dialectica

International Journal of Philosophy

Contents

EUROPEAN SOCIETY FOR ANALYTIC PHILOSOPHY

Vol 76 : 4
December 2022

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

Editorial Board

Jérôme Dokic, EHESS, Paris, France
Pascal Engel, EHESS, Paris, France
Manuel García-Carpintero, Universitat de Barcelona, Spain
Diego Marconi, Università di Torino, Italy
Carlos Moya, Universitat de Valéncia, Spain
Martine Nida-Rümelin, Université de Fribourg, Switzerland
François Recanati, Collège de France, Paris
Marco Santambrogio, Università degli Studi di Parma, Italy
Peter Simons, Trinity College Dublin, Ireland
Gianfranco Soldati, Université de Fribourg, Switzerland
Marcel Weber, Université de Genève, Switzerland

Editors

Fabrice Correia, University of Geneva
Philipp Blum, University of Lucerne
Sharon Casu, University of Fribourg (managing editor)
Marco R. Schori, University of Bern (managing editor)

Review Editors

Stephan Leuenberger and Philipp Blum

Editorial Committee

Philipp Blum (né Keller), Claudio Calosi, Sharon Casu, Zoé Christoff, Fabrice Correia, Matthias Egg, Martin Glazier, Aleks Knoks, Arturs Logins, Jörg Löschke, Giovanni Merlo, Robert Michels, Ryan Miller, Jonathan Mitchell, Paolo Natali, Donnchadh O'Conaill, François Pellet, Edgar Phillips, Jan Plate, Stephanie Rennick, Thomas Schindler, Sebastian Schmidt, Marco R. Schori, David Schroeren, Mike Stuart, Daniel Vanello.

Consulting Board

Johannes Brandl (Salzburg), João Branquinho (Lisboa), Elke Brendel (Bonn), Ingar Brinck (Lunds), Eros Corazza (Ikerbasque and Carleton), Josep Corbi (Valéncia), Michael Esfeld (Lausanne), Dagfinn Føllesdal (Stanford and Oslo), Frank Jackson (Australian National University, Canberra), Max Kistler (Paris I), Max Kölbel (Wien), Jan Lacki (Genève), Karel Lambert (Irvine), Paolo Leonardi (Bologna), Fraser Macbride (Manchester), Josep Macià (Barcelona), Genoveva Martí (Barcelona), Élisabeth Pacherie (Institut Jean Nicod, Paris), David Piñeda (Girona), Wlodek Rabinowicz (Lund), Barry Smith (Buffalo), Christine Tappolet (Montréal), Neil Tennant (Ohio State), Mark Textor (King's College London), Achille Varzi (Columbia University), Alberto Voltolini (Torino), Timothy Williamson (Oxford).

dialectica

vol 76
issue 4

December 2022

Contents

Responsibility First: How to Resist Agnosticism about Moral Responsibility

László Bernáth & Tamás Paár

We argue against the view that one should suspend belief in the existence of moral responsibility. We start with a simple argument based on the claim that the existence of obligations entails the existence of moral responsibility. If this is true, then agnosticism about moral responsibility is incoherent. However, this simple argument is insufficient. It can be repaired by focusing on agents who rationally believe in a particular conception of obligation (the "Responsibility First View" (RFV)). On that conception, non-moral obligations that are not appropriately related to moral obligations can be freely ignored and the property of being morally responsible is identical to the property of fulfilling all necessary conditions for bearing moral obligations. Those agents who rationally hold RFV can still rationally believe in moral responsibility even if they lack direct evidence for the existence of moral responsibility.

Even if we lack evidence for the existence of moral responsibility or if scientific research makes it unlikely that moral responsibility is real, one can still rationally maintain belief in it as long as one adopts a specific view of moral obligations (the "Responsibility First View"). Or so we will argue.

We first outline the case for agnosticism about responsibility (section 1), then we sketch a simple objection against it (section 2) and the reasons why the simple objection fails (section 3). Next, we outline the Responsibility First View (section 4) and we reformulate the case against agnosticism in a form that is not subject to the earlier difficulties (section 5). Finally, we consider free will, fairness and the circle of responsible agents.

1 Agnosticism about Moral Responsibility

There are three basic epistemological stances about moral responsibility. Some believe that normal human adults are often morally responsible, others deny that we know they are. The second approach, in turn, has two distinct versions. Members of the first group maintain that nobody is ever morally responsible (see Strawson, G. 1994; Pereboom 2001; Levy 2011), and they imply that this belief is justified for all of us. Members of the second group argue that we don't have enough evidence to tell. Our evidence is not decisive with regard to the existence of moral responsibility. Call philosophers who belong to this group "agnostics about moral responsibility." Typically, they suspend belief in the existence of moral responsibility, and they think that others should join them in doing so. A number of philosophers have put forward arguments to (roughly) that effect. They do not explicitly deny the reality of responsibility, but they argue that our most popular (compatibilist or incompatibilist) theories of moral responsibility make it unlikely that we could tell whether anyone is morally responsible (Byrd 2010, 2021; Sehon 2013, 2016; Kearns 2015).[1]

[1] Note that although the papers cited here are centered around arguments that seem to support responsibility agnosticism, there are important differences between the accounts of the authors. Byrd (2010) embraces agnosticism about moral responsibility, claiming that the current debate should lead us to accept that we do not *now* know whether anyone is ever morally responsible. Nevertheless, in Byrd (2021) he seems to mitigate the strength of his argument, arguing in his conclusion that even if at present agnosticism about free will is reasonable, there seems to be hope that we can overcome our ignorance *in the future*. Kearns argues mostly for agnosticism about *free will*, but he also believes that "the thesis that we (don't) have free will [...] entails the moral claim that we are (not) morally responsible" (2015, 249), hence what he says amounts to a version of moral responsibility agnosticism too. What he insists is that we do not *know* whether there is free will or not, not that we are *unjustified* in believing it, nevertheless, by "not know" he means that our justification is so weak that it doesn't even meet a "low standard" (2015, 236). But having such a weak justification in a given question might very well warrant suspending belief. Sehon (2013, 2016) seems to be the furthest from the position of these agnostics, as he develops a certain variant of compatibilism in order to counter his own challenge against belief in moral responsibility. However, there is a good reason to consider and also to answer the agnostic *arguments* of these three authors together. The reason is that their agnostic arguments are logically independent from answers that they might come up with answering or at least mitigating moral responsibility agnosticism itself, as Sehon (2016) does. Naturally, one could be consistent in accepting their arguments that support agnosticism about moral responsibility while rejecting ways they may propose to evade this kind of agnosticism. (This is why it is no surprise that Sehon 2013 basically employs agnostic arguments, without offering a solution.) Therefore, one can scrutinize these arguments independently of the full-blown theories of the aforementioned authors. It is worth noting that their ways to avoid the agnostic conclusion seem

Those who try to resist agnosticism seek to show that we do have sufficient evidence—be it moral (van Inwagen 1983, 206–223; Coffman 2016), phenomenological (Guillon 2014), conceptual (Latham 2019), transcendental (Lockie 2018), or practice-based (Strawson, P. F. 1969)—to make our belief in the reality of moral responsibility justified. Agnostics, however, retort that such pieces of evidence are unreliable and open to objections. In this paper, we do not engage with that debate.[2] Rather, for the sake of argument, we take it for granted that there is no sufficient evidence for the reality of moral responsibility, and we try to show that one can rationally attribute moral responsibility to herself and others even in that case. Those who prefer to argue against responsibility agnosticism more directly, and believe that there is sufficient evidence in favor of moral responsibility, may still welcome our argument as an additional way to counter the agnostic.

The argument for agnosticism can be formulated in the following way:

(**AR1**) Nobody is justified in believing that the metaphysical conditions of moral responsibility are ever satisfied.

(**AR2**) If you are not justified in believing that a necessary condition of *X*'s existence is satisfied, then you are not justified in believing that *X* exists.

Therefore,

(**AR3**) Belief in the existence of moral responsibility is unjustified.

On a common interpretation of justification, the following principle is true:

to be controversial. For example, Sehon (2016) offers a non-standard, non-causal and at present unpopular blend of compatibilism that has met with serious criticism (Mele 2019). If this kind of criticism is correct, the argument of this paper may still be sound, as it offers a different way to counter agnosticism about moral responsibility. Furthermore, even if the ways to respond to moral responsibility agnosticism that are suggested by those who themselves employ agnostic arguments might work, our argument does not lose its significance: it is an additional way to counter the kind of agnosticism in question that could strengthen belief in moral responsibility even more.

[2] It may be worthwhile to point out that our argument is somewhat akin to the transcendental arguments for free will and responsibility such as Robert Lockie's recent arguments (2018). For example, like Lockie's argument, we argue for the rationality of believing in moral responsibility is based on some analysis of conditions for bearing obligations. However, transcendental arguments aim to show that all rational (human) agents should believe in free will and responsibility (and Lockie's transcendental arguments share the same ambitions) whereas the argument we present attempts to show only that some agents can rationally believe in moral responsibility.

S. We should suspend those of our beliefs that are not justified.[3]

And so one can conclude that

ARC. We should suspend belief about the existence of moral responsibility.

The agnostic ascribes an epistemic obligation to those who assess the evidence regarding the existence of moral responsibility. The core intuition of our paper is that there is a serious tension between suspending beliefs about moral responsibility and ascribing epistemic obligations to oneself and others—intuitively, someone who is not morally responsible cannot have obligations. One can argue for this in two ways. First, one could argue that no one can be obliged to suspend belief about moral responsibility. Alternatively, one can say that holding a specific conception of moral responsibility makes it irrational to believe in obligations to suspend belief about moral responsibility. In the next section, we explore the first idea in order to see if a simple and intuitive argument could support it.

2 The Simple Objection

It might seem *prima facie* plausible that moral responsibility is a precondition of having obligations. If it is, then ascribing obligations while suspending belief about moral responsibility is irrational.

