, it seems impossible for the difference between the fact and the aggregate to be explained by reference to any of these constituents (2015, 56). But as argued above, the difference between the fact and the aggregate just is the non-mereological unity of the constituents in the fact, and this unity is explained by the essence of the property bearer.

7 Brutely Unified Facts

I have outlined an account of the unity of grounded facts, and of ungrounded facts where the properties are essential to the property-bearers. However, it is plausible that if ungrounded facts obtain, not all of them are essentially unified: for instance, the fact that a fundamental particle stands in a certain spatiotemporal location (Dasgupta 2014, 579), or the fact that a simple entity *a* is *F* (where *F*ness is e.g. a maximally determinate shade of colour). That is, in addition to grounded facts and facts which are ungrounded and essentially unified, there is at least logical space for a third category of facts, facts such that there is nothing in virtue of which their components are unified. Let us term these *brutely unified* facts.[39]

The possibility of brutely unified facts raises two issues for the position I wish to defend. The first is the general question of whether such facts are possible; the second is whether allowing for such facts weakens my position compared to other responses to the problem of unity. I shall consider the second issue in the next section; for the remainder of this section, I shall discuss the first.

An assumption made by some in the literature is that if the unity of a (supposed) fact cannot be explained, then we have reason to think that this fact cannot exist (e.g. Vallicella 2000, 248; Betti 2015, 103; Maurin 2015, 201). I do not share this assumption. I think it is true of any fact that we can ask why it obtains or why its constituents are arranged as they are, but if it turns out that a positive answer cannot be provided for certain facts, this does not in itself give us reason to doubt that such facts obtain.

[39] Thanks to Francesco Spada and to an anonymous reviewer for drawing my attention to the possibility of such facts. It may be that that there are facts which do not belong to any of the three categories I have distinguished (i.e. facts which are ungrounded and not essentially unified, but which are not brutely unified either). That said, it is not obvious what would unify such facts, and so I shall set aside this possible further category.

Vallicella offers three arguments against the possibility of brutely unified facts. First, he claims that the view that there are such facts leads to "the contradiction that a fact both is and is not a whole of parts" (2002, 20), i.e. an aggregate of its constituents. The argument is as follows:

> A fact *is* a whole of parts in that there is nothing 'in' it but its parts. For a fact is a complex, and a complex is composed of constituents. Analysis of *aRb* can yield nothing beyond *a*, R, and *b*. A fact is *not* a whole of parts in that the existence of the parts does not entail the existence of the whole. Thus a fact is more than the mere sum of its parts. This 'more' is something real, and yet it cannot be, or be grounded in, any further constituent of the fact. [...] it seems to be a contradiction to say of a whole that it is an entity in addition to its parts when it is composed of them. (Vallicella 2002, 20)

The problem with this argument is that it equivocates on the first claim, that "there is nothing 'in' a fact but its parts" (i.e. its constituents). This claim could be interpreted as meaning "a fact has no constituent other than its parts, e.g. *a*, *R* and *b*." But it could also be interpreted as meaning "a fact is reducible to or nothing over and above its parts", where this would entail, among other things, that a fact obtains if its parts all exist. Interpreted in the first way, the first claim would be accepted by the proponent of the compositional conception; but interpreted in this way, the first claim does not lead to a contradiction with the second claim, that a fact is more than the aggregate of its parts. Interpreted in the second way, the first claim would lead to a contradiction with the second claim; but interpreted in this way, the first claim would not be accepted by the proponent of the compositional conception. Thus, Vallicella's first argument is either a non-sequitur or it begs the question against the proponent of the compositional conception (by assuming that a fact is nothing over and above its constituents).

Vallicella's second argument starts with two facts, *Fa* and *Gb*, which ex hypothesi have no constituent in common. Valicella notes "each fact is precisely a *fact*, which suggests that they have the universal *being a fact* (facthood) in common" (2002, 21–22). But if they have no constituent in common "then facthood is not a common constituent; how then do we explain the circumstance that they are *both* facts? How do we explain the common categorical status?" (2002, 22). Since Vallicella thinks it cannot be a brute fact that both are facts, nor can either of these facts itself be a brutely unified fact.

doi: 10.48106/dial.v74.i11.05

Given a sparse conception of properties (see fn. 2), it is not at all clear that there is any good reason to accept that there is a property *being a fact*, or that there are facts of the form: Fa is a fact. There are certainly *truths* of the form "Fa is a fact." What explains their being true is precisely the factual ontological structure of Fa, i.e, a's instantiating Fness.

But assume that there is such a property as the property *being a fact*. Presumably this property will be instantiated by all and only facts, and therefore will be something which all and only facts have in common. But why assume that it must be a *constituent* of every fact? On the contrary, it seems obviously mistaken to assume that a fact such as the apple's being red must be partly composed of the property *being a fact*. The proponent of the compositional conception of facts has no need to assume that properties are constituents of the entities which instantiate them.[40] And this is true even if the entities which instantiate properties are themselves facts.

Vallicella's third argument is as follows:

> (i) if the difference between a fact and its constituents is a brute fact, then it is possible that two facts share all constituents. (ii) But it is not possible that two facts share all constituents. Therefore, (iii) the difference between a fact and its constituents is not a brute fact; it has an ontological ground. (2002, 22)

The proponent of the compositional conception will accept neither (i) nor (ii). As regards (i), if a fact is brutely unified then the difference between this fact and the aggregate of its constituents is simply that the fact is a *fact*, that is, it consists of the constituents arranged in a certain way. Vallicella claims,

> if a fact's being a fact is what distinguishes it from its constituents, then a fact's being a fact is what ultimately distinguishes it from other facts even if there also happens to be a difference in constituents. Each fact, just in virtue of its being a fact, differs from every other fact. (2002, 23)

But on the compositional conception, this is false. That Fa is a fact, i.e. a complex of constituents arranged in a specific way, is not what distinguishes it from Gb (which, after all, is just as much a fact). What distinguishes the

[40] This is a well-known view of properties, (e.g. Armstrong 1989, 77), but it is not one which I accept, and more importantly it is not one to which the proponent of the compositional conception is committed.

two is precisely that they have different constituents. More generally, what distinguishes each fact from the aggregate of its constituents is different to what distinguishes each fact from any other fact (the latter is given by the identity-conditions of facts outlined in section 2).

As regards (ii), it is widely thought possible for certain distinct facts to share the same constituents (as with facts including asymmetric relations—see section 2 above). Vallicella dismisses this response as question-begging against (ii), but this claim is highly doubtful.[41] The general point behind rejecting (ii) is that a fact is composed of constituents unified in some specific way (e.g. a particular instantiating a universal, or two particulars being related in a certain order), and that in certain cases the same constituents can be unified in more than one way, giving rise to distinct facts.[42]

8 Comparing Different Accounts of Unity

The account I offer of the unity of facts has two significant limitations compared to alternatives such as those offered by Betti or Vallicella. First, it is a disunified account, proposing different answers to the problem of unity for different facts (e.g. grounded versus ungrounded); in contrast, Betti and Vallicella each offer a unified account.[43] Second, the account I propose is limited in scope, if it is accepted that there can be ungrounded facts which are not essentially unified (this is the second issue mentioned at the start of the previous section). My account does not provide a positive answer to the question of what unifies these facts, whereas the positions defended by Betti and by Vallicella promise to do so.

Each of these limitations is important, but I do not think that they are decisive. While all facts share the ontological structure described in section 2, there are important differences between, e.g. facts which are grounded and

[41] In a footnote, Vallicella clarifies that what is question-begging is to appeal to aRb and bRa's being distinct facts in support of the claim that facts obtain (2002, 24, n. 51). This specific dialectical move might beg the question, but what I am discussing in this paragraph in the main text is not whether facts obtain, but whether we should accept claim (ii), that it is not possible that distinct facts share all constituents.

[42] An alternative counterexample to (ii) appeals to a plausible condition on the critieria of identity for facts, that these criteria are time-indexed (see section 2). If a instantiates Fness at t 1, ceases to instantiate it at t 2, and instantiates it again at t 3, it seems perfectly reasonable to say that there are two distinct facts composed of a and Fness; one obtained at t 1 and ceased to obtain at t 2, the other obtained at t 3.

[43] Though this may not be true of Vallicella's proposal (see e.g. 2000, 258n45).

facts which are ungrounded, and between facts which are essentially unified and facts which are not. Once these differences are made clear, the cost of a disunified account is diminished; or, to put it another way, once the differences between facts are made clear, it is less obvious that we should expect to find a single account which explains the unity of each fact.

Furthermore, it seems to me to be a mistake to assume from the outset that the problem of unity can be solved for every fact. Once we acknowledge that there are different types of fact, the possibility is opened that there are facts for which no positive answer can be given to the question "Why does this fact obtain?". Granted, it is methodologically preferable to be able to explain the unity of each fact. That is, all things being equal, we ought to prefer a theory which allows for a positive answer to each question of this form to one which does not. But are all things equal?

I contend they are not; the account I offer has advantages over the main alternatives. My account relies on grounding and on certain properties being essential to their bearers. While grounding and essential properties are by no means uncontroversial, each is a relatively familiar and well-developed idea, and there are reasons for accepting each idea which are independent of any role they might play with regard to the unity of facts. In contrast, the accounts offered by both Vallicella and Betti rely on ontological posits which have not been widely discussed or systematically clarified, each of which is ad hoc, and each of which faces independent considerations against it.

To develop this point, consider first some of the problems facing Betti's ontological posit, bearer-specific properties. First, on Betti's view the identities of properties are implausibly fragile. For instance, consider two entities, *a* and *b*, each of which instantiates the property *being the determinate shade of red x*. Now consider a counterfactual situation where *a* does not exist. On Betti's view, in this counterfactual situation *b* would not instantiate the property *being the determinate shade of red x*, since that property can only exist if it is instantiated by *a*. Rather, in that situation *b* would instantiate the distinct (though presumably qualitatively identical) property, *being the determinate shade of red x∗*. This is surely the wrong result; it seems to me that I can understand what it would be for that very property, *being the determinate shade of red x*, to exist and to be instantiated in a situation where *a* did not exist.[44]

[44] This point is even clearer if one accepts that there are determinate quantitative properties, e.g. *being the determinate length x*.

Second, Betti's position entails that our knowledge of what properties are is constrained to an implausible degree. Her view requires that one can only be said to know which property is in fact instantiated if one knows each and every entity which bears it (e.g. we can only know we are dealing with the property *being the determinate shade of red x* and not the property *being the determinate shade of red x∗* if we know that *a* exists). Again, this seems implausible.[45]

Vallicella appeals to a single entity, U, to unify all contingently unified facts. He assumes that U cannot necessarily unify these facts, as this would mean they were not contingently unified. Therefore, U must contingently unify them. As Vallicella puts it,

> U must have the power of contingent self-determination: it must have the power to contingently determine itself as operating upon its operand. In other words, if U is the ground of the contingent unity of a fact's constituents, then U contingently *grounds its grounding* of the unity of the fact's constituents. (2000, 255)

Vallicella's model for this contingent self-determination is our own free will. Specifically, he suggests that the contents of our thoughts are unified in conscious acts, as when one judges that *a* is *F* (2000, 255). Indeed, the entities Vallicella proposes as candidates to play the role of U are God and transcendental consciousness (2000, 252–53).

There is an ambiguity in this account as it stands. Consider one's unifying the contents of a specific thought (say, that *a* is *F*) in an act of judging. What is the unifier here? One's ability to judge is not sufficient to explain the unity of the contents of this thought, since one could exercise this ability without thinking that very thought. Alternatively, the unifier could be a particular exercise of this capacity, e.g. a particular act of judging (that *a* is *F*). But this answer immediately leads to a further problem. A particular act of judging will involve either oneself standing in a relationship to something, e.g. the contents of one's act of judging, or it will involve one instantiating a specific property, e.g. the property *judging that a is F*. Either of these will involve the obtaining of a fact, and furthermore this fact will be contingent. What explains the obtaining of such facts? (Note that the answer cannot be "one's

[45] Bearer-specific properties would not be so controversial if they were assumed to be tropes; however, as mentioned in section 3 Betti rejects this assumption. Furthermore, ruling out any universal properties or relations brings its own problems (see Armstrong 1989; Lowe 2006).

doi: 10.48106/dial.v74.i1.05

power to freely judge"—what is being asked for is an explanation of one's exercising this power on a specific occasion.) Similarly, U may have the power of contingent self-determination, but its having this power is not sufficient to explain the unity of each fact; what is also needed is an explanation of why this power is exercised as it is.[46]

None of this is to suggest that the accounts offered by Betti or by Vallicella cannot work, or that their posits cannot be ultimately defended. But each of their accounts faces serious theoretical problems. The account I offer, though limited in important respects, relies on more familiar and well-established ontological ideas. For this reason, it deserves to be taken as seriously as any other proposed solution to the problem of unity.[*]

Donnchadh O'Conaill
University of Fribourg
donnchadh.oconaill@unifr.ch

References

Armstrong, David M. 1989. *Universals: An Opiniated Introduction*. Boulder, Colorado: Westview Press.

———. 1997. *A World of States of Affairs*. Cambridge: Cambridge University Press.

Audi, Paul. 2012. "Grounding: Toward a Theory of the *In-Virtue-Of* Relation." *The Journal of Philosophy* 109 (12): 685–711. doi:10.5840/jphil20121091232.

———. 2015. "Explanation and Explication." In *The Palgrave Handbook of Philosophical Methods*, edited by Christopher John Daly, 208–30. London: Palgrave Macmillan.

Bennett, Karen. 2013. "Having a Part Twice Over." *Australasian Journal of Philosophy* 91 (1): 83–103. doi:10.1080/00048402.2011.637936.

Betti, Arianna. 2015. *Against Facts*. Cambridge, Massachusetts: The MIT Press.

[46] Vallicella notes that if U is God, then on a standard conception God necessarily has His attributes, e.g. He is necessarily omniscient (2000, 258, n. 45). But God presumably does not necessarily have the property *judging that a is F*; nor does He necessarily stand in a unifying relation to the fact that a is F (and if He did, the fact that a is F would not be contingent).

[*] Thanks to Francesco Spada, Jani Hakkarainen, Antti Tiainen, Arianna Betti, Tuomas Tahko, Henrik Rydéhn, Michael O'Sullivan and a number of referees for their comments on various drafts of this paper. Thanks also to audiences at the Université de Neuchâtel philosophy colloquium (December 2016) and the Dynamis workshop at the University of Tampere (March 2017) for discussion. My work on this paper was supported by the Academy of Finland (Kulttuurin ja Yhteiskunnan Tutkimuksen Toimikunta, grant number 274715) and the Swiss National Science Foundation (grant number 166320).

Blackburn, Simon. 2005. “Regress.” In *The Oxford Dictionary of Philosophy*, edited by Simon Blackburn, 2nd ed., 313. Oxford: Oxford University Press.

Bliss, Ricki Leigh. 2013. “Viciousness and the Structure of Reality.” *Philosophical Studies* 166 (2): 399–418. doi:10.1007/s11098-012-0043-0.

Cameron, Ross P. 2016. “Do We Need Grounding?” *Inquiry* 59 (4): 382–97. doi:10.1080/0020174x.2015.1128848.

Clark, Romane. 1988. “Vicious Infinite Regress Arguments.” In *Philosophical Perspectives 2: Epistemology*, edited by James E. Tomberlin, 369–80. Oxford: Basil Blackwell Publishers.

Correia, Fabrice, and Alexander Skiles. 2019. “Grounding, Essence, and Identity.” *Philosophy and Phenomenological Research* 98 (3): 642–70. doi:10.1111/phpr.12468.

Dasgupta, Shamik. 2014. “The Possibility of Physicalism.” *The Journal of Philosophy* 111 (9–10): 557–92. doi:10.5840/jphil20141119/1037.

Dixon, T. Scott. 2018. “Upward Grounding.” *Philosophy and Phenomenological Research* 97 (1): 48–78. doi:10.1111/phpr.12366.

Dodd, Julian. 1999. “Farewell to States of Affairs.” *Australasian Journal of Philosophy* 77 (2): 146–60. doi:10.1080/00048409912348901.

Eklund, Matti. 2019. “Regress, Unity, Facts, and Propositions.” *Synthese* 196 (4): 1225–47. doi:10.1007/s11229-016-1155-4.

Fine, Kit. 1982. “First-Order Modal Theories III – Facts.” *Synthese* 53 (1): 43–122. doi:10.1007/BF00500112.

———. 1995. “Ontological Dependence.” *Proceedings of the Aristotelian Society* 95: 269–90. doi:10.1093/aristotelian/95.1.269.

———. 2012. “Guide to Ground.” In *Metaphysical Grounding. Understanding the Structure of Reality*, edited by Fabrice Correia and Benjamin Sebastian Schnieder, 37–80. Cambridge: Cambridge University Press.

Glazier, Martin. 2017. “Essentialist Explanation.” *Philosophical Studies* 174 (11): 2871–89. doi:10.1007/s11098-016-0815-z.

Hall, Ned. 2004. “Two Concepts of Causation.” In *Causation and Counterfactuals*, edited by John David Collins, Ned Hall, and Laurie A. Paul, 225–76. Cambridge, Massachusetts: The MIT Press.

Lewis, David. 1986. “Against Structural Universals.” *Australasian Journal of Philosophy* 64 (1): 25–46. doi:10.1080/00048408612342211.

Lowe, Edward Jonathan. 2006. *The Four-Category Ontology: A Metaphysical Foundation for Natural Science*. Oxford: Oxford University Press.

Maurin, Anna-Sofia. 2007. “Infinite Regress – Virtue or Vice?” In *Hommage à Wlodek. Philosophical Papers Dedicated to Wlodek Rabinowicz*, edited by Toni Rønnow-Rasmussen, Björn Petersson, Jonas Josefsson, and Dan Egonsson. Lund: Lunds Universitet, Filosofiska Instititonen.

———. 2015. "States of Affairs and the Relation Regress." In *The Problem of Universals in Contemporary Philosophy*, edited by Gabrielle Galluzzo and Michael J. Loux, 195–214. Cambridge: Cambridge University Press.

Meinertsen, Bo Rode. 2008. "A Relation as the Unifier of States of Affairs." *Dialectica* 62 (1): 1–19. doi:10.1111/j.1746-8361.2007.01127.x.

Nolan, Daniel Patrick. 2001. "What's Wrong with Infinite Regresses?" *Metaphilosophy* 32 (5): 523–38. doi:10.1111/1467-9973.00206.

O'Conaill, Donnchadh. ms. "Grounding and the Unity of Facts." Unpublished manuscript.

Orilia, Francesco. 2006. "States of Affairs: Bradley Vs. Meinong." In *Meinongian Issues in Contemporary Italian Philosophy*, edited by Venanzio Raspa, 213–38. Meinong Studies / Meinong Studien 2. Heusenstamm b. Frankfurt: Ontos Verlag.

Perovic, Katarina. 2016. "Neo-Armstrongian Defence of States of Affairs. A Reply to Vallicella (2016b)." *Metaphysica* 17 (2): 143–61. doi:10.1515/mp-2016-0010.

Raven, Michael J. 2012. "In Defence of Ground." *Australasian Journal of Philosophy* 90 (4): 687–701. doi:10.1080/00048402.2011.616900.

———. 2013. "Is Ground a Strict Partial Order?" *American Philosophical Quarterly* 50 (2): 193–202.

Rodríguez Pereyra, Gonzalo. 1998. "Searle's Correspondence Theory of Truth and the Slingshot." *The Philosophical Quarterly* 48 (193): 513–22. doi:10.1111/1467-9213.00119.

———. 2015. "Grounding Is Not a Strict Order." *The Journal of the American Philosophical Association* 1 (3): 517–34. doi:10.1017/apa.2014.22.

Schaffer, Jonathan. 2016a. "Ground Rules: Lessons from Wilson." In *Scientific Composition and Metaphysical Ground*, edited by Kenneth Aizawa and Carl Gillett, 143–70. New Directions in the Philosophy of Science. London: Palgrave Macmillan.

———. 2016b. "Grounding in the Image of Causation." *Philosophical Studies* 173 (1): 49–100. doi:10.1007/s11098-014-0438-1.

Skiles, Alexander. 2015. "Against Grounding Necessitarianism." *Erkenntnis* 80 (4): 717–51. doi:10.1007/s10670-014-9669-y.

Strawson, Peter Frederick. 1950. "Truth." *Proceedings of the Aristotelian Society, Supplementary Volume* 24: 129–56. doi:10.1093/aristoteliansupp/24.1.111.

Trogdon, Kelly. 2018. "Grounding-Mechanical Explanation." *Philosophical Studies* 175 (6): 1289–309. doi:10.1007/s11098-017-0911-8.

Vallicella, William F. 2000. "Three Conceptions of States of Affairs." *Noûs* 34 (2): 237–59. doi:10.1111/0029-4624.00209.

———. 2002. "Relations, Monism, and the Vindication of Bradley's Regress." *Dialectica* 56 (1): 3–36. doi:10.1111/j.1746-8361.2002.tb00227.x.

———. 2016a. "Facts: An Essay in Aporetics." In *Metaphysics and Scientific Realism. Essays in Honor of David Malet Armstrong*, edited by Francesco Federico Calemi, 105–32. EIDE – Foundations of Ontology 9. Berlin: Walter De Gruyter.

———. 2016b. “Review of Betti (2015).” *Metaphysica* 17 (2): 229–41. doi:10.1515/mp-2016-0017.

Watzl, Sebastian. 2017. *Structuring Mind: The Nature of Attention and How It Shapes Consciousness*. Oxford: Oxford University Press.

Wieland, Jan Willem. 2013. “Infinite Regress Arguments.” *Acta Analytica* 28 (1): 95–109. doi:10.1007/s12136-012-0165-1.

Wieland, Jan Willem, and Arianna Betti. 2008. “Relata-Specific Relations: A Response to Vallicella (2002).” *Dialectica* 62 (4): 509–24. doi:10.1111/j.1746-8361.2008.01167.x.

Wilsch, Tobias. 2015. “The Nomological Account of Ground.” *Philosophical Studies* 172 (12): 3293–3312. doi:10.1007/s11098-015-0470-9.

Wilson, Jessica M. 2014. “No Work for a Theory of Grounding.” *Inquiry* 57 (5–6): 535–79. doi:10.1080/0020174X.2014.907542.

doi: 10.48106/dial.v74.i1.05

Strevens's Counterexample to Lewis's "Causation as Influence", and Degrees of Causation

Joshua Goh

Sungho Choi has criticised Michael Strevens's counterexample to David Lewis's final theory of "token" causation, causation as "influence." I argue that, even if Choi's points are correct, Strevens's counterexample remains useful in revealing a shortcoming of Lewis's theory. This shortcoming is that Lewis's theory does not properly account for *degrees* of causation. That is, even if Choi's points are correct, Lewis's theory does not capture an intuition we have about the *comparative* causal statuses of those events involved in Strevens's counterexample (we might, for example, intuit that Sylvie's ball-firing is *as much*/*more*/*less* a cause of the jar's shattering as/than is Bruno's ball-firing).

Sungho Choi (2005, 106–13) has criticised Michael Strevens's (2003, 4–7, 11–17) counterexample to David Lewis's (2000) final theory of "token" causation, causation as "influence" (hereafter, "CaI"). I argue that, even if Choi's points are correct, Strevens's counterexample remains useful in revealing a shortcoming of CaI. This shortcoming is that CaI does not properly account for *degrees* of causation. This paper proceeds as follows. Section 1 articulates CaI. Section 2 articulates Strevens's counterexample to CaI, and Choi's criticism of Strevens's counterexample. Section 3 argues that, even if Choi's points are correct, CaI does not capture an intuition we have about the *comparative* causal statuses of those events involved in Strevens's counterexample (we might, for example, intuit that Sylvie's ball-firing is *as much*/*more*/*less* a cause of the jar's shattering as/than is Bruno's ball-firing).

1 CaI

CaI involves three ideas. The first idea is the “alteration” of an event. Consider this event *E*: the vase’s shattering. Lewis defines an “alteration” of *E* as “either a very fragile *version* of *E* or else a very fragile *alternative event* that is similar to *E*, but numerically different from *E*” (2000, 188, emphasis mine).

To elucidate, an event is considered “fragile” if we impose stringent conditions for its occurrence (if we say that any change in one of its details turns it into a numerically different event) (Lewis 2000, 185–86). One alteration of *E* is *E*’s *actual* alteration: exactly when and how the vase shattered. The other alterations of *E* are un-actualised (one example: the vase shattering one millisecond later, and into more pieces).

The second idea is “influence.” Let *C* and *E* be two single, distinct, actual events. Lewis holds that *C* “influences” *E* iff

> there is a substantial range C_1, C_2, ... of different not-too-distant alterations of *C* (including the actual alteration of *C*) and there is a range E_1, E_2, ... of alterations of *E*, at least some of which differ, such that if C_1 had occurred, E_1 would have occurred, and if C_2 had occurred, E_2 would have occurred, and so on. (Lewis 2000, 190)

Idea three concerns the relationship between influence and causation. According to Lewis, *C* is a cause of *E* iff *C* directly influences *E*, or there is a chain of stepwise influence (hereafter, “i-chain”) leading from *C* to *E* (that is, a sequence of (actual) events *C*, D_1, D_2, ..., D_n, *E*, such that *C* influences D_1, D_1 influences D_2, ..., $D_{(n-1)}$ influences D_n, and D_n influences *E*) (Lewis 2000, 191; see also Lewis 1973, 563).

Let’s observe CaI in action. Consider this scenario: Sylvie throws a rock at a vase. Beside her, Bruno laughs. Here, CaI delivers the intuitive result that Sylvie’s throw is a cause of the vase’s shattering, while Bruno’s laughter is not. This is because Sylvie’s throw has substantial direct influence on the vase’s shattering. That is, there are many different, not-too-distant alterations of Sylvie’s throw (e.g. her throwing one millisecond later/with slightly more force) upon which alterations in the vase’s shattering (i.e. the vase’s shattering one millisecond later/into more pieces) counterfactually depend. Bruno’s laughter, however, has no substantial direct influence on the vase’s shattering. *Maybe* one distant alteration of Bruno’s laughter is so infectious that it delays Sylvie’s throw (and hence, the vase’s shattering) by a second. Nevertheless, no not-too-distant alteration of Bruno's laughter appears to alter the vase's shattering.[1] Moreover, one cannot identify any I-CHAIN leading from Bruno's laughter to the vase's shattering.

2 Strevens's counterexample to CaI; Choi's criticism

Here is Strevens's counterexample to CaI (2003, 4–7, 11–17):

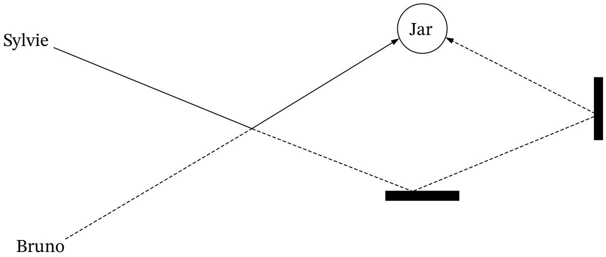

Figure 1: Solid line: actual trajectory of Sylvie's ball. Dotted line: actual trajectory of Bruno's ball.

> SCE. At time t_1, and using identical rifles, Sylvie and Bruno fire at a jar intrinsically identical, minute lead balls. Sylvie, who never misses, shoots so that her ball will ricochet two times prior to striking the jar. Bruno shoots directly at the jar. The balls, however, collide in mid-air at time t_c. Consequently, they *perfectly* exchange trajectories and spin (we thus take the motion of the balls to be that of two point particles; this admittedly requires something like a fortuitous gust of wind at t_c) (2003, 5, fn. 2). Stipulate moreover that the speeds of the two balls are always identical (and extremely high). Ultimately, Sylvie's ball shatters the jar, and Bruno's ricochets, then flies through thin air.[2]

[1] Unfortunately, Lewis is vague about what it takes for an alteration of an event to qualify as "not-too-distant." He says that, for some particular alteration of an event, whether or not we think it to be "not-too-distant" may be a matter of "mood" (2000, 197).

[2] Strevens, I think, mistakenly calls SCE a case of "late cutting" pre-emption (2003, 17, fn. 11). Standard late cutting involves the following: an effect; one pre-empting cause; one (non-causal)

doi: 10.48106/dial.v74.i1.06

Let SF stand for Sylvie's firing, BF for Bruno's firing, and JS for the jar's shattering. For two reasons, Strevens argues that CaI delivers this *un*intuitive result: SF is not *at all* a cause of JS. First, SF has no substantial direct influence on JS (2003, 4–5, 12–13). After all, hold fixed BF, and consider an alteration of SF in which Sylvie fires one millisecond earlier/later, or one in which her rifle points one degree to the left/right. Given the properties of both balls, these alterations result in: no collision → Bruno's ball striking the jar (before Sylvie's ball finishes ricocheting) → *no* alteration to JS. Second, there appears no I-CHAIN leading from SF to JS (2003, 5–7, 13–14). This second point, however, is where Choi (2005, 110–13) most seriously disagrees.

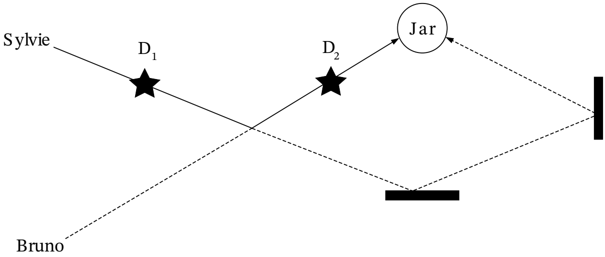

Figure 2:

Referring to Figure 2, and using both Choi's and Lewis's terminology (Choi 2005, 110–11; Lewis 1986a, 2:244–49), let D_1 and D_2 be the (fragile) events whose occurrence conditions consist of all the intrinsic and spatio-temporal properties satisfied by the region that Sylvie's ball occupies at, for D_1, time t_2 before t_c, and for D_2, time t_3 after t_c.

Strevens claims that D_1 has no substantial influence on JS. After all, alter, say, the spatio-temporal properties of Sylvie's ball at t_2. This results in: no collision → no alteration to JS. Strevens also claims: SF has no substantial influence on D_2. After all, alter, say, the timing, or direction of SF. This results in: no collision → the occurrence condition of D_2 being satisfied by *Bruno's* ball (Strevens notes that, on Lewis's metaphysics, it isn't a violation of the occurrence condition of D_2 if the ball at D_2's spatio-temporal region loses the property of "belonging to" Sylvie (2003, 7); said property, after all, is extrinsic).

pre-empted alternative (see Lewis 2000, 182–84). SCE involves an effect that has, intuitively, *two causes*.

Choi, however, claims that Strevens is twice mistaken. (i) D_1 *does* influence JS. After all, alter the *mass*, or *shape*, of Sylvie's ball at t_2. Admittedly, if t_2 were, say, right after t_1, then these alterations result in: Sylvie's ball taking a different post-t_2 trajectory (balls of different mass/shape encounter different amounts of air resistance) → no collision. However, stipulate that t_2 is *right before* t_c. Then, neither alteration prevents the balls' collision. Both, however, alter the manner of the collision, and resultantly the manner of JS. Furthermore, (ii) SF *does* influence D_2. After all, alter the *surface properties*, or *electrical charge*, of the ball Sylvie fires. Neither alteration prevents the balls' collision. Both, however, in altering an intrinsic property of Sylvie's ball at t_3, alter D_2.