Why is it plausible that moral responsibility is a precondition of having obligations? One could appeal to the idea that only morally responsible agents can have normatively binding obligations. Consider the following argument:

THE SIMPLE OBJECTION.

(SO1) If nobody is morally responsible, then nobody has normatively binding obligations.

(SO2) If nobody has normatively binding obligations, then nobody has a normatively binding obligation to suspend any of her beliefs.

[3] For example, one of the most prominent moral skeptics writes: "To call a belief 'justified' is to say that the believer ought to hold that belief as opposed to suspending belief, because the believer has adequate epistemic grounds for believing that it is true (at least in some minimal sense)" (Sinnott-Armstrong 2008, 48).

(**SO3**) If nobody has a normatively binding obligation to suspend any of her beliefs, then nobody has a normatively binding obligation to suspend belief in moral responsibility.

(**SO4**) If nobody is morally responsible, then nobody has a normatively binding obligation to suspend belief in moral responsibility.

Either there are morally responsible agents or not. By **SO4**, if there are no such agents, then nobody is obliged to suspend belief in moral responsibility. On the other hand, if there *are* morally responsible agents, some of whom are obliged to suspend some of their beliefs which could not be the case were they not morally responsible, then, one might argue, nobody can have a good reason to suspend belief in moral responsibility. For—assuming that moral responsibility is a precondition of having normatively binding obligations—an agent cannot ascribe to herself an obligation to suspend belief in moral responsibility unless she also takes herself to be morally responsible. That sounds incoherent, so no one can consistently believe, in the light of **SO4**, that there is an obligation to be agnostic about moral responsibility. In short:

> Rationality Premise. If **SO4** is true, then nobody has a normatively binding obligation to suspend belief in moral responsibility.

If the argument so far is sound, it follows that

> SOC. No one has a normatively binding obligation to suspend belief in moral responsibility.

The conclusion of the Simple Argument is a threat to agnosticism because it implies that nobody has a normatively binding obligation to suspend her belief in moral responsibility, even if there is no direct evidence for the reality of moral responsibility. On the other hand, if nobody is in fact morally responsible, then the lack of evidence for the existence of moral responsibility does not matter, since the lack of evidence fails to have normatively binding consequences. The agnostic's claim that we ought to suspend belief in moral responsibility is thus refuted.

3 Why the Simple Objection Fails

One can challenge the Simple Objection on a number of grounds. Here we take into account two objections to **SO1** and one against the Rationality Premise.

doi: 10.48106/dial.v76.i4.01

The first problem about **SO1** is the following. Even if being morally responsible is a precondition of having moral obligations, some normatively binding obligations might not be moral in nature, and so having them does not entail being morally responsible. Indeed, normatively binding obligations come in many varieties. One might have epistemic, aesthetic, prudential, legal, as well as role obligations. Moreover, on many theories of epistemic obligation, epistemic obligations are not moral at all (see, for instance, Feldman 1988; Russell 2001). If the obligation to suspend judgment about moral responsibility is a non-moral, epistemic obligation, then one could have it even if one is not morally responsible. Hence **SO1** seems to be false.

Another important objection to **SO1** is that moral responsibility may not be a precondition of having moral obligations, or so the agnostic could argue. She could rightly claim that if we conceive of moral obligations in a certain way, then it is logically possible for agents who are not morally responsible to be nonetheless morally obliged to do something. For example, one might conceive of moral obligation in a consequentialist fashion and say that we have a moral obligation to maximize pleasure and minimize suffering. And it is possible that whether or not anyone is morally responsible, suspending belief in moral responsibility would minimize the amount of suffering. Thus, it could be the case that someone bears a (consequentialist) moral obligation to suspend belief in moral responsibility regardless of whether she is a morally responsible agent (see Smilansky 1994; Pereboom 2001; Waller 2004; Trakakis 2007).

Further, the Rationality Premise is open to the objection that there is a gap between the truth of a proposition and rationally believing that proposition. Even if **SO1**, **SO2**, and **SO3** are true, it does not necessarily follow that everyone is rational in believing any of those premises. An agent's epistemic position might be such that her evidence either contradicts one of **SO1–SO3** or does not justify any of them. The agent's evidence may even be such that it is rational for her to believe in the soundness of the agnostic's argument. So even if **SO4** is true, there could be a normatively binding obligation to suspend belief in moral responsibility. Therefore, the Rationality Premise appears to be false.

In order to avoid these difficulties, we need to modify the Simple Objection. Instead of talking about normatively binding obligations, we will focus on more specific ones. To evade the second challenge, we will base the argument on a particular conception of moral responsibility, one that, if rationally upheld, renders it irrational to ascribe to oneself an obligation to be agnostic about moral responsibility. Finally, to avoid the difficulties with the Rationality

Premise, we will defend only those agents' beliefs who rationally accepted such a conception. The next section describes the conception that we will work with, the Responsibility First View, in detail.

4 The Responsibility First View

Consider the following famous passage from Wittgenstein:

> Supposing that I could play tennis and one of you saw me playing and said "Well, you play pretty badly" and suppose I answered "I know, I'm playing pretty badly but I don't want to play any better," all the other man could say would be "Ah, then that's all right." But suppose I had told one of you a preposterous lie and he came up to me and said, "You're behaving like a beast" and then I were to say "I know I behave badly, but then I don't want to behave any better," could he then say "Ah, then that's all right"? Certainly not; he would say "Well, you ought to want to behave better." (1965, 5)

When someone says "well, you play pretty badly", in most cases she is not merely offering a description but implies roughly the following: "you should do something about it if you don't want to look ridiculous." In Wittgenstein's story, the player in effect replies that he does not care about this implied "should." Using contemporary terms, one could say that the implied "should" expressed a prudential obligation to prevent an undesirable outcome, such as being ridiculed. Note that even if we suppose that the player would be unhappy if someone actually ridiculed him, he could nonetheless reply "that's all right," because being imprudent is not an unacceptable normative error. It seems to be implausible to think that one should avoid prudential errors with all her strength in every situation or that one should feel remorse if she made such an error. Sometimes, it is all right not to care about prudential obligations even if they actually bind the agent. In other words, it might be OK to neglect them even if violating them constitutes a basis for some negative treatment (such as ridicule).

However, the second example suggests that violating an obligation *is* a normative error that is unacceptable to such a great extent that one should feel remorse and should avoid repeating the error with all her strength. These violations are just not "all right"; they cannot be shrugged off. Wittgenstein and many other philosophers claim that moral obligations fall into this category.

doi: 10.48106/dial.v76.i4.01

Violating them results in unacceptable normative errors. Further, Wittgenstein's paper seems to imply that *only* the violation of moral obligations results in such an error. We will call this idea the Moral Primacy Thesis (MPT).

MPT is central to our case, so we would like to express it more precisely (incidentally explaining why the term "moral primacy" is apt). The following definition of "all things considered obligations" will be useful for that purpose:

> All things considered obligation to $\Phi =_{df}$. An obligation which is not overridden by any other obligation (in the given case) and which prescribes doing Φ to agent *S* in a way that *S* should avoid violating the obligation with all her strength; and if *S* fails to observe the obligation to Φ, then *S* should feel remorse.

We follow here Searle (1978) and many other philosophers who used the term "all things considered obligation." Nevertheless, we add that a genuine all things considered obligation to perform a specific action must have a normative weight that makes its violation normatively unacceptable. If an obligation does not have the significant normative weight, then—all things considered—it is permissible to ignore it.

Philosophers often talk about obligations that have a tendency to constitute all things considered obligations. They call these *pro tanto* or *prima facie* obligations (Ross 1930). These tend to constitute all things considered obligations if other, stronger obligations do not override them. (The paradigmatic examples are moral obligations.) However, for our purposes, it is better to not commit ourselves to any specific understanding of *pro tanto* or *prima facie* obligations because not only the difference between *pro tanto* and all things considered obligations is relevant for our argument but the difference between obligations that can constitute all things considered obligations in themselves and obligations that can do this only by the help of other obligations. So, instead of talking about *prima facie* and *pro tanto* obligations, we will use the term "strong obligation", defined as follows:

> Strong obligation to $\Phi =_{df}$. An obligation that constitutes an all things considered obligation to Φ (in the given case) unless it is overridden by some other strong obligation(s) to do something else.

So in some cases, Strong obligations to Φ constitute an all things considered obligation to Φ, and in other cases, strong obligations to Φ do not constitute an all things considered obligation to Φ (if they are overridden by other strong obligations).[4]

In addition, there are obligations that fail to constitute all things considered obligations in spite of the fact that nothing overrides them. For instance, in many cases, prudential obligations do not constitute all things considered obligations even though the agent has no other kind of obligation. This is precisely the case in Wittgenstein's example: although the agent has a prudential obligation to play tennis better, he is free to ignore and violate it. We call these obligations *weak* obligations.

> WEAK OBLIGATION TO Φ $=_{df}$. An obligation that does not constitute an all things considered obligation to Φ unless it appropriately relates to a strong obligation in the given case.

We intentionally use the vague term "appropriately relates." It is a complicated question when and how strong obligations turn weak obligations into all things considered obligations. For the present purposes, what matters is that this is certainly possible—whatever the details. For instance, if the tennis player in Wittgenstein's example had previously promised his wife to do his best and avoid ridicule, and there was no strong obligation to override the obligation to keep his promise, then he would have an all things considered obligation to play better. In this case, his prudential obligation to play better would be an all things considered obligation, because it would be appropriately related to his moral obligation to fulfill his promise to avoid ridicule.

Using the terminology just introduced, we can now characterize MPT more precisely:

> MORAL PRIMACY THESIS (MPT). All moral obligations are strong obligations and every other kind of obligation is weak.

In other words, only moral obligations can constitute all things considered obligations without being appropriately related to other kinds of obligations. On the other hand, prudential, epistemic, role, legal, etc. obligations can only

[4] It is worthwhile to note that our notion of strong obligation resembles the Kantian notion of categorical imperative. The main difference is that our notion of strong obligation does not imply universalizability. That is, we do not deny the possibility that one may have an obligation that constitutes an all things considered obligation to Φ, although one cannot at the same time will that it becomes a universal law.

constitute all things considered obligations if they are appropriately related to moral obligations.

It is easy to see that one of the relevant consequences of MPT is the following:

> Normative Weakness of Non-Moral Obligations. Nobody has to avoid violating with all her strength those non-moral obligations that do not relate appropriately to any of her moral obligations and if someone fails to observe such an obligation, she should not feel remorse.

Attributing moral obligation to agents must have some conditions. For example, it would certainly be absurd to attribute moral obligations to beings that are incapable to act, because it makes no sense to say that they should avoid doing something with all their strength. Whatever the relevant conditions are, there is an obvious term, under MPT, for those beings who fulfill all of them: they are the morally responsible agents. So proponents of MPT are free to adopt the following thesis as a component of their moral framework:

> Responsibility Identity Thesis (RIT). The property of being morally responsible is identical to the property of fulfilling all conditions for bearing moral obligations.

RIT makes moral responsibility into a precondition of having moral obligations, and it also makes the former prior to the latter in a certain respect. Since moral obligations are, in turn, prior to any other type of obligations under MPT, we will call the combination of MPT and RIT the Responsibility First View.

5 The Primacy Argument

We can now turn to the revision of the Simple Objection. The proponent of RFV can answer the agnostic's challenge as follows: "You claim that I should suspend my belief in moral responsibility because I cannot prove that anyone meets the conditions for being morally responsible. However, based on my conception of morality and responsibility, if I am not morally responsible, then I do not have any moral obligations. And if I have no moral obligations, I do not have any obligations that I should fulfill with all my strength, any obligations that should seriously concern me. In technical terms, I do not have all things considered obligations. So if I am not morally responsible, then it is all right for me to disregard your demand about suspension of belief. And in case you claimed that I have an all things considered obligation to suspend my belief, an obligation which I cannot disregard without committing a normative fault I should regret, then I conclude on the basis of my conception of responsibility that I am a morally responsible being after all. Either way, I can rationally resist your challenge and keep believing in moral responsibility."

We would like to express this revised version of the Simple Objection more formally:

The Primacy Argument.

(**PA1**) No agent can rationally think that she has an all things considered obligation to suspend her belief that she fulfills the necessary conditions of having all things considered obligations.

(**PA2**) Someone who rationally upholds RFV cannot rationally think that she has an all things considered obligation to suspend her belief that she is morally responsible.

(**PA3**) If someone cannot rationally think that she has an all things considered obligation to suspend her belief that she is morally responsible, then she is rational to reject agnosticism about moral responsibility.

(**PAC**) Rejecting agnosticism about moral responsibility is rational for anyone who rationally upholds RFV.

Until the agnostic does not challenge the moral framework that proponents of RFV employ, she cannot undermine their belief in their own moral responsibility. The agnostic cannot challenge belief in moral responsibility by merely pointing out that evidence for the existence of moral responsibility is insufficient. What is more, if someone upholds RFV rationally, then it would be straightforward irrational for her to accept the agnostic's conclusion—unless she finds out that her own moral framework is untenable.

The proponent of RFV gains a huge dialectical advantage by deploying the Primacy Argument. Due to the Primacy Argument, the debate shifts from the sufficiency of evidence to the tenability of a specific moral framework. Defending the tenability of RFV seems to be much easier than defending the sufficiency of evidence regarding the existence of moral responsibility. Especially so if the proponent of the Moral Primacy Thesis, by investigating the nature of moral obligation and responsibility, comes to the conclusion that moral responsibility has heavy-weight metaphysical preconditions such as libertarian free will, since scientific evidence for libertarian freedom seems to be lacking. (We will say more about free will in the last section.)

Moreover, as far as we can tell, both MPT and RIT can be supported by considerable arguments. Even though some consequentialists deny that moral obligation implies moral responsibility, that principle seems to be fundamental and obvious for almost everyone—as even consequentialist critics note (Waller 2004, 427–428). And someone who upholds RIT can explain why that principle is true: being a morally responsible agent is the same as being a potential bearer of moral obligations.

MPT also has notable advantages. Many people find it plausible that moral obligations can override all other obligations. MPT explains why this is the case: the set of moral obligations is identical to the set of strong obligations. Additionally, MPT provides a substantive definition of moral obligation: moral obligations are those obligations that can constitute all things considered obligation without the involvement of other types of obligation. Another notable advantage of MPT is that it helps understanding why some obligations can be neglected without normative costs in certain cases but not in others. Take, for example, the highway code, which prescribes various patterns of behaviour (call them "legal obligations"). Some of those prescriptions can be non-culpably neglected in a completely abandoned city. Still, in most cases, violating them is normatively unacceptable. One can use MPT to explain this phenomenon by pointing out that the highway code contains weak obligations. In most cases, they are appropriately related to moral obligations (for example, to the obligation to secure the safety of human beings). However, they are not appropriately related to moral obligations in an abandoned city.

Of course, anyone, including the agnostic, can argue against RFV. Indeed, it seems that one can find not only prominent supporters of RFV (Kant seems to be the most obvious example) but able critics too. For instance, Bernard Williams criticizes an ethical system under the label "morality" that contains, among other things, RFV (see Williams 2006, 174–196), because he believes that moral systems with such a strong notion of moral obligations threaten personal integrity. Even though the investigation of such counterarguments that are based on such wide-ranging considerations about the relation between a whole system of morality and other basic values is out of the scope of our paper, we can deal with another argument against RFV that is rather closely related to the problem of moral responsibility.

Namely, RFV seems to imply that epistemic obligations are identical with or, at least, not independent of moral obligations which means, in turn, that anyone who accepts RFV and would like to believe in epistemic norms is forced to believe in moral responsibility no matter which crazy theory about conditions of moral responsibility turns out to be true. For example, if Derk Pereboom's analysis on the conditions of moral responsibility is correct, moral responsibility needs not only agent-causation (which, according to Pereboom, may be a logically incoherent concept), but either systematic breaking of the laws of nature or inexplicable harmony between micro-physical statistical laws and the free decisions of the agents (see Pereboom 2001). For sure, believing that these conditions are met in reality would be a high price to pay for holding RFV. Insofar as the price is so high, it seems to be not only irrelevant, but weird that the proponent of RFV can and even should rationally defend believing in moral responsibility and its monstrous metaphysics by moving the battlefield from metaphysics to metaethics. After all, forming rational beliefs and fulfilling epistemic norms aim at the truth, and it is not too probable that this way of belief-formation leads us to true beliefs.[5]

To be clear, RFV does not imply that epistemic obligations as such depend on (or are identical with) moral obligations. RFV does not exclude that they are totally unrelated to each other. What RFV implies is only that an epistemic obligation has to be appropriately related to some moral obligations in order to be true that agents have to avoid violating it with all their strength and if someone fails to observe an epistemic obligation which does not relate appropriately to any moral obligation, she should not feel remorse. In other words, in themselves, epistemic obligations do not have sufficient normative weight to constitute all things considered obligations. So, if one both accepts RFV and rejects moral responsibility based on her evidence-basis, she cannot rationally believe that there are moral and all things considered obligations, but she still can rationally think that there are (weak) epistemic obligations. What she cannot rationally believe is that neglecting any epistemic obligation cannot be OK in the same way as neglecting the prudential obligation not to ridicule oneself. That is, even if one accepts RFV and, for instance, Pereboom's assessment of the evidence about free will and moral responsibility, she can rationally deny the existence of (a metaphysically rather extreme kind of) free will, moral responsibility, moral obligations, and all things considered obligations. The only thing that she cannot rationally maintain without re-

[5] We would like to thank an anonymous reviewer for pointing out this possible objection.

jecting RFV is the idea that anyone should suspend belief in those things to avoid committing an unacceptable normative error that cannot be shrugged off. In other words, if there is a proponent of RFV who tries to heroically find the truth no matter the cost and finds that her evidence-basis strongly indicates the non-existence of moral responsibility, she can rationally believe that she has an epistemic (or even prudential) obligation to deny the existence of moral responsibility, but she cannot rationally think either that she has an all things considered obligation to reject moral responsibility or that anyone has an all things considered obligation to try to find the truth no matter the cost.

Nonetheless, none of this undermines our point that the proponent of RFV, if she wants to defend the belief in moral responsibility, can move the battlefield from metaphysics to metaethics, and the latter seems to be much more advantageous for her, especially if she also holds that the sufficient conditions of moral responsibility are metaphysically rather demanding. The more demanding these conditions are, the less plausible is the claim that the existence of moral responsibility is obvious and/or probable in the light of the given evidence, so moving the battlefield from metaphysics to metaethics provides more strategic advantage.

Note, parenthetically, that one can construct a modified version of the Primacy Argument even if both MPT and RIT are untenable. One need not appeal to morality (or moral responsibility) at all. Anyone who rationally believes in all things considered obligations has the epistemic right to sustain belief in a specific kind of responsibility. As the first premise of the Primacy Argument says, no agent can rationally think she has an all things considered obligation to suspend the belief that she fulfills the necessary conditions of having all things considered obligations. In other words, someone who rationally attributes all things considered obligations to herself must also accept that she fulfills the necessary conditions of having all things considered obligations. Since having strong obligations is one of those necessary conditions, the agent in question must also accept that she fulfills the necessary conditions of having strong obligations. It is reasonable to say that being responsible "in a strong sense" requires fulfilling all necessary conditions for bearing strong obligations, so anyone who rationally attributes all things considered obligations to herself can rationally attribute "strong responsibility" to herself as well. It seems that this argument for "strong responsibility" can be threatened only by arguments against the existence of all things considered obligations.