Combining (i), the fact that D_1 influences JS, with the (safe) claim that SF influences D_1, and combining (ii), the fact that SF influences D_2, with the (safe) claim that D_2 influences JS, Choi concludes that there are (at least) two I-CHAINs leading from SF to JS—one "via" D_1 (I-CHAIN$_1$), and one "via" D_2 (I-CHAIN$_2$). Thus, CaI delivers the intuitive result that SF is a cause of JS, and "[SCE] spells no trouble whatsoever for [CaI]" (2005, 113).

3 CaI, SCE, and Degrees of Causation

I think, however, that even if Choi's points are correct, SCE still spells some trouble for CaI. In what follows, I argue that, even if Choi's points are correct, CaI does not capture an intuition we have about the *comparative* causal statuses of SF and BF. Thus, insofar as my argument succeeds, SCE remains useful in revealing the failure of CaI to properly account for *degrees* of causation.[3] [4]

Here is the intuition I have in mind:

[3] In the contemporary literature, there exists the idea that CaI can account for, or at least play a role in our understanding of, degrees of causation. Lewis himself, for example, thinks that degrees of causation track degrees of influence (2000, 191). Another example is found in Woodward (2010). Woodward doesn't find CaI promising as an analysis of "causation *simpliciter*" (2010, 304). Nevertheless, he suggests that CaI can play a role in "distinguish[ing] [...] *among* causal relationships" (2010, 304). In more detail, Woodward connects the "specificity" of causal relationships in biological contexts to influence (2010, 301–8). And while he doesn't explicitly state that degrees of causation track degrees of "specificity", he does state that where C_1 and C_2 are both causes of some effect E, if the causal relationship between C_1 and E is more "specific" as compared to the causal relationship between C_2 and E, then possibly we are justified if we "single out or 'privilege' the causal role of [C_1]" (2010, 316). See also Braham and Van Hees (2009, 331, n16), who discuss one point of similarity between their measure of degrees of causation, and CaI.

[4] There is another scenario in which, even if Choi's points are correct, SCE spells trouble for CaI. Say we modify SCE so that both balls detect and decimate balls that aren't intrinsically similar to

doi: 10.48106/dial.v74.i1.06

> Comparative Intuition. SF is *(at least) as much* a cause of JS as is BF.[5]

I think that Comparative Intuition is, and should be, held as strongly as is the (absolute) intuition that SF is *a* cause of JS. A question arises: what buttresses our intuition in Section 1 that Sylvie's rock-throw is a cause of the vase's shattering, while Bruno's laughter is not? One answer is the following: informed (only) of Sylvie's rock-throw, I can *predict*, *explain*, and *blame* someone for the vase's shattering. Informed (only) of Bruno's laughter, I can do none of these things. However, and to use Jonathan Schaffer's terminology, note that "the core *epistemic*, *explanatory*, and *ethical* connotations of causation" (2001, 12–13, emphasis mine) are no *more* present in the claim that "BF caused JS," than they are in the claim that "SF caused JS." Suppose the jar were a national treasure. First, and to endorse Lewis's view that we *don't* ordinarily consider events fragile (2000, 185–86; 1986b, 198), comparing a scenario in which I'm informed (only) of BF with one in which I'm informed (only) of SF, it's not as if I can only predict JS (here taken as a non-fragile event) in the former. Second, consider the question, "Why did the jar shatter?" It is likely that most would find the answer "Because Sylvie fired" to be no more lacking than the answer "Because Bruno fired." Third, it'd be surprising if Judge blamed Bruno more than she did Sylvie. More likely, liability for the jar's damages would be apportioned equally.

Nevertheless, two considerations might motivate

> Counter Intuition. BF is more a cause of JS than is SF.

Consideration$_1$ is this asymmetry: had Sylvie not fired, nothing about JS would have changed. However, had Bruno not fired, the jar would've shattered slightly later, and in a slightly different manner. Consideration$_2$ is that JS occurred at a time, and in a manner more (and, in fact, exactly) in line with Bruno's, rather than Sylvie's, intention.

If, however, Consideration$_1$ and Consideration$_2$ are what motivate Counter Intuition, then Counter Intuition is misleading. Consider this scenario:

them. Then, D_1's influence on JS, and SF's influence on D_2, are eliminated. Consequently, CaI must deliver the *un*intuitive result that SF is not *at all* a cause of JS.

[5] One may worry that, as stated, Comparative Intuition (absurdly) implies that JS was caused twice over (once by SF, and once by BF). If so, one may read Comparative Intuition as saying that SF and BF contributed to *the* causing of JS to the same degree. On this reading, "degrees of causation" should be read as "degrees of causal contribution" (see Kaiserman 2016, 387–89).

> Unlucky President. At time t_1, Assassin$_H$ and Assassin$_R$ poison President's coffee. Assassin$_H$ uses poison *H*, which will induce heart failure at time t_4. Assassin$_R$ uses poison *R*, which will induce respiratory failure at time t_5. At time t_2, President drinks her coffee. At time t_3, however, poison *H* and poison *R* interact in President's system—poison *H* neutralises the respiratory-failure-inducing elements of poison *R*; poison *R* neutralises the heart-failure-inducing elements of poison *H*. But President isn't so lucky—she happens to be fatally allergic to some other element *e* of poison *H*. Element *e* induces in President respiratory failure at t_5, and she dies.

Considerations parallel to Consideration$_1$ and Consideration$_2$ are present in Unlucky President. In Unlucky President, we have Consideration$_1$*, which is this asymmetry: had Assassin$_H$ not poisoned President's coffee, nothing about President's death would have changed. However, had Assassin$_R$ not poisoned President's coffee, President would've succumbed to heart failure at t_4, and not respiratory failure at t_5. In Unlucky President, we also have Consideration$_2$*: President's death occurs at a time, and in a manner more (and, in fact, exactly) in line with Assassin$_R$'s, rather than Assassin$_H$'s, intention. However, does either Consideration$_1$* or Consideration$_2$* push us to think that "Assassin$_R$'s poisoning caused President's death"? No. Most intuitively, Assassin$_H$'s poisoning caused President's death. This shows that considerations like Consideration$_1$ and Consideration$_2$ aren't substantially relevant to causation. Thus, if Counter Intuition is motivated by Consideration$_1$ and Consideration$_2$, then Counter Intuition should be suppressed.

Comparative Intuition, then, is justifiably strong. But I now argue that CaI violates this intuition: it counts SF as (significantly) *less* a cause of JS than is BF.

What determines how much a cause BF is of JS? On CaI, it is (roughly) the amount of influence that BF has on JS (Lewis 2000, 92). What determines this amount? Centrally, it is the size of the range of alterations to BF that lead to changes in JS. Accounting for those types of alterations that Strevens *and* Choi consider, there are (at least) *four* types of alterations to BF that lead to said changes: alterations to the *timing* and *direction* of BF, and to the *mass* and *shape* of the ball Bruno fires.

What determines how much a cause SF is of JS? Because SF has no substantial direct influence on JS,[6] CaI must appeal to I-CHAIN$_1$/I-CHAIN$_2$. For each of these I-CHAINs, however, CaI is silent on whether the determinant is (A) the amount of influence that SF has on D_1/D_2 (the amount of influence present in "link"$_1$ of the I-CHAIN), (B) the amount of influence that D_1/D_2 has on JS (the amount of influence present in "link"$_2$ of the I-CHAIN), or (C) some weighted average of [(A)+(B)]. Nevertheless, let's first determine (A) and (B):

> "LINK"$_1$ OF I-CHAIN$_1$ (At least). *six* types of alterations to SF lead to changes in D_1 (alterations to the *timing* and *direction* of SF, and to the *mass*, *shape*, *surface properties* and *electrical charge* of the ball Sylvie fires);
>
> "LINK"$_2$ OF I-CHAIN$_1$ (At least). *two* types of alterations to D_1 lead to changes in JS (alterations to the *mass* and *shape* of the ball at D_1's spatio-temporal region);
>
> "LINK"$_1$ OF I-CHAIN$_2$ (At least). *two* types of alterations to SF lead to changes in D_2 (alterations to the *surface properties* and *electrical charge* of the ball Sylvie fires);
>
> "LINK"$_2$ OF I-CHAIN$_2$ (At least). *four* types of alterations to D_2 lead to changes in JS (alterations to D_2's *spatio-temporal* properties (this counts for two), and to the *mass* and *shape* of the ball at D_2's spatio-temporal region).

Let the "strength" of an I-CHAIN "link" be the amount of influence present in that "link." I now claim that, for I-CHAIN$_1$ and I-CHAIN$_2$, CaI must say that what determines how much a cause SF is of JS is the strength of the I-CHAIN's *weaker* "link." This follows from my next, more general, claim that if an event C is a cause of another event E because there is a (two-"link") I-CHAIN leading from C to E, then how much a cause C is of E supervenes upon the strength of said I-CHAIN's weaker "link." I will now evidence the just-mentioned general claim by constructing one (two-"link") I-CHAIN in

[6] Admittedly, if Sylvie fires early enough, her ball will ricochet and shatter the jar before Bruno's ball can. We can, however, all but eliminate this small amount of influence by adding to SCE that the jar is placed at its location right before it actually shatters.

each of two causal scenarios. I will then show that, in these I-CHAINS, varying the strength of the stronger "link" (while holding fixed that of the weaker "link") *doesn't* vary our intuitions about how much C is a cause of E. Varying the strength of the weaker "link" (while holding fixed that of the stronger "link"), however, *does*. The first I-CHAIN I construct will possess I-CHAIN$_1$'s *strong-weak* pattern of influence (i.e. C (SF) has no substantial direct influence on E (JS); C strongly influences some intermediate event D (D_1); D weakly influences E). The second will possess I-CHAIN$_2$'s *weak-strong* pattern of influence (i.e. C (SF) has no substantial direct influence on E (JS); C weakly influences D (D_2); D strongly influences E).

> *Scenario 1. Divorce*. Only two things elicit in Wife hatred for Husband (the first significantly more so than the second): (1) the memory of their first fight, which occurred in the rain; (2) the memory of their second fight, which occurred in the fog. Wife, nevertheless, has fallen for Paramour. Thus, she has decided that she will file for divorce from Husband on Thursday afternoon. On Wednesday afternoon, Husband goes on a drinking binge. Late Wednesday night, Husband arrives home. His drunkenness annoys Wife, and the two fight in their driveway. Because fog happens to descend, the fight is so serious to Wife that it (temporarily) lays her thoughts of Paramour to rest, and independently drives her to file for divorce on Thursday afternoon.

We can construct a *strong-weak* I-CHAIN$_{\text{Divorce}}$ with these three events: (C) Husband's drinking binge on Wednesday afternoon; (D) the fight late Wednesday night; (E) Wife's filing for divorce on Thursday afternoon. (1) C has no substantial direct influence on E—altering whether or not/how/what/how long Husband drinks changes nothing about Wife's filing for divorce. (2) C strongly influences D—altering whether or not/how long Husband drinks changes whether or not/at what time the fight occurs. (3) D weakly influences E—altering whether or not/how long Wife and Husband fight changes nothing about Wife's filing for divorce. However, if the fight had occurred in the rain, then Wife would've filed for divorce, say, earlier.

Does strengthening I-CHAIN$_{\text{Divorce}}$'s stronger "link" (C's influence on D) make us intuit that C is more a cause of E than before? No. Add to Divorce that the fight's topic is sensitive to the type of alcohol that Husband consumes—this *doesn't* make us intuit that Husband's drinking binge is more a cause of

doi: 10.48106/dial.v74.i1.06

Wife's filing for divorce than before. But what if we strengthen I-CHAIN$_{\text{Divorce}}$'s weaker "link" (D's influence on E)? Add to Divorce that the timing of Wife's filing for divorce is sensitive to whether or not (but not the extent to which[7]) Husband is drunk during the fight (perhaps Wife takes sober fights most seriously, and would've filed for divorce earlier if Husband had been sober during the fight[8])—contrary to before, this *does* make us intuit that Husband's drinking binge is more a cause of Wife's filing for divorce on Thursday afternoon (and not, say, early Thursday morning).

> *Scenario 2. Resolve*. Colonel is testing Recruit's resolve. Recruit possesses a button which, if pressed, activates a light which Gunman takes as a signal to shoot Prisoner. Gunman will only ever shoot at time t_2. Also, iff Recruit doesn't press the button by time t_1, Colonel will shoot Prisoner at t_2. The following three events occur: (C) Recruit presses the button at t_1; (D) Gunman fires at t_2; (E) Prisoner dies at t_3.

C-D-E form *weak-strong* I-CHAIN$_{\text{Resolve}}$: (1) C has no substantial direct influence on E—altering whether or not/how/when Recruit presses the button changes nothing about Prisoner's death at t_3. (2) C weakly influences D—altering how Recruit presses the button changes nothing about Gunman's firing at t_2. And neither does having Recruit press the button *before* t_1. However, if Recruit hadn't pressed the button (by t_1), Gunman wouldn't have fired. (3) D strongly influences E—altering whether or not/how Gunman fires changes whether or not/how Prisoner dies.

Consider these two possible additions to Resolve: (1) Gunman possesses many rifles to choose from, each of which inflicts death differently; (2) Recruit possesses another button which, if pressed, *prevents* Gunman's firing (Colonel will nonetheless shoot Prisoner at t_2 if this button is pressed[9]). Again, only that addition which strengthens the I-CHAIN's weaker "link" (addition (2)) makes us intuit that C is more a cause of E than before.

There is evidence, then, that in (two-"link") I-CHAINS, how much C is a cause of E supervenes upon the strength of the I-CHAIN's weaker "link." Consequently, unless one (a) reasonably explains why this doesn't apply to

[7] This stipulation denies the substantial direct influence of C on E.

[8] I think that an alteration of the fight in which Husband is sober requires no bigger a Lewisian "miracle" (1979, 468–69) than do those alterations of D_1 that Choi appeals to.

[9] This stipulation denies the substantial direct influence of C on E.

I-CHAIN$_1$ and/or I-CHAIN$_2$, or (b) denies that the causal status of *C* has something to do with I-CHAINS (or counterfactual dependence in general), then how much SF is a cause of JS supervenes upon the strength of "link"$_2$, for I-CHAIN$_1$, and "link"$_1$, for I-CHAIN$_2$.

This result, however, likely forces CaI to (counterintuitively) count SF as (significantly) *less* a cause of JS than is BF. After all, *four* types of alterations to BF count towards the influence that BF has on JS. Only *two* types of alterations to D_1 count towards the influence that D_1 has on JS. And only *two* types of alterations to SF count towards the influence that SF has on D_2. Certainly, it remains possible that for, say, I-CHAIN$_2$, the *total number* (as opposed to the number of *types*) of alterations to SF that lead to changes in D_2 is greater than the *total number* of alterations to BF that lead to changes in JS. But this would be surprising. Why think, for example, that there are (significantly) more surface properties that Sylvie's ball might have had, than there are angles at which Bruno might have fired? It also remains possible for the defender of CaI to try to identify more *types* of alterations to SF that lead to changes in D_2. This strategy, however, can only be a stopgap, unless it can be shown that, for each such newly-identified type of alteration to SF, there is no not-too-distant, hitherto-unidentified, type of alteration to BF that leads to changes in JS. Showing this would be difficult. After all, there appear many examples of the latter (e.g. altering properties like the muzzle velocity and barrel length of Bruno's *rifle* will affect the travel of his ball).