To sum up the Primacy Argument, anyone who rationally accepts RFV can rationally maintain her belief in moral responsibility even if she does not have sufficient direct evidence that anyone fulfills the metaphysical conditions of being morally responsible. Until the agnostic refutes MPT or RIT, one can rationally resist the agnostic challenge.

6 Free Will, Fairness, and Others

Various questions could be raised about our argument. We will look at three. First, one might ask how the dialectic is related to free will. We claimed that someone who rationally believes in RFV does not have to suspend her belief in moral responsibility even if she lacks direct evidence for it. Could RFV be used to defend belief in free will as well?

The answer to this question depends on one's conception of free will. There are two basic approaches in the literature. According to the first, having free will means fulfilling a subset of conditions that guarantee necessary control over one's morally relevant actions (Clarke 1992). The present argument obviously extends to the defense of free will conceived this way. If someone rationally accepts RFV and also rationally thinks that she fulfills all necessary conditions for being morally responsible, then she cannot rationally believe that she fails to fulfill a subset of those conditions, namely those that are necessary for control. So our argument supports belief in free will for those who rationally believe RFV and identify having free will with fulfilling a subset of necessary control conditions for being morally responsible.

However, there is another prevalent conception, according to which free will is the ability to do otherwise (van Inwagen 1983). Our argument can be extended to this case as well, but only if one rationally upholds that the ability to do otherwise is a necessary precondition of being morally responsible. Given strong evidence that moral responsibility depends on free will of the second sort, then rational belief in RFV (together with the evidence in question) can ground rational belief in the existence of free will. And if the proponent of RFV has sufficient evidence that moral responsibility has further metaphysical conditions, she can also rationally believe that she fulfills all those further conditions, regardless of how demanding they are metaphysically.

These possible extensions of the Primacy Argument are especially significant if one takes into account that many philosophers and scientists insist that there is no sufficient scientific or other evidence for macro-level psychological indeterminism (which is a precondition of libertarian free will)

doi: 10.48106/dial.v76.i4.01

or the presence of compatibilist-friendly causal determinism in the brain. In light of the possible extensions of the Primacy Argument, proponents of RFV can rationally believe in responsibility-relevant free will (of either the incompatibilist or compatibilist sort) even in the absence of sufficient direct scientific or other evidence.

This last point regarding the absence of evidence leads us on to a potential objection implied by Scott Sehon. He emphasizes that we treat responsible and irresponsible agents very differently. If, for example, someone pushes another person into the traffic, we treat her act very differently depending on whether she was or was not responsible. If she was, then her act “certainly looks incredibly reprehensible and maybe even the stuff of an attempted murder charge” (Sehon 2013, 369). But if we know that the pusher is not responsible, we would not call her action “reprehensible” and would not make her face serious charges. Sehon adds that “[it] would be manifestly unfair to regard the agent as responsible if our degree of certainty on the matter is quite low” (2013, 36). One could extend this point and argue that if we lack strong direct evidence for moral responsibility, then, out of fairness, we should suspend belief about whether anyone is ever responsible in a way that would render retribution justified.

Proponents of the Primacy Argument evidently disagree, as their supposedly rational belief in RFV makes them rational in holding that they can be morally responsible for their actions. It is important that Sehon brings up this issue in terms of fairness. The obligation to be fair with others is naturally understood to be a moral obligation, and hence a strong one—according to MPT. Those who uphold RFV will see the situation as follows. The obligation to be fair can only be attributed to morally responsible persons. If nobody is morally responsible, then the strong obligation to be fair cannot be attributed to anybody. And if that is the case, then nobody has to care about being fair to anybody. So if the proponent of RFV takes Sehon’s exhortation to be fair seriously, and if she thinks she has to care about it, then, in the light of MPT, she incidentally attributes a strong obligation to herself. As a result, she implies that she fulfills all the conditions of having strong obligations, including having moral responsibility. That is, for proponents of RFV, Sehon’s point can only have force if they take themselves to be morally responsible. They would need to assume, first, that they are morally responsible, and, second, they would have to suspend judgment about the existence of moral responsibility because of that very assumption—which seems incoherent. Thus, the argument that insufficient direct evidence for moral responsibility should make us suspend our belief in moral responsibility because it might lead to the unfair treatment of others makes no sense to those who hold RFV to be true.

The question of being fair to others and taking them to be morally responsible brings us to the crucial issue of the circle of agents whom one might attribute moral responsibility to, on the basis of the Primacy Argument. This is a crucial issue, as it might very well be the case that the individuals who accept RFV can attribute moral responsibility only to themselves and not to anyone else. This is because **PA1** takes only the agent's own perspective into account. The agent is considering her own obligations and moral responsibility, and the reason why she doesn't have to become an agnostic is that, were she to take agnosticism as a strong obligation, she would thereby attribute moral responsibility to *herself*. (As we have indicated, she might even go on to attribute free will as well.) But the incoherence would arise only in her own case, so the Primacy Argument's conclusion applies only to her: she is free to go on believing that she, for one, is morally responsible. And clearly, she can believe in the existence of moral responsibility on the basis on *that*, since moral responsibility exists even if only one agent has it.

Extending this rather small circle of responsible agents might look unreasonable or unfair indeed. However, there could be ways to do it. Remember that the agent in question reasonably believes in her own responsibility. If she considers agents who seem to be like her in every relevant respect, she may take them to be morally responsible as well. Nevertheless, the reasonableness of this move depends on two crucial factors. First, the agent must have a rationally held theory of what the relevant respects are. Second, were she to deem morally responsible any agent other than herself, her judgment that that person is similar to herself in every relevant respect must also be rational. This means that reasonably extending the circle of morally responsible agents to others is logically possible, but could be difficult in practice. Fortunately, there might be an easier way. It seems reasonable to think that all fully developed human beings have the same metaphysical structure. Insofar as this assumption is reasonable, a proponent of the Primacy Argument can extend the circle of morally responsible agents to all fully developed human adults who fulfill the non-metaphysical and empirically verifiable conditions of moral responsibility, whether or not she can identify the precise metaphysical conditions for being morally responsible.

Note that extending the circle of responsibility poses a challenge not only with regard to other agents, but also with regard to the agent who can rationally believe in moral responsibility based on RFV and PA. This is because The

doi: 10.48106/dial.v76.i4.01

Primacy Argument does not imply that the agent is morally responsible all the time. It only permits the agent to believe that she is morally responsible in her present state. Nevertheless, what we have said previously about the possibilities of extending the circle of morally responsible agents can also be used to extend this temporal limitation. This means that if an earlier or later state of the agent seems to be similar in every relevant respect to her present state, then she may take it that she was or is going to be morally responsible at those times. However, it might not be clear in every case that these conditions are fulfilled. Therefore, our argument is compatible with accepting that even though we are reasonable in thinking that we are morally responsible some of the time, we could be also reasonable in thinking that we are not responsible at other times, or thinking that we should be agnostics about the question whether we are morally responsible in certain situations.*

László Bernáth
Research Centre for the Humanities
Budapest
bernath.laszlo@abtk.hu

Tamás Paár
Central European University
Budapest
paar.tamas@gmail.com

* We would like to express our indebtedness to the anonymous referees of this paper, their insightful comments helped us developing our position a great deal. We also thank Dániel Kodaj for his suggestions and the stimulating exchanges we had on many occasions. László Bernáth was supported by the János Bolyai Research Scholarship of the Hungarian Academy of Sciences (grant no. BO/00432/18/2), the OTKA (Hungarian Scientific Research Fund by the National Research Development and Innovation Office) Postdoctoral Excellence Programme (grant no. PD131998), and another OTKA research grant (grant no. K132911). Tamás Paár was supported by the ÚNKP-19-3-III-PPKE-24 New National Excellence Program of the Ministry for Innovation and Technology. His research was supported by the University of Oxford project "New Horizons for Science and Religion in Central and Eastern Europe" funded by the John Templeton Foundation. (The opinion expressed in the publication are those of the author(s) and do not necessarily reflect the view of the John Templeton Foundation.) We owe special thanks to the organizers and the committee of the Oxford University Essay Competition for Central and Eastern European Perspectives on Science, Theology and Philosophy in which this paper has been awarded first prize.

References

BYRD, Jeremy. 2010. “Agnosticism about Moral Responsibility.” *Canadian Journal of Philosophy* 40(3): 411–432, doi:10.1080/00455091.2010.10716729.

—. 2021. “What Should we Believe About Free Will?” *Erkenntnis* 86(3): 505–522, doi:10.1007/s10670-019-00116-3.

CLARKE, Randolph. 1992. “Free Will and the Conditions of Moral Responsibility.” *Philosophical Studies* 66(1): 53–72, doi:10.1007/bf00668395.

COFFMAN, E. J. 2016. “Incompatibilist Commitment and Moral Self-Knowledge: The Epistemology of Libertarianism.” in *Philosophical Issues 26: Knowledge and Mind*, edited by Christoph KELP and Jack LYONS, pp. 78–98. Oxford: Wiley-Blackwell, doi:10.1111/phis.12066.

FELDMAN, Richard H. 1988. “Epistemic Obligations.” in *Philosophical Perspectives 2: Epistemology*, edited by James E. TOMBERLIN, pp. 235–256. Oxford: Basil Blackwell Publishers, doi:10.2307/2214076.

GUILLON, Jean-Baptiste. 2014. “Van Inwagen on Introspected Freedom.” *Philosophical Studies* 168(3): 645–663, doi:10.1007/s11098-013-0159-x.

KEARNS, Stephen. 2015. “Free Will Agnosticism.” *Noûs* 49(2): 235–252, doi:10.1111/nous.12032.

LATHAM, Andrew J. 2019. “The Conceptual Impossibility of Free Will Error Theory.” *European Journal of Analytic Philosophy* 15(2): 99–120, doi:10.31820/ejap.15.2.5.

LEVY, Neil. 2011. *Hard Luck. How Luck Undermines Free Will & Moral Responsibility*. Oxford: Oxford University Press, doi:10.1093/acprof:oso/9780199601387.001.0001.

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

MELE, Alfred R. 2019. “Causalism: On Action Explanation and Causal Deviance.” in *Explanation in Action Theory and Historiography. Causal and Teleological Approaches*, edited by Gunnar SCHUMANN, pp. 45–58. London: Routledge, doi:10.4324/9780429506048-2.

PEREBOOM, Derk. 2001. *Living Without Free Will*. Cambridge: Cambridge University Press, doi:10.1017/CBO9780511498824.

ROSS, William David. 1930. *The Right and the Good*. Oxford: Oxford University Press.

RUSSELL, Bruce. 2001. “Epistemic and Moral Duty.” in *Knowledge, Truth, and Duty: Essays on Epistemic Justification, Responsibility, and Virtue*, edited by Matthias STEUP, pp. 34–47. Oxford: Oxford University Press, doi:10.1093/0195128923.003.0003.

SEARLE, John R. 1978. “Prima Facie Obligations.” in *Practical Reasoning*, edited by Joseph RAZ, pp. 81–90. Oxford: Oxford University Press. Reprinted in van Straaten (1980, 238–259), doi:10.1093/acprof:oso/9780199693818.003.0007.

SEHON, Scott R. 2013. “Epistemic Issues in the Free Will Debates: Can we Know when We Are Free?” *Philosophical Studies* 166(2): 363–380, doi:10.1007/s11098-012-0044-z.

—. 2016. *Free Will and Action Explanation. A Non-Causal, Compatilist Account*. Oxford: Oxford University Press, doi:10.1093/acprof:oso/9780198758495.001.0001.

SINNOTT-ARMSTRONG, Walter. 2008. “Framing Moral Intuitions.” in *Moral Psychology, Volume 2. The Cognitive Science of Morality: Intuition and Diversity*, edited by Walter SINNOTT-ARMSTRONG, pp. 47–76. Cambridge, Massachusetts: The MIT Press.

SMILANSKY, Saul. 1994. “Ethical Advantages of Hard Determinism .” *Philosophy and Phenomenological Research* 54(1): 355–363, doi:10.2307/2108494.

STRAWSON, Galen. 1994. “The Impossibility of Ultimate Moral Responsibility.” *Philosophical Studies* 75(1–2): 5–24. Reprinted in Strawson, G. (2008, 319–336), doi:10.1007/bf00989879.

—. 2008. *Real Materialism, and Other Essays*. Oxford: Oxford University Press, doi:10.1093/acprof:oso/9780199267422.001.0001.

STRAWSON, Peter Frederick. 1969. *Meaning and Truth*. Oxford: Oxford University Press. Reprinted in Strawson, P. F. (1971, 170–189) and in Strawson, P. F. (2004, 131–146).

—. 1971. *Logico-Linguistic Papers*. London: Methuen & Co. Reprinted as Strawson, P. F. (2004).

—, ed. 2004. *Logico-Linguistic Papers*. 2nd ed. Aldershot, Hampshire: Ashgate Publishing Limited, doi:10.4324/9781315250250.

TRAKAKIS, Nick N. 2007. “Whither Morality in a Hard Determinist World? .” *Sorites* 19(1): 14–40, http://lorenzopena.es/Sorites/Issue_19/sorites19.pdf.

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

VAN STRAATEN, Zak, ed. 1980. *Philosophical Subjects: Essays Presented to P.F. Strawson*. Oxford: Oxford University Press.

WALLER, Bruce N. 2004. “Virtue Unrewarded: Morality without Moral Responsibility.” *Philosophia: Philosophical Quarterly of Israel* 31(3–4): 427–447, doi:10.1007/bf02385194.

WILLIAMS, Bernard Arthur Owen. 1985. *Ethics and the Limits of Philosophy*. Hammersmith: Fontana Press.

—. 2006. *Ethics and the Limits of Philosophy*. London: Routledge. With a commentary on the text by A.W. Moore; original publication: Williams (1985), doi:10.4324/9780203969847.

WITTGENSTEIN, Ludwig. 1965. “A Lecture on Ethics.” *The Philosophical Review* 74(1): 3–12, doi:10.2307/2183526.

Self-Knowledge and Interpersonal Reasoning

Benjamin Winokur

Many philosophers contend that we often possess "privileged" and "peculiar" self-knowledge of our mental states. Self-knowledge is privileged insofar as it is systematically more secure than the knowledge that others have of one's propositional attitudes, and it is peculiar insofar as it is systematically obtained in a way that is only suited for delivering self-knowledge. Focusing on privileged and peculiar self-knowledge of propositional attitudes like beliefs, I offer an account of its instrumental value. On my account, privileged and peculiar self-knowledge of one's propositional attitudes enables one to be a more efficient and reliable interpersonal reasoner.

Self-knowledge of one's current mental states often seems interesting—if not outright puzzling—for at least two reasons. First, such self-knowledge often seems to be *privileged*, for it seems to be systematically (though not universally) more secure than the knowledge one has of others' mental states. Second, it often seems to be *peculiar*, for it seems to be systematically (though, again, not universally) obtained in a way that is only suited for delivering self-knowledge, hence, *not* by whatever means enable one to acquire knowledge of *other* minds (Byrne (2018), 4–9). The standard project in contemporary theorizing about self-knowledge is to vindicate these appearances by unearthing the special security and sources of self-knowledge. However, others have argued that we do not actually possess any privileged and peculiar self-knowledge (hereafter "PPSK"), at least when it comes to self-knowledge of propositional attitudes like belief (Gopnik 1993; Carruthers 2011; Cassam 2014). These PPSK-skeptics typically understand self-knowledge and other-knowledge of propositional attitudes as on a par in terms of their security, source, or both.

In reply, some PPSK-realists have offered competing interpretations of the putative evidence against realism about PPSK of propositional attitudes Keeling (2019b), while others have pushed back against the non-privileged and non-peculiar accounts of self-knowledge that are favoured by many skeptics (Coliva 2016; Keeling 2018; Marcus and Schwenkler 2019; Andreotta 2022). The stakes of these debates are hard to grasp if we are unsure "what, if anything, of value we fail to possess if these skeptics are right" (Peterson 2021, 365). For this reason, I will argue that PPSK of one's propositional attitudes—chiefly, our beliefs—is instrumentally valuable for the efficiency and reliability of a widespread activity in our social-epistemic lives, that of *interpersonal reasoning*. Some readers may interpret my arguments as providing further support for PPSK-realism if they believe that interpersonal reasoning is in fact a highly efficient and reliable activity in our actual lives. Other readers might reach the more modest conclusion that interpersonal reasoning is a more effective enterprise *to the extent that we possess PPSK*, whether or not we really possess PPSK, and hence whether or not interpersonal reasoning is a particularly efficient and reliable activity for us to undertake. Either way, the significance of debates between PPSK-skeptics and PPSK-realists can be better appreciated in light of what follows.

Here is the layout for my paper. In section 1 I draw initial inspiration from two earlier accounts of PPSK's instrumental value. The first, due to Sydney Shoemaker (1988, 1996), concludes that social cooperation in general requires each of us to possess PPSK of many of our propositional attitudes. The second, due to Charles Siewert (2003), concludes that PPSK is indispensable to social cooperation *whenever this depends on justifying one's actions to others*. Justifying one's actions to others can be one way of reasoning with others, that is, reasoning interpersonally. But it is *only* one way of reasoning interpersonally. I thus consider, in section 2, whether *all* interpersonal reasoning might benefit from PPSK. My argument is that PPSK does indeed play beneficial roles in all interpersonal reasoning. In section 3 I address objections to my account. In section 4 I consider another recent account of PPSK's instrumental value, one that emphasizes its role in our capacity for "epistemic control," and I show how my account complements that account. In section 5 I conclude.

1 Cooperation and Privileged, Peculiar Self-Knowledge

Is PPSK instrumentally valuable? Some philosophers have argued that it is.[1] Indeed, some have argued that it is instrumentally *indispensable*. Here is Shoemaker, who writes of self-knowledge by "self-acquaintance" instead of privileged and peculiar self-knowledge:[2]

> When one is engaged in a cooperative endeavor with another, it is essential to the efficient pursuit of the shared goal that one be able to communicate to the other information about one's beliefs, desires and intentions [...] When in such circumstances one conveys one's beliefs to another, this is not merely for the purpose of conveying what one takes to be information about the world, namely the contents of the beliefs; it is also for the purpose of giving him information about oneself which will assist him in predicting one's behavior and so in coordinating his own behavior with it, and also to enable him to correct those of one's beliefs he knows to be mistaken [...] And here the utility of self-knowledge depends crucially on its being acquired by self-acquaintance; if I had to figure out from my behavior what my beliefs, goals, intentions, etc. are, then in most cases it would be more efficient for others to figure this out for themselves than to wait for me to figure it out and then tell them about it. (1988, 185–186)

Shoemaker argues that PPSK is indispensable for efficiently cooperating with other human beings. For, if others could know one's mind in the same way and as reliably as one knows one's own mind, one would be far less efficient at soliciting others' cooperation. This is because it would just as often be up to others to figure out one's mind, and to decide on this basis whether cooperation was worthwhile. As a result, one would frequently fail to solicit others' cooperation of one's own accord.

[1] Peterson (2021, 1) thinks that the question of PPSK's value has been ignored by epistemologists working on self-knowledge. While I myself hope to contribute an answer to this question, I think that this assessment of the extant literature is somewhat exaggerated given the views that I discuss in this section, among others [see, e.g., Burge (1996); Nguyen (2015); Sorgiovanni (2019); Winokur (2021a), Winokur (2021b). Peterson's own account of PPSK's instrumental value is discussed in section 4. He also discusses the potential *intrinsic* value of PPSK, a topic that I do not broach here.