I end by blocking one last maneuver that the defender of CaI might perform. Consider:

> "THRESHOLD" OPERATION OF CAI. Causation isn't a scalar relation. That is, there are no degrees of causation—either an event *C* is a cause of another event *E*, or it isn't. Thus, if the strength of the weaker "link" of I-CHAIN$_1$/I-CHAIN$_2$ determines anything, it's simply *whether or not* SF is a cause of JS. That said, in both I-CHAINS, said strength meets that minimum amount of influence *x* required to establish causation. So there is a sense in which CaI *does* capture Comparative Intuition—SF is "as much" a cause of JS as is BF in that neither firing can be said to be more or less a cause than the other. (On "Threshold" Operation, then, any influence that *C* has on *E* exceeding *x* is ignored.)

doi: 10.48106/dial.v74.i1.06

Besides its diverging from Lewis's writing[10], there are (at least) two reasons to reject "Threshold" Operation.

First, causation *is* plausibly a scalar relation. After all, this appears to be the "common sense", or "ordinary", view. For one thing, Hitchcock and Knobe offer experimental evidence for their claim that "ordinary causal judgments of subjects" come in degrees (2009, 602). For another thing, Michael Moore argues that the *law* treats causation as scalar (2009, 71, 118–23; see also Braham and Van Hees 2009, 324). Thus, in tort law, the idea of "degrees of causal contribution" is both taken as sensible, and employed widely. We see this especially in negligence cases in which the doctrine of *divisible harm* is invoked so as to apportion liability amongst several defendants according to the degree of causal contribution each makes to some *in*divisible harm (Moore 2009, 118–19). In one such case[11]— *Moore v. Johns-Manville Sales Corp* 781 F 2d 1061 (5th Cir 1986)—liability for each plaintiff's asbestosis was apportioned according to the degree to which each (defendant) manufacturer's (asbestos-containing) products caused the plaintiff's asbestosis (i.e. each defendant's "degree of relative causation"). Therefore, if we think that our concept of causation should accord with how causation is employed "ordinarily," we should also think that causation *is* a scalar relation.

Second, determining the value of x appears impossible. After all, x cannot be some one particular value. This is because we can easily conceive of one pre-emption case in which (the event intuited as) the pre-empting cause *doesn't* exhibit x amount of influence on the effect, and another pre-emption case in which (the event intuited as) the (non-causal) pre-empted alternative *does* (see Dowe 2000, 6–7). One may then suggest that one determine x on a case-by-case basis. This, however, would require one to establish some standard set of case features relevant to determining x (so as to ensure that our determinations of x are not *ad hoc*). At this point, however, I simply cannot see what these features might be.[*]

Joshua Goh
University College London
hseng.goh.14@ucl.ac.uk

[10] (2000, 191) indicates that Lewis thinks causation *is* a scalar relation; (2000, 188–89) sees Lewis establish causation with reference to comparative, and not absolute, standards.

[11] Moore (2009, 119, fn. 36) contains more case examples.

[*] For invaluable input, thanks to Arif Ahmed, Luke Fenton-Glynn, two anonymous referees from the University of Cambridge, and three anonymous referees for *Dialectica*.

References

BRAHAM, Matthew, and Martin VAN HEES. 2009. “Degrees of Causation.” *Erkenntnis* 71 (3): 323–44. doi:10.1007/s10670-009-9184-8.

CHOI, Sungho. 2005. “Understanding the Influence Theory of Causation: A Critique of Strevens.” *Erkenntnis* 63 (1): 101–18. doi:10.1007/s10670-005-0607-x.

DOWE, Phil. 2000. “Is Causation Influence?” http://fitelson.org/269/Dowe_ICI.pdf. Unpublished manuscript.

HITCHCOCK, Christopher R., and Joshua KNOBE. 2009. “Cause and Norm.” *The Journal of Philosophy* 106 (11): 587–612. doi:10.5840/jphil20091061128.

KAISERMAN, Alex. 2016. “Causal Contribution.” *Proceedings of the Aristotelian Society* 116 (3): 387–94. doi:10.1093/arisoc/aow013.

LEWIS, David. 1973. “Causation.” *The Journal of Philosophy* 70 (17): 556–67. doi:10.2307/2025310. Reprinted, with a postscript (Lewis 1986b), in Lewis (1986a, 2:159–213).

———. 1979. “Counterfactual Dependence and Time’s Arrow.” *Noûs* 13 (4): 455–76. doi:10.2307/2215339.

———. 1986a. *Philosophical Papers*. Vol. 2. Oxford: Oxford University Press.

———. 1986b. “Postscript to Lewis (1973).” In *Philosophical Papers*, 172–213. Oxford: Oxford University Press.

———. 2000. “Causation as Influence.” *The Journal of Philosophy* 97 (4): 181–97. doi:10.2307/2678389.

MOORE, Michael S. 2009. *Causation and Responsibility. An Essay in Law, Morals, and Metaphysics*. Oxford: Oxford University Press.

SCHAFFER, Jonathan. 2001. “Causation, Influence, and Effluence.” *Analysis* 61 (1): 11–19. doi:10.1093/analys/61.1.11.

STREVENS, Michael. 2003. “Against Lewis’s New Theory of Causation: A Story with Three Morals.” *Pacific Philosophical Quarterly* 84: 398–412. doi:10.1046/j.1468-0114.2003.00182.x.

WOODWARD, James F. 2010. “Causation in Biology: Stability, Specificity, and the Choice of Levels of Explanations.” *Biology and Philosophy* 25 (3): 287–318. doi:10.1007/s10539-010-9200-z.

doi: 10.48106/dial.v74.i1.06

Consistency, Obligations, and Accuracy-Dominance Vindications

Marc-Kevin Daoust

Vindicating the claim that agents ought to be consistent has proved to be a difficult task. Recently, some have argued that we can use accuracy-dominance arguments to vindicate the normativity of such requirements. But what do these arguments prove, exactly? In this paper, I argue that we can make a distinction between two theses on the normativity of consistency: the view that one ought to be consistent and the view that one ought to avoid being inconsistent. I argue that accuracy-dominance arguments for consistency support the latter view, but not necessarily the former. I also argue that the distinction between these two theses matters in the debate on the normativity of epistemic rationality. Specifically, the distinction suggests that there are interesting alternatives to vindicating the strong claim that one ought to be consistent.

The normativity of the following formal coherence requirements is contentious:

> Belief Consistency. If A believes that *p*, it is false that A believes that ¬*p*.[1]
>
> Credal Consistency. If A has a credence of X in *p*, then A has a credence of (1-X) in ¬*p*.

Do we fall under an obligation to satisfy these requirements?[2] Many philosophers like John Broome (2013, ch. 13) are convinced that the above requirements are normative, but cannot find a satisfactory argument in favour of such a conclusion. Other philosophers are less optimistic. For instance, Niko

[1] This requirement is sometimes called “Pairwise Consistency”, as in Easwaran (2016).

[2] See Way (2010) for an overview of this debate. See Fitelson (2016) on epistemic teleology and coherence requirements. See Bona and Staffel (2018) on accuracy and approximation of Bayesian requirements of probabilistic coherence. See also Pettigrew (2013, 2016a).

Kolodny (2005; 2007b; 2007a, 230–31) has argued that there is no reason to be consistent. According to him, what matters from an epistemic point of view is acquiring true beliefs (or acquiring beliefs that are likely to be true on the evidence) and avoiding false beliefs (or avoiding beliefs that are likely to be false on the evidence). However, a perfectly consistent system of beliefs (or credences) can be entirely false, inaccurate or improbable on the evidence. So, consistency requirements are not normative, in the sense that one does not necessarily have a reason to be consistent.

Recently, a new strategy has emerged to vindicate the normativity of Consistency. This strategy relies on accuracy-dominance principles, which roughly say that if state *Y* is better than state *X* at every possible world, one ought to avoid state *X*. However, there is a weak and a strong interpretation of what is entailed by the accuracy-dominance arguments. According to the strong interpretation, accuracy-dominance arguments entail that one ought to be consistent. Joyce, for instance, argues that:

> It is thus established that degrees of belief that violate the laws of probability are invariably less accurate than they could be. Given that an epistemically rational agent will always strive to hold partial beliefs that are as accurate as possible, this vindicates the fundamental dogma of probabilism [according to which degrees of belief must make conformity to the axioms of probability]. (1998, 600)

According to the weak interpretation, accuracy-dominance arguments merely entail that ought not to be inconsistent. Easwaran, for instance, says that “we can use dominance to *eliminate*” the inconsistent doxastic options (2016, 826, emphasis added). In other words, dominance is here used to argue against inconsistency. Thus, we can make the following distinction between two views:

> Normativity+. Given the accuracy-dominance arguments, A ought to be consistent.

> Normativity−. Given the accuracy-dominance arguments, A ought not to be inconsistent.

This paper argues that, while accuracy-dominance arguments can vindicate Normativity−, they do not necessarily vindicate Normativity+. Specifically, accuracy-dominance arguments vindicate Normativity+ when supplemented with a contentious hypothesis concerning the relationship between reasons for and reasons against. Hence, accuracy-dominance arguments do not vindicate Normativity+ *on their own*.

In Section 1, I clarify the debate on the normativity of Consistency. In Sections 2 and 3, I present two important arguments in the debate surrounding the normativity of Consistency: accuracy-dominance arguments and Kolodny's objection from truth-conduciveness. Both arguments are veritistic: They assume that only true beliefs bear final epistemic value, and only false beliefs bear final epistemic disvalue. I argue that, under the assumption that veritism is true, the only way to make sense of both arguments is to make a distinction between Normativity+ and Normativity− (i.e. to deny that both views are coextensive). Then, I argue that accuracy-dominance arguments fail to vindicate Normativity+.

This is not necessarily bad news. In conclusion, I explain why this might be an occasion to adjust our expectations in the debate on the normativity of formal coherence requirements. Many people think that there is something bad or suboptimal with inconsistent combinations of attitudes. The mistake might have been to try to explain this assumption in terms of *an obligation to be consistent*. Being in a position to vindicate Normativity− while remaining neutral on Normativity+ could be advantageous in the debate on the normativity of formal coherence requirements.

1 The "Why-Be-Consistent?" Challenges

There are many putative explanations of why one ought to have *some* consistent combinations of beliefs. They stem from the normative authority of truth, knowledge or reasons, as in the following:

> Truth Vindication. One ought to believe *p* if and only if *p*. Truth is consistent (or: Inconsistent propositions cannot be true simultaneously). So, one ought to have some consistent combinations of beliefs (e.g. the true ones).
>
> Knowledge Vindication. One is epistemically permitted to believe *p* if and only if one is in a position to know that *p*. Knowledge is consistent (or: Propositions that one is in a position to know

doi: 10.48106/dial.v74.i1.07

> cannot be inconsistent with each other). So, one is only epistemically permitted to believe consistent combinations of beliefs.
>
> Reasons Vindication. One is epistemically permitted to believe *p* if and only if one has sufficient epistemic reason to believe *p*. One never has sufficient epistemic reason to believe *p* and sufficient epistemic reason to disbelieve *p* simultaneously. So, one is only epistemically permitted to believe consistent combinations of beliefs.[3]

Philosophers like Broome (2013) and others are worried that the above putative vindications do not fully vindicate the normativity of Consistency. Some consistent combinations of beliefs may include some false, unjustified or unreasonable beliefs. Even if consistent agents sometimes believe propositions that are false, unjustified or unreasonable, it seems that they satisfy a distinct obligation to have consistent beliefs (e.g. an obligation that does not boil down to truth, knowledge or reasons). In other words, perhaps the agent is unjustified, mistaken or unreasonable, but one could still say: *At least he or she is consistent*. Here, the putative obligation to be consistent will not come from truth, knowledge or reasons.[4]

So, according to some philosophers, the above vindications are somehow incomplete. Perhaps we can easily argue that agents ought to have *some* consistent combinations of beliefs, but finding a vindication of Consistency that covers all the possible consistent combinations of beliefs has proved to be a difficult task.

It should also be noted that the normativity of Consistency is part of a broader debate on the normativity of *structural rationality*. Structural rationality allegedly requires of agents not to be incoherent—for example, not to be akratic, not to have intransitive preferences, and so forth (Worsnip 2018a, 2018b). So, in addition to Consistency, there are other putative structural requirements of rationality, like:

[3] Kolodny (2007b) endorses this view. See Daoust (2020) for discussion.

[4] In fact, Broome (2013, ch. 11) is interested in the stronger claim that rationality is a *source* of normativity. So, he is not interested in offering a derivative vindication of consistency requirements, that is, a vindication of these requirements on other grounds (like truth, knowledge, or reasons). By contrast, dominance principles are often tied to rationality (see e.g. Joyce 1998).

> Inter-Level Coherence. Rationality requires that, if A believes that he or she has sufficient epistemic reason to believe *p*, then A believes that *p*.[5]

> Instrumental Principle. Rationality requires that, if A intends to ϕ, and A believes that ψ-ing is a necessary means to ϕ-ing, then A intends to ψ.[6]

Broome and others have tried to find compelling arguments for the claim that *structural rationality* has normative authority. However, structural rationality is neutral on whether one's beliefs should be true, reasonable or amount to knowledge. Some entirely false and unreasonable belief systems can satisfy the requirements of structural rationality. So, at least given the agenda of these philosophers, a good vindication of the normativity of Consistency should cover the cases in which one's beliefs are false or unreasonable.

An interesting feature of accuracy-dominance arguments is that they remain neutral on whether one's beliefs should be true, reasonable or amount to knowledge. They focus on what is wrong with having some combinations of beliefs, regardless of the substantive properties of such beliefs.

2 Accuracy-Dominance and Consistency

Accuracy-dominance arguments for vindicating the normativity of Consistency come from decision theory and rely on the following principle:

> Strong Dominance. If an available state *X* is strongly dominated by an available state *Y* at every possible world, in the sense that state *Y* is better or has more value than state *X* at every possible world, one ought to avoid state *X*.

Strong Dominance has been used to vindicate probabilism, the view roughly stating that an agent's rational credences should satisfy the probability ax-

[5] Coates (2012) and Lasonen-Aarnio (2020) have argued that responding correctly to one's evidence sometimes entail believing "P, but I am irrational to believe P", which is an incoherent combination of attitudes. They conclude that such incoherence is not necessarily irrational. See Greco (2014), Horowitz (2014), Kiesewetter (2016), Littlejohn (2018), Titelbaum (2015) and Worsnip (2018a) for various responses to this view.

[6] See, among others, Broome (2013, sec.9.4), Kiesewetter (2017, ch. 10) and Way (2013) on the Instrumental Principle.

doi: 10.48106/dial.v74.i1.07

ioms. With respect to some inaccuracy measures such as the Brier score, probabilistically inconsistent agents have access to a credence function that is less inaccurate (and thus less epistemically disvaluable) at every possible world (Joyce 1998; Leitgeb and Pettigrew 2010; Pettigrew 2016a).

For the sake of simplicity, I will leave aside dominance for credence and focus on dominance for belief (these arguments have the same structure, but dominance arguments for belief are more accessible).

There is a plausible explanation of why inconsistent combinations of beliefs are strongly dominated. An agent can take different doxastic attitudes towards p, as in the following:

(i) Believing p and not disbelieving p,
(ii) Disbelieving p and not believing p,
(iii) Neither believing nor disbelieving p,
(iv) Believing p and disbelieving p.

The question is whether (iv) is strongly dominated. To answer this question, we need to determine the epistemic value of (iv) at every possible world. In veritistic frameworks, only true beliefs have final epistemic value and only false beliefs have final epistemic disvalue. Accordingly, T is the epistemic value of having a true belief (for $T > 0$), F is the epistemic disvalue of having a false belief (for $F < 0$), and the epistemic value of not believing p (or not disbelieving p) is 0.[7] Finally, assume that $T \leq -F$, which amounts to endorsing a conservative account of epistemic value. The conservative constraint on epistemic value is plausible.[8] As Dorst says:[9]

> [An epistemically rational agent] will be doxastically conservative... Why? Well here's a fair coin—does she believe it'll land heads? Or tails? Or both? Or neither? Clearly neither. But if she cared more about seeking truth than avoiding error, why not believe both? She'd then be guaranteed to get one truth and one

[7] I'm glossing over some inessential subtleties here. It is possible to assign a value to not believing p (or to withholding judgment on whether p), but ultimately, we would get exactly the same results. See Easwaran (2016, sec.C) and Dorst (2019, 10, n. 12).

[8] But this constraint might not stem from accuracy-first epistemology. See Steinberger (2019) and the next footnote.

[9] In addition to Dorst's argument, see Easwaran (2016), Easwaran and Fitelson (2015) and Pettigrew (2016b) for similar arguments in favour of the conservative account of epistemic value. See Steinberger (2019) on why alternatives to conservatism are compatible with accuracy-first epistemology.

> falsehood, and so be more accurate than if she believed neither... Upshot: we impose a *Conservativeness* constraint to capture the sense in which Rachael has 'more to lose' in forming a belief than she does to gain. (2019, 11)

Then, we can determine the possible values of each option at every possible world. Since the value of these options is solely determined by p's truth value, we need to consider the worlds in which p is true and the worlds in which p is false, as in Table 1.

Table 1: An agent's doxastic options with respect to p

Doxastic options / possible world	p is true	p is false
Believing p and not disbelieving p	T	F
Disbelieving p and not believing p	F	T
Neither believing nor disbelieving p	0	0
Believing p and disbelieving p	$T+F$	$T+F$

Finally, in accordance with Table 1, we can conclude that inconsistent combinations of beliefs are strongly dominated. The following reasoning supports such a conclusion:

(1) $T \leq -F$ (conservative assumption). Accordingly, $T+F<0$.
(2) Following (1) and Table 1, believing p and disbelieving p simultaneously has an epistemic value of less than 0 at every possible world.
(3) However, following Table 1, neither believing nor disbelieving p has an epistemic value of 0 at every possible world.
(C) Therefore, following (2) and (3), inconsistent combinations of beliefs such as believing p and disbelieving p are strongly dominated: another available option (neither believing nor disbelieving p) is more valuable at every possible world.[10]

Hence, one ought to avoid being inconsistent.

[10] Similar arguments can be found in Easwaran (2016§B) and Pettigrew (2016b, 256). Dorst (2019, 31, esp. proposition 3) argues for a similar but contextualist view.

3 Truth-Conduciveness, Reasons For and Reasons Against

3.1 *Kolodny's Objection From Truth-Conduciveness*

The above argument states that inconsistent combinations of beliefs are dominated, which means that one ought not to be inconsistent. Naturally, this seems to suggest that one ought to be consistent. But this equivalence is less obvious than it seems.

To see why, consider Kolodny's argument against the normativity of Consistency. According to him, one does not necessarily have an epistemic reason to be consistent. Rather, what matters from an epistemic point of view is having true beliefs and avoiding false beliefs, and satisfying Consistency does not guarantee a better ratio of true to false beliefs. In fact, some perfectly consistent sets of beliefs are entirely false (or improbable on the evidence). Kolodny summarizes his argument in the following way:

> From the standpoint of theoretical deliberation—which asks 'What ought I to believe?'—what ultimately matters is simply what is likely to be true, given what there is to go on. [...] [But] formal coherence may as soon lead one away from, as toward, the true and the good. Thus, if someone asks from the deliberative standpoint 'What is there to be said for making my attitudes formally coherent as such?' there seems, on reflection, no satisfactory answer. (2007a, 231)

In other words, if one merely satisfies Consistency, one is not more likely to end up forming true beliefs and avoiding false beliefs. So, the mere satisfaction of Consistency does not improve one's ratio of true to false beliefs. In view of the foregoing, Kolodny thinks that it is false that one falls under an obligation to be consistent.[11]

3.2 *Comparing the Objection from Truth-Conduciveness and Accuracy-Dominance Arguments*

Kolodny argues that there is no reason to be consistent. His argument relies on the fact that being consistent does not guarantee a good ratio of true to false

[11] Elsewhere, Kolodny (2005) raises some objections against the normativity of other structural requirements, such as Inter-Level Coherence.

beliefs. By way of contrast, accuracy-dominance arguments suggest that there is good reason not to be inconsistent. If one is inconsistent, one is strongly dominated, in the sense that one has access to a better option at every possible world. For instance, if one believes *p* and disbelieves *p* simultaneously, one will necessarily improve one's situation by neither believing nor disbelieving *p*.

Accuracy-dominance arguments and Kolodny's objection from truth-conduciveness are both veritistic.[12] Indeed, they presuppose that only true beliefs bear final epistemic value, and only false beliefs bear final epistemic disvalue. Nevertheless, such arguments apparently support incompatible conclusions concerning the normativity of Consistency: Kolodny argues that veritism entails the denial of the normativity of Consistency, whereas accuracy-dominance arguments support the normativity of Consistency. This is puzzling.

Perhaps Kolodny and accuracy-dominance theorists do not endorse the same version of veritism. Veritism says that only true beliefs have final epistemic value, and only false beliefs have final epistemic disvalue. However, when it comes to epistemic obligations and permissions, these assumptions concerning epistemic value might translate in many different ways. For instance, perhaps agents ought to maximize their *total* epistemic score (e.g. the total balance of epistemic value they get from their doxastic states), or perhaps agents ought to maximize their *expected* epistemic score. For clarity, consider the following example: Suppose *p* is very likely relative to a body of evidence E. But as it happens, *p* is false. Then, believing *p* (or having a high credence in *p*) might maximize expected epistemic value with respect to E. But disbelieving *p* (or having a low credence in *p*) will maximize epistemic value *tout court*.

Yet, it is implausible that a difference in how Joyce and Kolodny understand veritism is the reason why they disagree. Kolodny's argument can be reformulated in many different ways. Consider the following possibilities: (i) Suppose agents ought to maximize *expected* accuracy. Then, Kolodny could say: Some consistent combinations of beliefs can minimize expected accuracy (believing the most improbable propositions can be consistent). (ii) Suppose agents ought to optimize their ratio of true to false beliefs. Then, Kolodny could argue that some agents with a very bad ratio of true to false beliefs are consistent. (iii) Suppose agents ought to maximize *total* accuracy. Then, Kolodny could say: Some consistent combinations of beliefs can minimize

[12] See notably Goldman (2015) and Whiting (2010) on veritism.

accuracy (believing false propositions only can be consistent). As we can see, Kolodny's objection is malleable.[13]

Another possibility is that Kolodny and accuracy-first theorists have a different understanding of what "ought" means. We can make a distinction between normativity in the rule-following sense (as in: Relative to domain D, A ought to X) and normativity in the reason-involving sense (as in: A has a reason to X).[14] For example, the rules of etiquette require of agents to be polite, but agents might lack a reason to be polite. By way of analogy with the rules of etiquette, perhaps accuracy-first theorists are merely interested in arguing that the rules of rationality require consistency. This would be compatible with Kolodny's view—namely, that agents do not have a reason to be consistent. Both views would then be compatible with each other.

It is true that accuracy-first theorists see Consistency as a demand of rationality. However, it is implausible that accuracy-first theorists are *merely* concerned with normativity in the rule-following sense. Accuracy-first theorists like Joyce tie norms of rationality to epistemic value, as in the following:

> The Norm of Truth. An epistemically rational agent must strive to hold a system of full beliefs that strikes the best attainable overall balance between the epistemic good of fully believing truths and the epistemic evil of fully believing falsehoods (1998, 577).
>
> The Norm of Gradational Accuracy. An epistemically rational agent must evaluate partial beliefs on the basis of their gradational accuracy, and she must strive to hold a system of partial beliefs that, in her best judgment, is likely to have an overall level of gradational accuracy at least as high as that of any alternative system she might adopt (1998, 579).

Satisfying the requirements of rationality is different from, say, satisfying the requirements of etiquette. The former has a privileged relationship to value. Epistemically rational agents want to optimize their overall balance of epistemic value. Accordingly, it would be surprising that Joyce and others are merely concerned with normativity in the rule-following sense. Specifically,

[13] I thank a referee for inviting me to discuss this possibility.

[14] See Parfit (2011, 144–48) on this distinction.

it would be surprising that, while rationality has some sort of privileged relationship to value, it is merely normative in the rule-following sense.[15]

Under the assumption that Kolodny and accuracy-dominance theorists agree upon a specific version of veritism and the meaning of "ought," the natural reaction is to think that at least one of the above arguments is mistaken—either the objection from truth-conduciveness is inconclusive, or accuracy-dominance arguments fail. After all, how can there be no reason to be consistent and reasons against being inconsistent? If there is something wrong with being inconsistent, there must be something good with being consistent!

However, this natural reaction presupposes that there is always a connection between (i) reasons for being consistent (as in Normativity+) and (ii) reasons against being inconsistent (as in Normativity−). Call this the Coextensivity Thesis, as in the following:

> COEXTENSIVITY THESIS. Arguments in favour of Normativity− count as arguments in favour of Normativity+ (and vice versa).

Those who endorse the Coextensivity Thesis think that (i) and (ii) express the same normative relation.

If the Coextensivity Thesis were correct, then Kolodny's objection from truth-conduciveness would be inconclusive. Under the assumption that the Coextensivity Thesis is correct, two kinds of considerations can vindicate the view that one ought to be consistent—namely, reasons be consistent and reasons against being inconsistent. Kolodny argues for the *absence* of reasons in favour of being consistent. But if the Coextensivity Thesis is correct, *such considerations are just half of the story*. We also need to consider whether there are reasons against being inconsistent in the balance, since they count as reasons for being consistent. Accuracy-dominance arguments entail that one ought not to be inconsistent. So, even if Kolodny is right that there is no reason to satisfy Consistency, this does not entail that it is false that one ought to be consistent. Insofar as there are arguments against inconsistency (as suggested by accuracy-dominance arguments), there is a reason to be consistent.

However, if the Coextensivity Thesis is false, then accuracy-dominance arguments are compatible with the objection from truth-conduciveness. Here is why. Kolodny argues that there is no reason to be consistent: he denies that one ought to be consistent, as in Normativity+. However, if the Coextensivity

[15] I thank a referee for inviting me to clarify this possibility.

doi: 10.48106/dial.v74.i1.07

Thesis is false, we can deny Normativity+ without denying Normativity−. In other words, even if it is false that one ought to be consistent, perhaps one ought not to be inconsistent. The same goes for accuracy-dominance arguments. According to such arguments, inconsistent combinations of beliefs are dominated. So, one ought not to be inconsistent. But if the Coextensivity Thesis is false, this does not entail that one ought to be consistent.

3.3 *Reasons to be Consistent and the Coextensivity Thesis*

So, is the Coextensivity Thesis true? This depends on what “a reason to be consistent” means. Suppose, like Kolodny, that “a reason to be consistent” concerns each individual consistent option one has (see section 3.1. That is, suppose that “a reason to be consistent” means something like “a consideration that counts in favour of *each* individual consistent options one has.” For Kolodny, nothing can be said in favour of some consistent combinations of attitudes. So, under this interpretation of what “a reason to be consistent” means, we do not necessarily have a reason to be consistent.

Relative to this interpretation of what “a reason to be consistent” means, the Coextensitivity Thesis does not seem plausible. For reasons found in Snedegar (2018), we can make a distinction between reasons for Consistency (as in Normativity+) and reasons against inconsistency (as in Normativity−). The distinction comes from the following account of reasons for and reasons against endorsed by Snedegar:

> My view puts a strong condition on reasons for and a weak condition on reasons against. For some objective to provide a reason for an option, that option has to do the best with respect to the objective. For some objective to provide a reason against an option, that option only has to do worse than some alternative. (2018, 737)

Snedegar roughly argues that the problem with views that lump together reasons against and reasons for is that there can be good reasons not to ϕ, even if there are worse alternatives to ϕ-ing.[16] For instance, suppose that I am trying to decide what to drink. I might have conclusive reason not to drink gin, but this does not entail that I have a reason to drink any beverage that

[16] See Snedegar (2018) for more details.

isn't gin. I should definitely not drink petrol, even if petrol isn't gin. This is compatible with my having conclusive reason not to drink gin.

Snedegar's observation sits well with accuracy-dominance arguments discussed in Section 2. Indeed, recall the options agents have in Table 1. Clearly, there is conclusive reason not to go for the inconsistent option, since neither believing nor disbelieving *p* is better than being inconsistent at every possible world. However, this does not entail that there is a reason in favour of every alternative to the inconsistent option. For instance, disbelieving *p* when *p* is true (or believing *p* when *p* is false) is worse than being inconsistent. So, as in the gin and petrol case, reasons against inconsistency are logically weaker than reasons for Consistency.

This suggests that accuracy-dominance arguments do not vindicate Normativity+ on their own. Of course, when combined with the Coextensivity Thesis, these arguments support Normativity+. But Kolodny's interpretation of what "a reason to be consistent" means conflicts with the Coextensivity Thesis. So, while accuracy-dominance arguments support Normativity−, it is an open question whether they also support Normativity+.

Here is a response to my argument on behalf of the accuracy-dominance theorist. We can regroup the consistent options in Table 1 under a single option. Call this the consistent option. With respect to the consistent option, Snedegar's distinction does not apply. If there is conclusive reason not to go for the inconsistent option, and the only option left is the "regrouped" consistent option, then reasons against inconsistency favour the consistent option. So, could there be a sense in which the Coextensivity Thesis is true?[17]

My response to this objection goes as follows. This way of framing the problem cannot make sense of Kolodny's objection concerning some consistent options. *There is something wrong with some consistent combinations of beliefs* —some consistent combinations of beliefs are entirely wrong or improbable on the evidence. Kolodny is right to point out that nothing can be said in favour of these combinations of attitudes. The only way to make sense of Kolodny's objection is *not* to regroup all the consistent options under a single label, precisely because relevant normative distinctions can (and should) be made between some consistent options.

At best, this reply shows that, under a different interpretation of what "a reason to be consistent" means, the Coextensivity Thesis is true. But Kolodny's argument still succeeds relative to another interpretation of this expression.

[17] I thank a referee for inviting me to discuss this objection.

doi: 10.48106/dial.v74.i1.07

When Kolodny discusses the normativity of Consistency, he discusses the normativity of the individual consistent options one has, including the ones that are entirely wrong or improbable on the evidence. The accuracy-dominance theorist can claim that one ought to be consistent, but that is simply because the expression "one ought to be consistent" here refers to something logically weaker than what Kolodny has in mind.[18]

3.4 *An Escape Route for the Accuracy-Dominance Theorist?*

The accuracy-dominance theorist could then offer the following objection. Suppose there is an accuracy-dominance argument against one's attitudes. Accordingly, one can identify at least one collection of attitudes that veritistically dominates one's current state. If agents can identify at least one set of attitudes that is better than their current state, then they have a reason to take the dominating set of attitudes, which will be consistent. Doesn't this support the view according to which one ought to be consistent? If agents ought to take dominating combinations of beliefs, and such combinations of beliefs are consistent, then this seems to entail that agents ought to be consistent.[19]

This objection carries weight depending on what accuracy-dominance arguments prove. Let me explain.

Suppose the contender is right. Then, accuracy-dominance vindications are akin to the Truth Vindication, the Knowledge Vindication or the Reasons Vindication discussed in Section 1. If one has inconsistent combinations of beliefs (say, one believes p and also believes $\neg p$), the Truth Vindication says that agents ought to maintain the true one (and abandon the false one), the Knowledge Vindication says that agents are only permitted to maintain the known one, and the Reasons Vindication says that agents are only permitted to maintain the reasonable one (and ought to abandon the *unreasonable* one). In any case, satisfying such norms means that agents will cease entertaining inconsistent combinations of beliefs.