[2] These, I submit, are just notational variants.

doi: 10.48106/dial.v76.i4.02

Reflecting on Shoemaker's argument, Siewert wonders whether rational animals "could engage in cooperation and assistance-seeking behaviour, even by generally acting in an attitude-revealing fashion, without representing their own minds to themselves" (2003, 139). In a different idiom: couldn't there be creatures that are exceptionally adept at *expressing*—i.e., showing, manifesting, displaying—their attitudes to their fellow creatures without also possessing *PPSK of the attitudes expressed*, and couldn't this enable equally efficient cooperation?[3] Contra Shoemaker, Siewert supposes that there could be such creatures. Still, he is optimistic about a nearby argument:

> For whether or not there can be social animals that act in a usefully self-revealing fashion while oblivious to their own psychologies, they could not engage in the practice of *justifying* such acts, without being able to represent, in their justifications, relevant facts about their own desires and beliefs [...] Now, if the reasons we would offer did not have us acting in ways revealing our actual beliefs and desires to others, we would be much less effective in securing others' cooperation and assistance in the satisfaction of our desires than we in fact are. (2003, 139)

On this argument, it is not that efficient cooperation always requires PPSK. Rather, such self-knowledge is required for cooperation *whenever such cooperation also depends on justifying one's actions to one's would-be cooperators*. For, lacking PPSK, our actions would often fail to cohere with the attitudes that we self-ascribe. In turn, we would be worse at justifying our actions because we would be worse at appealing to the actual beliefs, desires, and intentions that underwrite them. These inconsistencies might be noticed by others, and this might diminish their trust in us.

More recently, Jon Greco has written that:

> *Of course* thinking about one's first-order mental states is essential to activities involving coordination and cooperation [...] In particular, *giving one's reasons*, both epistemic and practical, is essential to various activities in which one must defend one's beliefs and actions, and having a grasp on such mental states oneself is essential to reporting them to others. [-Greco (2019), 52][4]

[3] This is my gloss on Siewert's argument. Like Bar-On (2004), I use "express" here to denote actions that express mental states, though I denote another sense of expression in section 3.

[4] See Müller (2019, 6) for a similar view.

Like Siewert, Greco claims that self-knowledge is essential for justifying one's actions (and beliefs) to others. However, Siewert argues that PPSK is indispensable to our widespread success in these matters, whereas Greco claims that "this kind of metacognitive activity can tolerate the same fallibility that we experience in cognition generally" (2019, 53).[5] He thus denies the importance of *privileged* self-knowledge (and is silent about *peculiar* self-knowledge). But he does not consider Siewert's argument, and so it is hard to know whether his position would change upon further reflection.

This difference between Greco and Siewert set aside, notice that they both focus on a certain kind of interpersonal reasoning. Here, 'interpersonal reasoning' denotes exchanges of assertions between interlocutors, or exchanges of questions and assertions, toward a discursive end. For instance, one might reason interpersonally in order to acquire new rational attitudes, or to subject one's already-held attitudes to the scrutiny of other agents, or to persuade other agents to adopt one's already-held attitudes. I say that Siewert and Greco are focused on a certain kind of interpersonal reasoning because they only focus on cases in which agents reason interpersonally about one another's actions or attitudes. In other words, neither philosopher focuses on cases in which agents aim to justify "agent-neutral" propositions to one another, these being propositions whose contents do not refer to any particular agent's actions or attitudes. One such proposition is:

> Runaway climate change is a worsening phenomenon.

It is to be contrasted with the sorts of propositions that Siewert and Greco focus on, namely "agent-specific" propositions like:

> I should continue to be vigilant about my fossil fuel consumption.

[5] Greco also ventures a response to the possibility of efficiently cooperative animals lacking self-knowledge: "One might object that non-human animals are also social in a sense that implies coordination and cooperation, and they manage their social lives without citing their mental states in explanations to themselves or their cohorts. But this objection misses the point that human social agency is also rational agency. It involves rationalizing one's thoughts and actions by means of giving one's reasons—i.e., overtly giving one's reasons—to oneself and to others" (2019, 53). This too is reminiscent of Siewert's view. But while Greco denies that cooperation among non-human animals involves rational agency, Siewert thinks that non-human animals could count as rational agents in so cooperating.

The latter proposition, but not the former, requires agents to provide self-referential information, this being information that justifies *the agent herself* to act in such-and-such a way or have such-and-such an attitude. Such information will naturally include “relevant facts about their own desires and beliefs” (Siewert 2003, 139) whereas reasoning about agent-neutral propositions simply requires providing first-order evidence about the agent-independent world, e.g., evidence of rapidly melting arctic ice. But because reasoning about either sort of proposition can be conducted interpersonally, Siewert and Greco will have shown—at most—that self-knowledge matters for interpersonal reasoning about agent-specific matters (whether such self-knowledge is privileged and peculiar as Siewert claims, or not, as Greco claims). I note this here because I will argue in section 2 that PPSK plays a role in both agent-specific and agent-neutral interpersonal reasoning.

Before I get there, I want to make two preliminary points. First, the reader may have wondered whether that Shoemaker’s and Siewert’s arguments establish what they purport to since, on close inspection, they seem to emphasize the importance of privileged access but not, in addition, peculiar access. This is because each argument insists that the special security of agents’ self-knowledge is what facilitates cooperation with other people, and yet this does not obviously entail that agents must exploit a peculiar means of achieving such security. In what follows I will provide arguments for the importance of peculiar self-knowledge as well, thereby going beyond the arguments considered thus far.

Second, it should be noted that some philosophers deny that interpersonal reasoning of *any* kind (i.e., whether about agent-neutral or agent-specific propositions) requires self-knowledge of *any* kind (i.e., whether privileged and peculiar, non-privileged and non-peculiar, or any other combination). For example, Robert Brandom writes that there is “nothing incoherent in descriptions of communities of judging and perceiving agents, attributing and undertaking propositionally contentful commitments, giving and asking for reasons, who do not yet have available the expressive resources *I* provides” (1994, 559). If these communities lack articulate use of the first-person singular, then they cannot self-ascribe and hence self-know their attitudes.[6]

[6] See also Strijbos & De Bruin (2012). The importance of this claim depends on assuming that self-knowledge requires linguistically articulate self-ascriptive thought, and some friends of “tacit” self-knowledge might dispute this (e.g., Boyle 2011, 2019). Alternatively, it could be granted that there is such a thing as self-*consciousness* that does not involve linguistically articulate self-ascriptive thought (cf. Musholt 2015, chap. 4). Even if this is a tenable view, I am focusing

Similarly, Ladislav Koreň claims that we can reason interpersonally by exercising a "practical competence" with linguistic devices like "no," "but," and "so," thus manifesting a "sensitivity" to rational connections between claims without having "metarepresentational" beliefs about the rational connections between one's own attitudes or one's interlocutor's attitudes (2023, 5 (NEW PAGENUMBER)). Finally, Annalisa Coliva offers the following thought experiment:

> Take a subject who is able to judge that P, give evidence in favour of it and withdraw from it if required and, therefore, has the first-order belief that P based on judgement. Suppose you ask her "Do you believe that P?" and she is unable to answer. You conclude that she does not have the concept of belief. (2016, 191)

This is a situation in which one interlocutor reasons interpersonally while, *ex hypothesi*, lacking the conceptual wherewithal to self-ascribe the attitudes that her assertions express. Coliva adds that any such agent will at least possess "the *ability* to differentiate between, for instance, believing P and P's being the case, by being *sensitive to the fact that* her point of view may be challenged [...]" (2016, 192, emphasis mine). On my reading, the emphasized terms suggest that such an agent utilizes *pre-metarepresentational* capacities in the service of interpersonal reasoning; these abilities and sensitivities enable her to reason with others without forming second-order beliefs about her first-order beliefs or her interlocutor's first-order beliefs.

These philosophers clearly reject Greco's claim that "thinking about one's first-order mental states is essential to activities involving coordination and cooperation," given that interpersonal reasoning is itself a coordinated and cooperative endeavour. But do they extend this rejection as far as to deny that interpersonal reasoning *with an aim to justifying one's own actions and attitudes* requires self-knowledge or, at the very least, some form of self-representation like a self-belief? As Steven Levine makes clear in a response to Brandom, it is hard to see how they could cogently deny this. Levine begins by acknowledging the possibility of agents who reason interpersonally insofar as the assertions at issue are first-order assertions of the form "that-P," these being expressions of agent-neutral propositions in the sense described above. As

on what epistemologists in this area ordinarily focus on, i.e., *explicit* self-knowledge involving linguistically articulate self-ascriptive thought (*pace* also those who view self-knowledge as an *ability*—cf. Campbell (2018)).

doi: 10.48106/dial.v76.i4.02

regards assertions of these propositions, “the performer can justify the statement without explicitly claiming that it is he who is justifying the statement […] because this assertion concerns an objective state of affairs that can be justified by agent-neutral reasons” (2009, 111). However:

> […] is this the case when the assertion that is being challenged concerns an agent’s *own* action or perception? Here what is being challenged is, for example, one’s entitlement to perform an action or one’s entitlement to claim that one’s perception is veridical. In either case, the justificatory reasons offered cannot be agent-neutral in the way that reasons justifying the assertion “that-P” are. (2009, 111)

So Levine is in league with Siewert and Greco in arguing that, when one’s own actions are challenged by an interlocutor, one cannot merely avail oneself of agent-neutral reasons. Instead, one must avail oneself of agent-specific reasons, which will include facts about one’s own psychology. The only question is whether Levine would side with Siewert in understanding these exchanges as requiring PPSK on the part of whoever seeks to justify her own attitudes, or with Greco in denying any indispensable role for such epistemically high-grade self-knowledge.