The contender makes a similar point. If one has inconsistent combinations of beliefs, one should go for the option dominating inconsistent combinations of beliefs. But if that is right, the accuracy-dominance argument merely entails

[18] My response might not convince some readers. In any case, we can draw a lesson from this discussion. We have learned that the expression "a reason to be consistent" is ambiguous. Some readings of this expression are a problem for Kolodny's argument, and other readings of this expression conflict with vindicating Normativity+.

[19] I thank a referee for bringing this objection to my attention.

that agents ought (or have reasons) to have *some* combinations of beliefs, not *any* consistent combination of beliefs. In other words, the argument leaves out some consistent combinations of beliefs.

This brings us back to the discussion in Section 1. What do we expect from a good vindication of the normativity of Consistency? For many philosophers, a good vindication of Consistency should cover all the possible consistent combinations of beliefs. If the contender is right, then accuracy-dominance arguments can explain the significance of some consistent combinations of beliefs—namely, the dominating ones. But this is not what we were looking for. The explanation should apply to *all* the consistent combination of beliefs. To be clear: Some philosophers might not be interested in this specific interpretation of the "Why-Be-Consistent" debate. It should be clear that, with respect to other understandings of the question, the contender is right.

4 Conclusion and Implications in the Debate on the Normativity of Structural Rationality

This paper supports the view that there are two theses concerning the normativity of Consistency: Normativity+ and Normativity−. While accuracy-dominance arguments support Normativity−, they might not necessarily support Normativity+. This is so, because the Coextensivity Thesis might be false. In fact, one way to reconcile Kolodny's objection from truth-conduciveness with accuracy-dominance arguments is to deny the Coextensivity Thesis.

These clarifications concerning Normativity+ and Normativity− allow us to rethink the debate on the normativity of structural rationality. Indeed, a popular strategy for arguing against the normativity of structural rationality is to point out that there is no reason to satisfy some specific rational requirements (such as Consistency). Kolodny's objection from truth-conduciveness is a good illustration of such arguments. These arguments are compelling if we focus on Normativity+. But this might be a mistake. Perhaps that, when it comes to formal requirements like Consistency, the only view we should try to vindicate is Normativity−.

The argument of this paper allows us to make sense of some pre-theoretically correct assumptions structural requirements of epistemic rationality such as Consistency. Plausibly, there is something wrong, suboptimal or disvaluable with inconsistent combinations of beliefs. The mistake might have been to try to explain this assumption in terms of *an*

doi: 10.48106/dial.v74.i1.07

obligation to be consistent. But if I am right, we might only be able to explain this assumption in terms of *an obligation not to be inconsistent*. Hence, requirements like Consistency might merely be normative in a weak sense.

The good news is that we can now make sense of such a possibility. If the Coextensivity Thesis is false, it makes perfect sense to say that one ought not to be inconsistent without also saying that one ought to be consistent. There might not be something good with being structurally rational, but it seems patently clear that there is something bad with being structurally irrational.[*]

Marc-Kevin Daoust
Harvard University
mk.daoust@live.ca

References

Bona, Glauber de, and Julia Staffel. 2018. "Why Be (Approximately) Coherent?" *Analysis* 78 (3): 405–15. doi:10.1093/analys/anx159.

Broome, John A. 2013. *Rationality Through Reasoning*. Oxford: Wiley-Blackwell.

Coates, Allen. 2012. "Rational Epistemic Akrasia." *American Philosophical Quarterly* 49 (2): 113–24.

Daoust, Marc-Kevin. 2020. "The Explanatory Role of Consistency Requirements." *Synthese* 197: 4551–69. doi:10.1007/s11229-018-01942-8.

Dorst, Kevin. 2019. "Lockeans Maximize Expected Accuracy." *Mind* 128 (509): 175–211. doi:10.1093/mind/fzx028.

Easwaran, Kenny. 2016. "Dr. Truthlove or: How I Learned to Stop Worrying and Love Bayesian Probabilities." *Noûs* 50 (4): 816–53. doi:10.1111/nous.12099.

Easwaran, Kenny, and Branden Fitelson. 2015. "Accuracy, Coherence, and Evidence." In *Oxford Studies in Epistemology*, edited by Tamar Szabó Gendler and John Hawthorne, V:61–96. Oxford: Oxford University Press.

Fitelson, Branden. 2016. "Coherence." http://fitelson.org/coherence/coherence_duke.pdf. Unpublished manuscript.

Goldman, Alvin I. 2015. "Reliabilism, Veritism, and Epistemic Consequentialism." *Episteme* 12 (2): 131–43. doi:10.1017/epi.2015.25.

Greco, Daniel. 2014. "A Puzzle about Epistemic Akrasia." *Philosophical Studies* 167 (2): 201–19. doi:10.1007/s11098-012-0085-3.

Horowitz, Sophie. 2014. "Epistemic Akrasia." *Noûs* 48 (4): 718–44. doi:10.1111/nous.12026.

[*] This research was supported by the Social Sciences and Humanities Research Council (grant #756-2019-0133). Thanks to Samuel Dishaw, Branden Fitelson, Daniel Laurier and Justin Snedegar for helpful comments.

JOYCE, James M. 1998. “A Nonpragmatic Vindication of Probabilism.” *Philosophy of Science* 65 (4): 575–603. doi:10.1086/392661.

KIESEWETTER, Benjamin. 2016. “You Ought to ϕ Only If You May Believe That You Ought to ϕ.” *The Philosophical Quarterly* 66 (265): 760–82. doi:10.1093/pq/pqw012.

———. 2017. *The Normativity of Rationality*. Oxford: Oxford University Press.

KOLODNY, Niko. 2005. “Why Be Rational?” *Mind* 114 (455): 509–63. doi:10.1093/mind/fzi509.

———. 2007a. “How Does Coherence Matter?” *Proceedings of the Aristotelian Society* 107: 229–63. doi:10.1111/j.1467-9264.2007.00220.x.

———. 2007b. “State or Process Requirements?” *Mind* 116 (462): 371–85. doi:10.1093/mind/fzm371.

LASONEN-AARNIO, Maria. 2020. “Enkrasia or Evidentialism? Learning to Love Mismatch.” *Philosophical Studies* 177: 597–632. doi:10.1007/s11098-018-1196-2.

LEITGEB, Hannes, and Richard PETTIGREW. 2010. “An Objective Justification of Bayesianism i: Measuring Inaccuracy.” *Philosophy of Science* 77 (2): 201–35. doi:10.1086/651317.

LITTLEJOHN, Clayon. 2018. “Stop Making Sense? On a Puzzle about Rationality.” *Philosophy and Phenomenological Research* 96 (2): 257–72. doi:10.1111/phpr.12271.

PARFIT, Derek. 2011. *On What Matters. Volume One*. Oxford: Oxford University Press. Edited and introduced by Samuel Scheffler.

PETTIGREW, Richard. 2013. “Accuracy and Evidence.” *Dialectica* 67 (4): 579–96. doi:10.1111/1746-8361.12043.

———. 2016a. *Accuracy and the Laws of Credence*. Oxford: Oxford University Press.

———. 2016b. “Jamesian Epistemology Formalised: An Explication of ‘the Will to Believe’.” *Episteme* 13 (3): 253–68. doi:10.1017/epi.2015.44.

SNEDEGAR, Justin. 2018. “Reasons for and Reasons Against.” *Philosophical Studies* 175: 725–43. doi:10.1007/s11098-017-0889-2.

STEINBERGER, Florian. 2019. “Accuracy and Epistemic Conservatism.” *Analysis* 79 (4): 658–69. doi:10.1093/analys/any094.

TITELBAUM, Michael G. 2015. “Rationality’s Fixed Point (or: In Defense of Right Reason).” In *Oxford Studies in Epistemology*, edited by Tamar Szabó Gendler and John Hawthorne, V:253–94. Oxford: Oxford University Press.

WAY, Jonathan. 2010. “The Normativity of Rationality.” *Philosophy Compass* 5 (12): 1057–68. doi:10.1111/j.1747-9991.2010.00357.x.

———. 2013. “Intentions, Akrasia, and Mere Permissibility.” *Organon F* 20 (4): 588–611.

WHITING, Daniel. 2010. “Should I Believe the Truth?” *Dialectica* 64 (2): 213–24. doi:10.1111/j.1746-8361.2009.01204.x.

WORSNIP, Alex. 2018a. “The Conflict of Evidence and Coherence.” *Philosophy and Phenomenological Research* 96 (1): 3–44. doi:10.1111/phpr.12246.

———. 2018b. “What Is (In)coherence?” In *Oxford Studies in Metaethics*, edited by Russ Shafer-Landau, XIII:184–206. Oxford: Oxford University Press.

Review of Soames (2018)

Fraser MacBride

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

A New Vision is the sequel to Soames' *The Analytic Tradition in Philosophy, Volume I: The Founding Giants* (Princeton UP, 2014). *Founding Giants* covered Frege, Moore and Russell. *New Vision* covers Wittgenstein's *Tractatus*, the rise of logical empiricism and its downfall, the advances in logic due to Gödel, Tarski, Church and Turing, Tarski's theory of truth, and contrasting approaches to ethics and meta-ethics in the 1930s. Soames describes his goal as being to identify major insights and achievements, distinguishing them from major errors or disappointments. His declared focus is explication and evaluation of arguments in the texts of Wittgenstein, Carnap *et al*. Thereby Soames conceives of himself as "arguing with the greats" rather than historians of analytic philosophy. He thereby seeks to avoid the perils of antiquarianism which besets history of philosophy when it is bowed down by too much attention to historical-textual detail, whilst his engagement with the secondary literature is sparse.

I do believe that it is possible to do insightful history of philosophy by interrogating dead philosophers as though they were walking amongst us—possible because it's actually been done. Exemplars of this kind of work are Jonathan Bennett's *Kant's Analytic* (1966) and *Kant's Dialectic* (1974), volumes which have stood the test of time, proving fruitful for philosophers and historians of philosophy alike. But I don't think that there's a simple equation which determines that more history, more textual detail means less philosophy—because sometimes more of that is just what's needed to channel the philosophy of our forebears. It's because Soames hasn't done enough to get the history and the texts right that I think he quite often gets their philosophy wrong.

Soames' story in *New Vision* is, as he says, a "complicated" one— understandably so because his aim is to engage directly with the arguments of the greats and they gave a lot of arguments. As a consequence, *New Vision* might better be characterised as a collection of interrogative episodes rather than as an extended dialogue. To provide an impression of the whole, I'm going to evaluate one such episode in which Soames attempts to strike up an argument with Wittgenstein.

In *New Vision* Soames takes Wittgenstein to task for what he describes as "among the darkest and most implausible aspects of the *Tractatus*", Wittgenstein's metaphysics of simples and atomic facts configured from them, ideas which Soames does not consider to have had much interest or influence anyway (Soames 2018, 23). Where does Soames think Wittgenstein went wrong? To be blunt: because Wittgenstein had the ill-fortune to come before Kripke. Soames credits Kripke with the land mark discovery that metaphysical and epistemic modalities needn't march in step but have the potential to diverge, so propositions might be necessary whilst being *a posteriori* and *a priori* though contingent. For Soames this discovery was one of the most remarkable achievements of analytic philosophy in the 20th century. But coming before Kripke, Soames claims, Wittgenstein mistakenly identified necessarily true propositions with propositions knowable *a priori*. According to Soames it's this very mistake, "the notorious tractarian collapse of the modalities", that led Wittgenstein down the false path to his misbegotten metaphysics of simples and atomic facts (Soames 2018, 14).

Wittgenstein famously advanced his atomism by arguing that if there were only complexes all the way down, "then whether a proposition had sense would depend on whether another proposition was true" (2.0211). This would be an intolerable consequence because, Wittgenstein continued, "[i]t would be impossible to form a picture of the world (true or false)" (2.0212). Since it is possible for us to form a true or false picture of the world, Wittgenstein concluded that the analysis of complexes must terminate in absolute simples. Soames reconstructs Wittgenstein's argument along the following lines.

Suppose S_1 is a statement affirming the existence of a complex designated by the logically proper name "*O*". In order for S_1 to "have sense", by which Wittgenstein means be true or false, S_1's constituent expressions, including "*O*", must have meaning. In order for "*O*" to have meaning, *O* must exist. Because *O* is a complex, *O* exists if and only if its parts (*a*, *b*, *c*) are arranged a certain way. Let S_2 be the statement whose constituent expressions include logically proper names for *O*'s parts and which says that *O*'s parts are so arranged. Then whether S_1 has sense depends upon whether S_2 is true. But in order for S_2 to have sense its constituent expressions must have meaning too,which they do only if *O*'s parts exist. Since *O*'s parts are complexes too, they exist if and only if their parts are arranged a certain way. Hence whether S_2 has sense depends upon whether another statement S_3 which says that the parts of *O*'s parts are so arranged is true, and so on without end. Represent this chain of meaning-truth dependencies as an unending sequence:

$$(S)\quad (S_1 \rightarrow S_2), (S_2 \rightarrow S_3), (S_3 \rightarrow S_4), \ldots$$

Now the key interpretative question is why does Wittgenstein take this regress of one sentence's meaningfulness presupposing the truth of another to be vicious? For Soames it's vital to appreciate that this regress presupposes a chain of necessary connections whereby the existence of a complex is analysed in terms of the existence and arrangement of its parts: necessarily *O* exists if and only if *O*'s parts exist and they're arranged a certain way, necessarily *O*'s parts exist iff the parts of *O*'s parts exist and they're arranged a certain way, and so on without end. We can represent this chain as an unending sequence of necessary conditionals:

$$(S_{\Box})\quad \Box(S_1 \rightarrow S_2), \Box(S_2 \rightarrow S_3), \Box(S_3 \rightarrow S_4), \ldots$$

According to Soames, we have seen, Wittgenstein presupposes that necessity and *a priori* knowability coincide. Hence, for Soames' Wittgenstein, ($S_{\Box}$) is equivalent to another non-terminating sequence of *a priori* knowable conditional:

$$(S_{\text{apriori}})\quad \textit{a priori}\text{ knowable }(S_1 \rightarrow S_2), \textit{a priori}\text{ knowable }(S_2 \rightarrow S_3), \textit{a priori}\text{ knowable }(S_3 \rightarrow S_4), \ldots$$

Soames now reasons that if there were no simples "it would follow that *knowing* that ["*O*"] means what it does" and hence knowing the meaning of the sentences in which "*O*" occurs, "would require *knowing* the proposition that *a*, *b* and *c* are composed in the right way" (p. 13). But the same reasoning can be repeated for its parts: "*knowing* that they exist and that propositions about them are meaningful, and have the senses that they do, would require *knowing* the existence of still further objects, as well as the meaningfulness of still further names for those objects and so on without end" (pp. 13–14). Soames concludes: "Thus, if there were no metaphysically simple objects, then one couldn't *know* the meaning of any sentence or perhaps whether it even had a meaning" (p. 14).

Soames' reconstruction of Wittgenstein's argument isn't plausible. Even supposing that ($S_{\Box}$) and (S_{apriori}) are equivalent it doesn't follow that this imposes a requirement upon what must be actually known by a speaker who grasps "*O*". A proposition's being knowable (*a priori* or otherwise) is quite different from its being known – possibility doesn't entail actuality. So even if it is *a priori* knowable that $S_1 \rightarrow S_2$, it doesn't follow that anyone actually knows this, much less that a speaker has to actually know S_2 in order to actually know S_1. Soames supposes that (S_{apriori}) imposes an unending, therefore unsatisfiable set of necessary conditions upon actually knowing that *O* exists. But because (S_{apriori}) covers only the weaker modality of what is knowable, it remains open that a speaker might know S_1 and not know S_2 even if $S_1 \rightarrow S_2$ is *a priori* knowable.

The upshot is that Soames fails to explain how the a priori knowability of $S_1 \rightarrow S_2$ etc. imposes a requirement upon what must be known by someone who understands "*O*". All that Soames establishes is that if there is complexity all the way down, then there is an indefinite potential for unpacking *O*'s complexity, a potential that can be realised by actually coming to know *a priori* $S_1 \rightarrow S_2$, $S_2 \rightarrow S_3$ etc. This might be a surprising view to hold. But since Soames hasn't shown that speakers would have to actually exhaust (*per impossibile*) the potential for unpacking *O*'s complexity in order to grasp "*O*"'s meaning, Soames sheds no light upon Wittgenstein's claim that if there was complexity all the way down, it would be impossible to say something about *O* (or any other object). So it's hard to see that Soames succeeds in striking up a conversation with Wittgenstein rather than talking past him.

Where Soames has gone adrift is failing to factor in Wittgenstein's own insistence that a non-terminating sequence of meaning-truth dependencies would make it impossible to "form", or more literally "draw up" ["entwerfen"], "a picture of the world (true or false)" (2.0212). By "picture of the world (true or false)", Wittgenstein doesn't simply mean "bearer of truth or falsity" but points us further into the interior of the *Tractatus* where a more demanding notion of a proposition and what it is to grasp a proposition awaits us – Wittgenstein's picture theory. It's because a non-terminating sequence of meaning-truth dependencies is incompatible with the possibility of a proposition in this more demanding sense that Wittgenstein concludes that there cannot be complexity all the way down (as I argue in 2018, 188–90).

Let me elaborate briefly upon this alternative interpretation. When we read further into the *Tractatus* we find that a proposition is a complete picture of reality in the sense that when a speaker understands a proposition, they have an exact knowledge of how objects must be arranged for that statement to be true or false and which arrangements of them are thereby left open.

And this is information a speaker can uptake with effortless facility: “The proposition is a picture of reality, for I know the state of affairs presented by it, if I understand the proposition. And I understand the proposition, without its sense having been explained to me” (4.021). So a speaker must already actually know everything she/he needs to know to understand how things must be arranged for a proposition to be true even if the proposition isn’t one she/he has heard before. But a speaker couldn’t have knowledge of what it takes for a proposition to be true (or false) and what is thereby left open if she/he had *per impossibile* to check and see whether a non-terminating sequence of meaning-truth dependencies was satisfied for every expression of their language. A speaker wouldn’t be in a position to know straightaway that the expressions of their language were meaningful but only have a supertask ahead of them. As finite agents, speakers could never confirm that more than an initial segment of the sequence was satisfied, so never be in a position to exercise the consummate facility with language with which Wittgenstein credits speakers.

By contrast to Soames’ account, this interpretation has the merit of making immediate contact with what speakers are required to know to understand a language and it makes sense of Wittgenstein’s argument at 2.0211-2.0212 in the wider context of Wittgenstein’s commitment to the picture theory. It’s a further consequence of this interpretation that what Soames describes as the “notorious Tractarian collapse of the modalities” plays no significant role in Wittgenstein’s argument – Soames’ original mistake was to read the *Tractatus* through “Kripke goggles.”

I have concentrated upon one interrogative episode of *New Vision* to give a representative impression, but I might have taken issue with other episodes where, it seems to me, Soames’ arguments falter for lack of engagement with the historical texts. Consider, for example, his dismissal of the Tractarian conception of a proposition as a propositional sign in its projective relation to the world in favour of his own cognitive act type theory. Or his criticism of the *Aufbau* that Carnap failed to realise that statements expressed in purely logical vocabulary have no empirical content when, Soames has forgotten, “$\exists x \exists y(x \neq y)$” consists of purely logical vocabulary but remains verifiable or falsifiable depending on how many things there are.

Fraser MacBride
University of Manchester
fraser.macbride@manchester.ac.uk

doi: 10.48106/dial.v74.i1.08

References

Bennett, Jonathan. 1966. *Kant's Analytic*. Cambridge: Cambridge University Press.

———. 1974. *Kant's Dialectic*. Cambridge: Cambridge University Press.

MacBride, Fraser. 2018. *On the Genealogy of Universals. The Metaphysical Origins of Analytic Philosophy*. Oxford: Oxford University Press.

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

Review of Oppy (2018)

Mario Schärli

Oppy, Graham, ed. 2018. *Ontological Arguments*. Cambridge: Cambridge University Press.

A shadow of criticism has followed ontological arguments for almost a thousand years and counting. Irrespectively, the arguments continue to intrigue philosophical thought, and no decline is in sight. In particular, modal versions of the argument formulated by Hartshorne, Lewis, Plantinga and Gödel in the 1960's and 70's have helped to dispel the widely held suspicion that a simple logical blunder lies behind ontological arguments. As a result, recent discussions have shifted from assessing the argument's validity towards its soundness and dialectical efficacy. This requires engaging with the philosophical issues inevitably raised by the argument, such as questions about the nature of concepts and arguments, existence and possibility. These have since stood at the forefront of the debate.

The concerns united by reference to ontological arguments form the subject matter of a recently published volume edited by Graham Oppy, himself one of the most prolific authors on the topic in the past 25 years. His informative introductory essay underlines important differences between the arguments commonly called "ontological." Oppy suggests abandoning the search for unity suggested by the description "the ontological argument." Instead, the commonalities should be viewed genealogically: "What is distinctive of ontological arguments is that their formulation has the right kind of connection to Anselm's argument" (p. 11). Hence, fruitful engagement with and criticism of ontological arguments proceeds by cases.

This sets the tone for the volume's first group of articles which are devoted to defenders and critics of the ontological argument, namely: Anselm, Aquinas, Descartes, Leibniz, Kant, Hegel, Gödel, Lewis, Plantinga, and Tichý. A second group of three essays dealing with overarching systematic issues surrounding the preceding arguments complements the volume. Here we find treatments of the relation between conceivability and possibility, the "fallacy" of begging the question, and the relation between existence, characterization and modality.

Overall, the volume provides readers with informative up-to-date discussions of ontological arguments of scholarly value by senior researchers in the field. (With the notable exception of M. Inwood's article on Hegel which lacks engagement with the literature on the subject.) At the same time, the essays are written in an accessible manner, rendering the volume suitable as an accompaniment to graduate-level courses on the subject. Due to limitations in space, I will refrain from summarizing and discussing all the contributions. For that purpose, Oppy's introduction (pp. 2–5) is well suited. Instead, I will focus on three contributions I found particularly worth discussing.

The majority of ontological arguments treated in the volume—Anselm's, Leibniz's, Gödel's, Plantinga's—are shown to be deductive in nature by their interpreters. A noticeable rift opens up between them and Descartes' argument, according to Lawrence Nolan's interpretation. His article represents an important scholarly contribution because it virtually reverses the standard deductive reading, and plausibly so.[1] Developing a suggestion hinted at by M. Gueroult and J. Barnes, Nolan interprets Descartes' so-called ontological argument as "the report of an intuition in the sense of a non-discursive, self-validating, intellectual apprehension" (p. 54). The aims Descartes pursues with the argument are persuasive rather than argumentative: all he points to serves the purpose of getting the meditator to have the relevant intuitive insight.

A strength of Nolan's reading is that it allows us to make good sense of passages (e.g. *Med.* V, AT VII 68–69 and *Princ.* I., §15), where Descartes clearly glosses the cognition of God as an intuitive insight; these have always been difficult to accommodate within deductive interpretations of the argument. Moreover, Nolan convincingly shows that his reading coheres with Descartes' skepticism towards a formal-deductive understanding of reasoning voiced in the *Rules* as well as the other philosophical doctrines he adheres to (pp. 57–65). However, the intuitive reading of the argument has to confront the following difficulty: what about the passages where Descartes overtly argues in a deductive manner?

Nolan uses two principal interpretive moves to provide a coherent picture in these cases (pp. 54, 66–71). First, he convincingly shows that the overtly argumentative passages, commonly taken to be Descartes' argument, are best read as rebuttals of possible criticisms. They allow the meditator's intuition not to be distracted by an unjustified conception, e.g. by understanding the

[1] Cf. also his earlier "The Ontological Argument as an Exercise in Cartesian Therapy" (2005).

distinction between essence and existence as a real rather than merely rational distinction. Second, he argues that, for historical reasons, Descartes aimed to present his philosophy in a manner adjusted to the scholarly discourse of his day, which put great emphasis on the syllogistic demonstrability of God's existence.

While the latter may be correct as a matter of historical fact, Nolan's line of interpretation may be bolstered by a more penetrating understanding of the relation between intuition and deduction. It is Descartes' view that deduction is necessary in case one does not have clear and distinct, intuitive insight at one's disposal (p. 61). Although this legitimizes ascribing priority to intuitive over deductive insight, it does not imply a merely historical explanation of the occurrence of deduction. Rather, one might—in line with Descartes' general manner of proceeding in the *Meditationes*—explain the deductive arguments as necessary steps towards intuitive insights. It helps to take into consideration what the condition for distinctly perceiving a given content is: being able to tell it apart from others (*Princ.* I., §45). If that is the case, then the arguments delivered to fend off criticisms are not merely negative or persuasive, but positively contribute to the distinctness of the meditator's perception and thus to its intuitiveness. This should not be understood as a criticism, but as additional support for Nolan's reading. In my view, Nolan's essay represents a lasting contribution to our understanding of Descartes. Moreover, it points to a version of the ontological argument that might merit systematic development in the light of recent advancements in the epistemology of intuition.[2]

Other than authors who defend the ontological argument, the volume features some of its most important critics in Aquinas, Kant, and Lewis. Among these, L. Pasternack's perceptive and well-informed article on Kant is one of the best discussions currently available. It sets the record straight on the nature of Kant's case against the ontological argument. Contrary to popular wisdom, the latter extends well beyond the familiar line "existence is not a real predicate". Pasternack distinguishes two main strands of criticism within Kant's argumentation in the *Critique of Pure Reason*, the first of which targets an analytic, the second a synthetic reading of the judgment "God exists" (p. 102). Kant argues that an ontological argument insisting on the analytic reading of the statement is dialectically flawed, i.e. does not add up

[2] First and foremost: Chudnoff (2013).

doi: 10.48106/dial.v74.i1.09

to an argument at all (pp. 104, 106)[3], while a synthetic reading rests on the thesis that existence is a "real predicate" which Kant disputes (pp. 106–115).

The soundness of the second part of Kant's criticism rests on an argument against existence being a real predicate, which Pasternack deems inconclusive. His rendering is as follows: if existence were a property of objects, then a concept specifying all the properties of the object, but lacking existence as a mark, would fail to fully articulate the object in question, leading to a "mismatch" between concepts and their objects. This argument is unconvincing for two related reasons. First, it leaves open why this mismatch should be deemed problematic according to Kant, which needs to be established for the argument to be sound. Pasternack appears to agree on this point, which leads to the second weakness of the argument: it is susceptible to the "obvious rebuttal" Pasternack puts forward. Basically, it consists in making the mismatch disappear by allowing existence to be a mark of concepts (p. 114).

However, a more convincing reading of Kant's argument is possible. Immediately after the passage Pasternack quotes in support of his reading, Kant writes: "Even if I think in a thing every reality except one, then the missing reality does not get added when I say the thing exists, but it exists encumbered with just the same defect as I have thought in it; otherwise something other than what I thought would exist" (A600/B628). It emerges from this sentence that the alleged "mismatch", i.e. a concept's not fully capturing all the properties its instances exhibit, is not what is at issue, at least as far as Kant perceives matters. On the contrary, his point concerns instantiation, or the relation between concepts and objects, in general. This addresses the first weakness of Kant's case as interpreted by Pasternack. But what is Kant's point then?

Kant argues that a concept's instantiation does not correspond to any *addition* of properties to it; rather, a concept's instantiation amounts to the object's having *just* the properties the concept specifies. A plausible way of construing this claim is: A concept's content consists in the conditions an object has to meet in order to count as an instance of it. Kant can be understood as showing that this view cannot be upheld if one understands existence as a property. The reasoning can be understood as follows. If existence were a property of objects and being an instance of a concept is to exist, then a concept's instances would

[3] It is, of course, a common criticism of ontological arguments that they are question-begging; e.g. Aquinas raises a similar point according to B. Leftow's reconstruction (pp. 47, 49, 51) and P. van Inwagen discusses the issue concerning the modal ontological argument in his contribution (pp. 238–249).

consequently have to bear the property "existence". If "falling under a concept" consists in an object's *conforming* to the conditions set by the concept and existence is one of these conditions, then existence would have to be a mark of the concept. But this would render some existence-judgments analytic—a view Kant takes himself to have refuted at this point in the discussion. If one grants this, it follows that existence is not a mark of any concept. But if it still holds that instances of concepts exist, then a concept's instantiation consists in an object's conforming to the conditions set *and* exhibiting the property "existence" *additionally*. As the latter is not part of a concept's content, this content's identity therefore cannot consist in a specification of what it takes to be its instance, no matter how completely or incompletely it captures an object's properties.

Kant therefore does not argue that a mismatch between concepts and objects is problematic as such, but that a specification of a concept's content in terms of conditions instances have to meet is impossible given that one accepts the following three theses:

(1) a concept is individuated by the conditions on objects to count as instances of it;
(2) existence is a property of objects;
(3) existence-judgments are synthetic.

Kant's point therefore is: the view that existence is a property is indicative of a misunderstanding of what concepts and "falling under a concept" are. Pasternack's rebuttal misses the mark in relation to this issue, for accepting existence as property and conceptual content *either* leads to the implausible view that all existence-judgments are analytic *or*, if they remain synthetic, precludes a conception of concepts as specifying the conditions of what it is to fall under them.

Alongside "the usual suspects", Pavel Tichý's work on the ontological argument makes an unexpected appearance in the volume. As is convincingly shown by G. Oddie's essay, Tichý offers one of the most penetrating and revealing interpretations of Anselm's *Proslogion* III, i.e. the passage serving as inspiration for what is known as "the modal ontological argument". Tichý delivers a logically valid reconstruction of Anselm's argument as well as an unfamiliar axiological criticism of its soundness.

The reading rests on Tichý's ontology, fundamental to which is the distinction between "two entirely different *types* of entity" (p. 199): *individuals*

doi: 10.48106/dial.v74.i1.09

(such as Donald Trump) and *offices* (such as "the President of the U.S.A"). Intuitively, offices are either occupied by an individual or not, where occupancy is to be understood as a property of the office rather than the individual. Formally, offices are partial functions mapping world-time pairs to individuals which are undefined when the office goes unoccupied. What an office is—its essence—is given as a set of conditions called *requisites* which have to be borne by occupants to count as such (p. 203).

Within this framework, the modal ontological argument aims to derive the necessary occupation of "the divine office" (p. 205), which Tichý interprets as "*that individual office such that no individual office is greater than it*" (p. 206). Anselm's formula, thus understood, singles out a second-order office, that is, an office occupied by a first-order office rather than an individual. Glossing over the details of the reconstruction, the *Proslogion* III argument derives necessary existence as a requisite of the greatest office, yielding the conclusion that the divine office is necessarily occupied. This yields a "valid" argument according to Oddie (p. 209).

Compared to the standard modal ontological argument known from the writings of Harthshorne and Plantinga, Tichý's interpretation of the argument has one key advantage. The standard version treats existence in all possible worlds as an essential property of God and derives God's existence from His/Her possible existence plus S5. The argument is often criticized for begging the question because the premise that God's existence is possible cannot be substantiated in a non-circular fashion. G. Priest's offers one way of putting the difficulty (p. 265).[4] According to the premises of the argument, the following two entailments hold: (1) God's actual existence follows from His/Her possible existence; (2) God's actual existence entails His/Her possible existence. "God exists" and "possibly, God exists" are therefore equivalent according to the argument's premises. Hence, presupposing the possibility of God's existence is question-begging insofar as it is equivalent to presupposing God's existence. By contrast, Tichý's reconstruction derives the necessary occupancy of the divine office via an axiological premise, namely: necessarily occupied offices are always greater than ones which are not (p. 208). The truth of this premise can be assessed independently, and hence God's necessary existence gets established in a more satisfactory way.