As aforementioned, I will soon argue that PPSK plays important roles in interpersonal reasoning about both agent-specific and agent-neutral propositions. But how can I be headed in this direction, having just traced a dialectic that only acknowledges a role for self-knowledge in interpersonally defending agent-specific propositions about one’s own actions or perceptions? In other words, if it is conceded to Brandom and others that agents can reason interpersonally about agent-*neutral* propositions without so much as a capacity for self-belief, then isn’t it foolish to contend that PPSK—let alone any other sort of self-knowledge—matters for such activity? Fortunately, there is no real problem here. My argument will be that PPSK contributes to interpersonal reasoning *for agents who in fact possess the capacity for representing themselves and their beliefs in higher-order thought*. This focus allows me to grant Brandom, Koreň, and Coliva their contention that some agents can reason interpersonally despite lacking this metarepresentational capacity.[7]

[7] There are other ways to dispute the indispensability of self-knowledge for interpersonal reasoning. For example, Roelofs (2017) argues that no such knowledge is required by interpersonal reasoners who are “evidentially unified” with and “cognitively vulnerable” to one another. Evidentially

What I will argue is that agents who *do* possess this capacity, such as most cognitively developed adult human beings, are systematically vulnerable to certain deficiencies in interpersonal reasoning if they lack PPSK.

2 Interpersonal Reasoning and Privileged, Peculiar Self-Knowledge

Oftentimes, cognitively developed adult human beings have knowledge—or at least beliefs—about their own attitudes, and they often have further beliefs about how their attitudes converge with or diverge from their peers. It is often these higher-order states of mind that motivate agents to reason with one another in the first place. After all, if one agent believes that there is a discrepancy between what she believes and what her interlocutor believes, this can help to explain why she bothers to try and settle the discrepancy through an interpersonal exchange of reasons.

For a hypothetical example, consider two interlocutors: Maya and Roman. Maya might aim to convince Roman that climate change is an existential threat to human civilization (note that this is an agent-neutral proposition: I emphasize the importance of this fact near the end of this section). My claim now is this: Maya would be in a precarious epistemic position, one that might undermine the efficiency of her reasoning with Roman, or one that might even make it better for her to *not* try to reason with Roman about this issue, if she did not possess PPSK.

Why so? It is easiest to begin by focusing on *privilege*. Here is the basic idea: if Maya were not in a systematically superior epistemic position regarding her beliefs about her attitudes than Roman was concerning his beliefs about Maya's attitudes, then Roman could more easily—i.e., with better epis-

unified agents are automatically attuned to one another's evidence without having to explicitly share it, while cognitively vulnerable agents can rationally cause changes in one another's minds through cognizing this unified evidence (they can induce such changes as *basic actions*). Evidential unity and cognitive seem conceptually possible, and they might even be achieved by actual agents who are wired to one another's brains in the right sorts of ways. The upshot is that neither party must have "I"-thoughts about their selves and attitudes in the course of interpersonal reasoning nor, for that matter, thoughts about others' selves and attitudes. Instead, by focusing strictly on first-order reasons, they can automatically adjust one another's attitudes. However, Roelofs admits that, for us, "it seems very unlikely…that such a close rapport could persist for very long, or cover very many topics" (2017, 17). We are simply not wired to one another in these ways, at least not with any real consistency. Accordingly, what I say below applies to ordinary agents who lack evidential unification and cognitive vulnerability.

doi: 10.48106/dial.v76.i4.02

temic grounds—convince Maya that her attitudes already align with his. In convincing Maya of this, Roman would be providing second-order grounds for skepticism about Maya's belief that she believes climate change to be an existential threat to human civilization. As a result, Maya would not even bother to reason with Roman about the first-order discrepancy, because her self-belief would change in such a way that she no longer took there to be any such discrepancy. Roman might alter Maya's self-belief in good faith by providing evidence that it is mistaken. But in other cases, Roman might operate in bad-faith by knowingly supplying Maya with misleading grounds for the same conclusion. Indeed, if Roman's testimony is a source of evidence all on its own then, given Maya's lack of privileged access to her own belief, her epistemic situation upon receiving Roman's testimony is immediately altered even if Roman supplies no independent evidence in favour of his testimony. In such cases we could say—perhaps somewhat overdramatically—that Maya has been taken as Roman's *epistemic hostage*. As an epistemic hostage, Maya succumbs to Roman's efforts (good faith or otherwise) to convince her that her self-ascribed attitudes are not really her own. Maya, being falsely convinced of this, is even cut off from opportunities to reason with agents *other than* Roman about climate change, given that she has been pre-emptively convinced that she does not disagree with those—like Roman—who are climate science deniers.

We might construe these situations as threats to Maya's epistemic autonomy. I say this because, plausibly, epistemic autonomy is at least partly a matter of being able to navigate various interpersonal reasoning contexts without having one's self-conception co-opted too easily by others. Indeed, this matters even if we are sometimes duped about the *first-order* issues by clever interlocutors who supply us with misleading evidence at *that* level of discourse (e.g., misleading statistics suggesting that climate change—of the anthropogenic variety, at least—is not taking place). An agent who is convinced by a clever interlocutor that the evidence for climate change is bad is still an agent who has assessed those reasons for herself and hence has been mislead on a basis that still deploys her own rational faculties to some degree. And while it is true that Maya might also deploy her own rational faculties in assessing Roman's claim that her *self-belief* is wrong, perhaps because Roman supplied her with good reasons (by *her* lights, at least) to do so, the *result* is that Maya lacks the self-knowledge that she needs in order to recognize that there is a discrepancy between her belief about climate change and Roman's belief about it, and *this* undermines her epistemic autonomy for reasoning with Roman about climate change itself.

Now, as aforementioned, this account of PPSK's instrumental value is most clearly geared toward *privileged* self-knowledge, since it is an argument about what happens when the epistemic security of Maya's self-beliefs is, as a general matter, no better than that of Roman's perspective unto Maya's mind. But the account can extend to *peculiarity* as well, at least if we construe the relationship between privilege and peculiarity in such a way that Maya's privilege *is due to* the peculiar way in which she knows her own mind (cf. Peterson 2021). For, if her self-beliefs are not generally acquired by a peculiar means that is generally available to her, then nothing prevents individuals like Roman from seizing upon the very same means to acquire knowledge of Maya's mind, and this makes it harder to understand why Maya's self-beliefs are, in general, so epistemically secure that Roman's contrary claims or beliefs do not give Maya strong reason to change what she believes about herself.

To bring this point into sharper relief, we can consider a putative foil for my account, namely Quassim Cassam's *Inferentialist* account of self-knowledge. According to Cassam, both self-knowledge and other-knowledge of agents' attitudes are acquired through inferences. On his view, there remains an epistemic asymmetry between self-knowledge and other-knowledge, but this asymmetry simply "boils down to a difference in the kinds of evidence that are available in the two cases" (2014, 150). More specifically, the evidence that one has about one's own attitudes is superior to the evidence that one has about others' attitudes because it includes sensations, memories, and other non-attitudinal mental goings-on that are not so easily accessed by one's peers. Applying this view to interpersonal reasoners like Maya, we might say that Maya's self-knowledge of her attitudes is privileged to some degree even if the same method—inference—is used by both Maya and Roman in coming to form beliefs about Maya's attitudes. So there is nothing peculiar about Maya's route to self-knowledge. But now one might insist that Maya cannot be easily taken as an epistemic hostage by Roman, even though she lacks a peculiar way of knowing herself, simply because she has especially good evidence about herself.

However, it could happen that such additional evidence is unavailable to Maya in any number of cases, for what reason can be given for thinking that Maya will always have access to special evidence, given that access to evidence in general is a contingent matter of one's epistemic position relative to a body of information? Peter Carruthers—another prominent Inferentialist—takes it that we have privileged self-knowledge of non-propositional-attitudinal mental states (2011), and contends that this can be used as a basis for inferring our propositional attitudes. However, privileged access to these other mental states can only provide a basis for inferring our propositional attitudes *when we are in such mental states*, and yet this itself is a contingent matter. Moreover, even stipulating that Maya has systematically better evidence about herself than Roman has about her, we would also need a general assurance that Maya infers the correct conclusions from this systematically superior evidence. It could happen that Maya has privileged access to the evidence about what she herself believes but cannot reliably *utilize* this evidence. At the very least, it could happen that she is, in general, no better at utilizing this evidence than Roman is at utilizing *his* evidence about Maya's attitudes. Indeed, philosophers like Carruthers seem to embrace this point when they claim that Inferentialist views best explain failures of self-knowledge.

Finally, Inferentialist views are vulnerable to what I call an *efficiency concern* and a *gridlock concern*. The efficiency concern is that, absent peculiar access, it could be generally appropriate for Roman to ask Maya to supply the grounds for her self-beliefs, and for Maya to ask Roman to do the same, just to be sure that they were operating in a case where Maya really did have (and had effectively utilized) this superior evidence. Engaging in this second-order interpersonal reasoning would significantly slow down their efforts to get to the first-order issues, thus rendering interpersonal reasoning about first-order issues a less efficient activity. The gridlock concern is that the second-order issue might not get resolved at all whenever both parties fail to reach a verdict about what Maya believes. One might attempt to circumvent these concerns by arguing that Maya's inferences are subpersonal or non-conscious, such that she cannot be expected to articulate them to Roman. But inferences that are not available for peer-review are also inferences that Maya might be required to lower her trust in, thus calling her self-beliefs into question all over again. To be sure, if some sort of Inferentialism is true, it may follow that agents like Maya often have better evidence and draw better inferences about their own attitudes than their interlocutors can draw about her attitudes, but the points I have been making suggest that such access will be *worse* for Maya than any form of access that renders the special epistemic security of her self-beliefs a non-contingent matter.