[4] Ways of stating and resolving the difficulty are discussed in the articles of J. Spencer, J. Rasmussen, P. van Inwagen in the volume. Rasmussen tries to make progress on the issue by providing an independent argument for God's possibility turning on the modal properties of value (pp. 183–185). I find his argument unconvincing, but due to limitations in space, I cannot give my reasons here.

However, this premise also renders the argument unsound according to Tichý. The claim that necessarily occupied offices are always greater than ones which are not is subject to counterexamples, one of which is: the office "*the discoverer of the incompleteness of arithmetic*" is contingently occupied, whereas the office "*the pick of the morally most depraved*", where "pick" refers to a choice function to be applied in case of a tie, is necessarily occupied. Yet, the former is plausibly "greater" than the latter (p. 212). Therefore, Oppy concludes with Tichý, the argument rests on an implausible axiology of existence. Attempts at weakening the relevant requisite (e.g. either being God or else the pick of the morally best) will, while ending up necessarily occupied, fail to prove the existence of God at all world-times, for God is not merely the relatively morally best being, but the absolutely best (pp. 212–213).

Oddie's simultaneously fascinating and accessible discussion of Tichý's reconstruction will hopefully lead to the recognition of what strikes me as the most convincing version of a *Proslogion* III-style modal ontological argument. Further discussion may delve deeper into the axiological questions raised by Tichý. As is always the case with criticisms resting on counterexamples, they may show *that*, but not explain *why*, some thesis is false. What principled reason against Anselm's axiology can be given?[*]

Mario Schärli
University of Fribourg
mario.schaerli@unifr.ch

References

Chudnoff, Elijah. 2013. *Intuition*. Oxford: Oxford University Press.

Nolan, Lawrence. 2005. "The Ontological Argument as an Exercise in Cartesian Therapy." *Canadian Journal of Philosophy* 35 (4): 521–62. doi:10.1080/00455091.2005.10716601.

[*] Research going into this article has been funded by the Swiss National Science Foundation, Grant No. P1FRP1_188051. I am grateful to Anna Katsman and Jelscha Schmid for critical remarks on a draft version of this article.

doi: 10.48106/dial.v74.i1.09

Review of Antonelli (2018)

Hamid Taieb

Antonelli, Mauro. 2018. *Vittorio Benussi in the History of Psychology: New Ideas of a Century Ago*, Studies in the History of Philosophy of Mind 21, Cham: Springer.

The history of phenomenology has not been a peaceful and autonomous process taking place independently of any competitors. On the contrary, from the very beginning of their inquiries, phenomenologists had to struggle with several rival explanatory schemes in psychology. The most important among them were physiological psychology (of various sorts) and psychoanalysis. Both of these scientific projects tried to minimize the importance of consciousness in the explanation of the mind, the first by treating consciousness as some sort of epiphenomenal outcome of brain and other nervous processes, the second by describing it as a blind domain, driven by underlying mental acts to which consciousness itself has no access. Interestingly, however, phenomenology did not ignore these two competing explanatory schemes; on the contrary, it entered into manifold discussion with them, trying to establish more and more precisely the "division of scientific labour" among these three approaches. Evidence of this engagement is plentiful. With respect to physiological psychology, the discussion goes as far back as Franz Brentano, who tried to combine his "descriptive psychology", also called "descriptive phenomenology", with "genetic psychology", that is, physiological psychology; and it has had a long and complex history, up to the most recent papers published in the journal *Phenomenology and the Cognitive Sciences*. With respect to psychoanalysis, phenomenologists such as Merleau-Ponty and Ricœur engaged in detail with the thought of Freud (who, by the way, had been a student of Brentano); there have also been more recent attempts to combine these two traditions, for example by Lohmar (2012). However, as shown by Mauro Antonelli, the first ecumenical hero in this history, who combined in a harmonious way all three disciplines—that is, phenomenology, physiological psychology, and psychoanalysis—was Vittorio Benussi.

In reading Antonelli's book, one comes to realize that Benussi, who is described as an "*Einzelgänger*" (p. 238), is a figure as important as he is unknown. Antonelli very nicely combines detailed analysis of Benussi's philosophy of mind with description of the historical and scientific background in which Benussi developed his work. Benussi's life was rich, but also "tragic", as Antonelli emphasizes. Born in Trieste in 1878, Benussi moved to Graz at the age of 18, where he studied with and was influenced by Meinong, and through him by Brentano, with whom Meinong had studied. In Graz, Benussi did not have a permanent academic position: he was a temporary assistant in Meinong's psychology laboratory and worked at the university library to earn enough money to live; but with access to Meinong's laboratory, becoming even its "*de facto* director" (p. 112), he developed his own research agenda. After Trieste was absorbed by Italy following the First World War, he became an Italian citizen, and as a result he lost his position as a librarian in Graz, and was forced to move to Padua. He then fell into a deep depression, despite being hired as a professor at the University of Padua soon after arriving in the city. He committed suicide in 1927 at the age of forty-nine by drinking cyanide, just as in a dream years earlier.

After a short but useful introduction (ch. 1), which explains the *raison d'être* for a monograph on Benussi, Antonelli presents the state of the art in psychology in the German-speaking world at the end of the 19th century and provides a brief overview of Brentanian and Meinongian philosophy and psychology (ch. 2). Following these helpful chapters of contextualization, and a biographical sketch of Benussi (ch. 3), Antonelli enters into the details of Benussi's work and impressive research program. Benussi is mostly known for having developed a theory of *Gestalt*. He was a member of the so-called "Graz School" of *Gestalt* theory, which was opposed to the "Berlin School" of Wolfgang Köhler and his associates. *Gestalten* are, roughly speaking, complex but unitary entities based on a series of elements, to which, however, they are not reducible; for example, a melody is a *Gestalt*, which is based on but not reducible to the series of sounds that compose it. Benussi emphasized the importance of subjective activity in the production of *Gestalten*, whereas the Berlin School had an objectivist account of them (see ch. 4.3 and 4.6, which present in detail Benussi's views, including his evolution on the topic, due in part to objections from the Berlin Gestaltists). However, as clearly shown by Antonelli, Benussi's research extended far beyond Gestalt theory; among the topics on which he worked were the classification of mental acts, the distinction between intentional content and object, sensory illusions, judgments and "assumptions" (or "pseudo-judgments"), the theory of "productive presentations" (which explains, among others things, the constitution of *Gestalten*), the relation between emotions and cognition, and time-consciousness (ch. 4.4); beyond these rather classical themes of Brentanian and Meinongian psychology (ch. 4.2), but also of Würzburgian *Denkpsychologie*, another source of inspiration for him, Benussi worked on testimony, including lie detection (ch. 4.7), unconscious mental phenomena, including their relation to dreams (ch. 4.5 and 5.4), and the influence of the body on emotions (ch. 5.2), as well as mental analysis (ch. 5.1) and hypnosis (ch. 5.3), these themes being mostly develop in his later, Padua period, perhaps due to the fact that he had no laboratory allowing him to continue his work on sensation and *Gestalt* (p. 261). On all these themes, the reader will find original and highly interesting developments, due first to Benussi's careful experimentations and analyses, founded on methodological reflections about psychology and its relation to philosophy (ch. 4.1), and second to Antonelli's clear and detailed reconstruction, made possible by an impressive knowledge of Benussi's work, including his *Nachlass* (which is presented at the end of the volume, along with a bibliography and a list of the lecture courses that Benussi delivered at the universities of Graz and Padua), and by a rare sense of synthesis. The Conclusion (ch. 6) shows that Benussi's work could be applied to draw interesting connections between phenomenology and enactivism on the one hand, and contemporary neurosciences, biology, and pragmatics on the other.

Obviously, it is impossible in this review to address all of the topics listed above. I would like to focus on one aspect of Benussi's work, namely, his account of emotions, which will also be the occasion to discuss some crucial methodological points that he defends about psychology. In the Brentanian tradition, an important psychological thesis, which is not based on any empirical-inductive generalization, but is meant to be an *a priori* truth, is that no emotion can take place without an underlying presentation: emotions are all *about* something, or have an object, and this object is provided to them by a presentation on which they, thus, depend. Interestingly, this thesis is attacked by Benussi, who holds explicitly that such a view is a mere philosophical *speculation* (pp. 277–278). His position is based on specific empirical findings, as he wanted psychology to rely on experience, and thus adopted a "theoretical minimalism" (pp. 145–147, Antonelli quoting an expression from Sadi Marhaba); in this respect, according to Antonelli, Benussi's approach is to be placed somewhere between the philosophical phenomenology of Husserl and the experimental phenomenology of Stumpf (p. 320).

doi: 10.48106/dial.v74.i1.10

What then was Benussi's empirical ground for his thesis of the non-intentionality of emotions? He applied his "analytic" method in psychology, the idea being that the mental life is a "harmonious coordination of autonomous elementary functions" (as Benussi puts it) that one can "disarticulate", pretty much on the model of vivisection (p. 262). One of the tools that Benussi used for performing these vivisections was hypnosis. Now, one state to which he was able to lead the persons on whom he was testing his hypotheses was that of "basic sleep", a state in which, supposedly, subjects had their "conscious intellectual life" interrupted while being still able to have some specific feelings. Once put in these states, the subjects were suggestible, and Benussi would invite them to have specific emotions, such as hate. When they came back to consciousness, they were asked to report what they experienced. Now, according to their testimonies, they did indeed experience specific emotions such as hate, but given the absence of intellectual awareness these emotions were deprived of any object (p. 278). In fact, the test subjects reported a series of "kinaesthetic and muscle sensations", which Benussi apparently took to be constitutive of emotions. All this was proof, for Benussi, that intentionality is not necessary to emotions, and thus that the philosophical thesis that emotions are based on an underlying presentation is speculative. Note that Benussi defended the view that emotions might be intimately linked with an "organic-visceral sensitivity" (as Antonelli puts it, p. 315), to the extent that they might be generated by viscera and other organs, including the lungs (pp. 303–304); as such, they would be the product of a "physiological unconscious" (p. 316). Benussi was thus connecting the mind closely to the body, and through it to the evolution of the species; in this, as Antonelli emphasizes, Benussi anticipated various contemporary theories, notably those of Antonio Damasio and Jaak Panksepp, and evolutionism more broadly.

These considerations about emotion are particularly interesting, as Benussi's views anticipate various contemporary hypotheses and debates. They also seem to develop an account of emotions very much like that of William James, for whom emotions are feelings of bodily processes. Now, in contemporary philosophy, the Jamesian account of emotions has been challenged in favour of a model which defends the intentionality of emotions. (For a good overview on contemporary theories of emotions, see Scarantino and DeSousa 2018.) It would be interesting to compare Benussi's views on emotions with those of contemporary philosophers, which Antonelli does not do, despite his general willingness to make such comparisons with more recent thinkers.

Independently of this, however, a question that is raised by this theoretical conflict about emotions is that of the delimitation of the scope of Benussi's research. Benussi criticizes speculative approaches to the philosophy of mind and praises empirical inquiries. However, the people with whom Antonelli compares him—not just Husserl, but also Brentano and Stumpf—all agree on one important point: they admit *a priori* truths in philosophy of mind, and they are very careful — especially Husserl—to distinguish this "eidetic phenomenology", which is about the nature or essence of mental acts and states, from empirical psychology, which is devoted to the study of the mental life of a determinate natural species (e.g. human beings). Benussi's attraction to empirical research might have led him to neglect this distinction too much. Indeed, the distinction does not play a major role in Antonelli's book. Keeping this distinction in mind, however, leads to a more accurate determination of the scope of one's psychological research, since it allows one to distinguish in one's inquiries between what belongs to a mental phenomenon as such, and what belongs to it insofar as it is implemented in a certain kind of living being. This might have important consequences for the way one describes and understands a given phenomenon. As regards emotions, couldn't one say that the feelings Benussi is pointing to are not themselves the emotion of, say, hate, but merely some bodily impressions that human beings contingently co-experience while feeling hate? In that case, what Benussi's subjects are reporting are these feelings, which they confuse with hate properly speaking simply because they are concomitant, while hate as such, by its very nature or essence, has another structure, being object-directed.

Such interrogations can be extended to all dimensions of psychology, and were in fact extended in this way by Husserl and others. As Antonelli shows, Benussi developed, in parallel to Husserl, a genetic phenomenology which studies how the subject passively and unconsciously constitutes the identity of perceptual objects despite constant perceptual variations, organizes the perceptual field, produces *Gestalten*, etc. But here too, Husserl pointed out the possibility of an *a priori* knowledge, since these processes have their own essential rules, which are independent of being instantiated in this or that natural species (see e.g. Husserl's *Passive Synthesis*, Hua 9, 121.34–123.28, and Elmar Holenstein's (1972, 22–25) study on association of ideas in Husserl). In sum, a question that remains open when reading Antonelli's book, in the discussion of emotions and elsewhere, is whether Benussi's criticism of "speculative" philosophy goes too strongly in the opposite direction, by blurring an important distinction found among other phenomenologists of

doi: 10.48106/dial.v74.i1.10

his time. And behind this question is the more fundamental one of whether it is legitimate to accept something like a "philosophical psychology" which supposedly has its own proper task that is distinct from that of empirical psychology. Perhaps Benussi underestimated the importance of this issue.

But these reflections should not distract us from the most important point: Antonelli's book is a fascinating, well-informed, and admirably clear study which should be read by everyone interested in the history of psychology and phenomenology. It also extends the canon in the philosophy of mind by rehabilitating an unduly neglected figure who managed to combine, long before others, the theoretical insights of phenomenology, physiological psychology, and psychoanalysis. There is no doubt that Vittorio Benussi's theoretical project remains highly relevant.

Hamid Taieb
University of Hamburg
hamid.taieb@hu-berlin.de

References

DeSousa, Ronald B. 2013. "Emotion." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta. Stanford, California: The Metaphysics Research Lab, Center for the Study of Language; Information. Revision, January 21, 2013, of the version of February 3, 2003.

Holenstein, Elmar. 1972. *Phänomenologie der Assoziation. Zu Struktur und Funktion eines Grundprinzips der passiven Genesis bei E. Husserl*. Dordrecht: Springer Netherlands.

Lohmar, Jagna, DieterandBrudzinska, ed. 2012. *Founding Psychoanalysis Phenomenologically. Phenomenological Theory of Subjectivity and the Psychoanalytic Experience*. Dordrecht: Springer Netherlands.

Scarantino, Andrea, and Ronald B. DeSousa. 2018. "Emotion." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta. Stanford, California: The Metaphysics Research Lab, Center for the Study of Language; Information. https://plato.stanford.edu/entries/emotion/. Version of September 25, 2018; new version of DeSousa (2013).

Published by *Philosophie.ch*

Verein philosophie.ch
Fabrikgässli 1
2502 Biel/Bienne
Switzerland
dialectica@philosophie.ch

https://dialectica.philosophie.ch/

ISSN 1746-8361

Dialectica is supported by the Swiss Academy of Humanities and Social Sciences.

Abstracting and Indexing Services

The journal is indexed by Arts and Humanities Citation Index, Current Contents, Current Mathematical Publications, Dietrich's Index Philosophicus, IBZ — Internationale Bibliographie der Geistes- und Sozialwissenschaftlichen Zeitschriftenliteratur, Internationale Bibliographie der Rezensionen Geistes- und Sozialwissenschaftlicher Literatur, Linguistics and Language Behavior Abstracts, Mathematical Reviews, MathSciNet, Periodicals Contents Index, Philosopher's Index, Repertoire Bibliographique de la Philosophie, Russian Academy of Sciences Bibliographies.

Contents