Now, even though I have been critiquing an Inferentialist rejection of peculiar access, I want to reiterate a general lesson for all would-be skeptics about such access. The lesson is that, if the same method—whether inferential or otherwise—is used for acquiring both self-knowledge and other-knowledge, then epistemic privilege will seem to be highly contingent. For, if two agents can come to know one agent's mind by the same means, then there need be no systematic barrier to their doing with equal epistemic pedigree. In the context of my account, this would mean that there is no strong assurance that agents are systematically warranted in retaining their self-beliefs when challenged by their interlocutors. And this, in turn, would mean that there is no general assurance that interpersonal reasoning *about the world*, rather than about the interlocutors' minds, can proceed smoothly. The efficiency and gridlock concerns also generalize: if Maya and Roman share the same method for arriving at a view about Maya's mind, then Roman might endeavour to interrogate Maya about whether their current context is one in which she has exercised the method more effectively, whether the method is inferential or not. This would slow down and (potentially) gridlock the discourse at the second-order level. Crucially, though, I am not claiming that the systematic protection provided by PPSK against these concerns is universal in scope. For my purposes, PPSK's instrumental value will have been demonstrated if it is our standard sort of self-knowledge. This would ensure that one is not *systematically, generally*, or *universally* vulnerable to innocent-yet-erroneous self-belief change, bad-faith epistemic hostage-takers, or to the efficiency and gridlock concerns, thus improving interpersonal reasoning's reliability and efficiency as a tool in our social-epistemic toolkit for understanding our shared world.

So goes my account. If correct, it shows that PPSK is instrumentally valuable for interpersonal reasoning, at least among those who are in a position to form beliefs about their own attitudes in the first place (again, a child who has yet to acquire the concept of belief cannot be erroneously convinced that she *shares a belief* with someone else). Notably, the account applies whether we imagine interpersonal reasoners as aiming to debate an agent-neutral proposition or an agent-specific one. I initially described Maya as aiming to convince Roman that climate change is an existential threat—this being an agent-neutral proposition—whereupon Roman steers the discourse to the second-order level in order to convince Maya that she does not really believe this in the first place. But the content of the proposition was incidental to the example. Had the proposition's content been agent-specific, e.g., about Maya's particular climate-focused actions or the belief-desire pairs that rationalize her actions, Roman might have proceeded in the same way. So, my account has

doi: 10.48106/dial.v76.i4.02

a broader scope than Siewert's: it applies to agent-specific *and* agent-neutral interpersonal reasoning.

3 Addressing Objections

In this section I reinforce my account by addressing four objections. The first objection is that legitimate challenges to our self-knowledge are in fact quite frequent, and that this provides evidence against the claim that PPSK frequently serves as an epistemic shield against erroneous self-belief change in our actual interpersonal reasoning practices. The second objection is that PPSK does not suffice to ensure that interpersonal reasoning is a reliable route for rational attitude adjustments.[8] The two final objections are specific defenses of the claim that factors beyond PPSK can protect interpersonal reasoners against erroneous self-belief change in interpersonal reasoning contexts.

The first objection turns on familiar cases of self-deception. Self-deception is ordinarily taken as a failure of self-knowledge in which an agent self-ascribes an attitude that she in fact lacks. Those who take us to have privileged self-knowledge surely ought to say something about this familiar phenomenon. If one does not take privileged access to be universal in scope, then it is at least logically possible to accommodate such cases. Alternatively, one might deny the ordinary view of self-deception by arguing that it does not involve false self-ascriptions (Bilgrami 2006; Coliva 2016). The apparent trouble for my account, however, is that accusations of self-deception are frequent and potentially epistemically legitimate in many cases, and yet these might be precisely the moves that our interlocutors use in order to convince us that our self-beliefs are false. If accusations of self-deception are epistemically legitimate and widespread, and if these accusations can spur agents to adjust their self-beliefs, then what protection does PPSK really provide here?

To begin my response, I want to reiterate a point from my introductory remarks about the dialectical ambitions of this paper, namely that readers need not be convinced that I have unearthed PPSK's actual functional role for interpersonal reasoners at this world. Secondly, when I say that PPSK provides an epistemic shield against challenges to one's self-beliefs in interpersonal reasoning, I do not deny that people might sometimes fail to take advantage of this shield—PPSK offers *epistemic* protection that may not be

[8] These first two objections were put to me by Rachel Cooper.

psychologically appreciated. Beyond these somewhat concessionary responses, the devil must reside in the details, since any further response depends on how we understand the cases at issue. Thus, consider a case in which Maya avows a love of comic books and Roman replies: “you do not love comic books; you’ve just tricked yourself into thinking that loving comic books makes you interestingly different.”[9] What might bring Maya to accept this accusation? Well, Maya might fixate on the thought that her interlocutor has better evidence about her mind than Roman has about it. If she wondered about her own evidence, and wondered about its inferential role in supporting her self-beliefs, she would be supposing her own self-ascription to be vulnerable to the same epistemic standards that Roman uses to evaluate her self-beliefs. If her self-knowledge is peculiar, however, she will not fixate on this thought, because her self-ascription is *not* based on the same epistemic standards.

In fact, our actual manner of proceeding tends to bear this out: one’s interlocutor judges one to be self-deceived about one’s love of comic books, and one responds *not* by attempting (and possibly failing) to offer higher-quality evidence *about what one believes*, but by offering reasons about *why comic books are loveable*. Indeed, one possible *explanation* of privilege and peculiarity is that one’s own take on the reasons for or against adopting some attitude (typically) determine one’s adoption of it. And if one self-ascribes this attitude with full knowledge of the first-order reasons that one takes to support it, one is entitled to make this self-ascription even if other people have evidence contravening one’s self-ascription (Bilgrami 2006; Coliva 2016).[10]

Moreover, if we have PPSK, *other* challenges to our self-beliefs may also be illegitimated, these being challenges where other agents do not accuse us of being self-deceived but, rather, as having made innocent (or “brute”[11]) errors about ourselves—errors that could only be made on the basis of innocent inferential or observational mistakes.

Here is another, final sense in which the devil is in the details. The objection under consideration is that accusations of self-deception are common, and that these accusations might frequently lead to (reasonable) changes in one’s self-beliefs. However, while such cases may indeed be common, they may only be common in the sense that *all of us* are *occasionally* susceptible to

[9] I owe this example to Rachel Cooper.

[10] Compare Schwengerer’s verdict on two cases he discusses (2021, 12). What I may owe my interlocutor, in this case, is an explanation of how my actions fail to live up to my self-ascribed attitude, *not* an explanation to the effect that the evidence shows that I have this attitude.

[11] For the operative notion of brute error, see Burge (1996) and Bar-On (2004).

them. On this explanation of their commonality, no single agent is liable to be the reasonable target of an overwhelmingly large number of self-deception accusations. There is something suspicious about anyone, even one's therapist, who would unrelentingly accuse one of self-deception across myriad cases by saying things like "you do not believe that $-P$, nor hope that $-Q$, nor desire to ϕ, nor love S…". This suspicion may well reflect a fact about us: that we have enough PPSK to be reasonable in *not* giving in to too many accusations of self-deception—accusations which, if legitimate, would force us to change our self-beliefs.

The second objection to my account is that PPSK does not improve the reliability of interpersonal reasoning even if it provides us with epistemic warrant to ignore (many) accusations of mistaken self-belief. Cases in favour of this objection are easy enough to set up. For example, maybe Maya claims that climate change is an existential threat to human civilization and Roman gives insufficient epistemic uptake to her assertion because he is prejudiced against women. Indeed, in this case, Maya may be the victim of a "testimonial injustice" (Fricker 2007). But I want to offer two observations. First, although the factors preventing Roman from reasoning with Maya have nothing to do with Maya's self-knowledge or Roman's beliefs about Maya's self-perspective, this does not change the fact that Maya would have an *additional* problem on her hands if Roman were generally in an epistemic position to make Maya erroneously change her self-beliefs. Second, to the extent that Roman's prejudiced behaviour does not prevent Maya from knowing herself, she is still in a position to congregate with less prejudiced individuals and to reason with *them* (or even to reason with Roman *indirectly* by reasoning with someone that Roman is *not* prejudiced against, and getting that individual to convey Maya's reasons to Roman). This point also applies to another concern, namely that Roman might simply say that *he* agrees with Maya when *he* does not (this being an inverse version of the epistemic hostage-taking tactic). Maya may not be able to rationally challenge this claim if Roman has PPSK, unless she has reason to deem him insincere, since she will then have strong reason to take Roman at his word. Once again, though, this would not put Maya in the position of being made to form a false belief about what she herself believes about the world, and hence she would not be prevented from discoursing with other agents about the contents of her beliefs about the world.

I now address two objections to the effect that something other than PPSK can explain why we are protected against epistemic hostage-taking. According to the first objection, what *really* protects Maya against Roman's nefarious machinations is the same thing as what explains her *first-person authority*, where what explains *this* is something other than PPSK. Roughly, "first-person authority" denotes two claims: (1) it is epistemically rational to presume the truth of speakers' present-tense self-ascriptions of mental states, and (2) it is typically epistemically irrational to interrogate the epistemic grounds of speakers' present-tense self-ascriptions (hereafter just 'self-ascriptions).[12] Now consider an "expressivist" explanation of first-person authority which contends that speakers' self-ascriptions ought to be presumed true and be insulated from requests for epistemic support because they express and thus *show* the self-ascribed mental states to one's hearers (Bar-On 2004). This explanation is available even if speakers do not *also* possess PPSK of the mental states that their self-ascriptions express. The objection, then, is that Maya's first-person authority gives Roman a strong reason not to challenge most of her self-ascriptions, such that PPSK is explanatorily superfluous in explaining why Roman is not likely to give Maya an erroneous basis for changing her self-beliefs.

Now, for all I have said, Maya's self-ascriptions may be first-person authoritative in virtue of what they express, whether or not Maya also has PPSK of what they express. Nevertheless, I argue that without *also* possessing PPSK, Roman could purposefully *ignore* the first-person authority of Maya's self-ascriptions in a bid to convince her that her attitudes converge with rather than diverge from Roman's. He might (rightly) take Maya to have expressed her first-order belief through a self-ascription but still claim that her self-belief is false. Hence, PPSK protects Maya against being manipulated by bad faith interlocutors who ignore her first-person authority, *however* that is to be explained, because PPSK ensures the general (and systematically superior) reliability of her self-beliefs relative to Roman's beliefs about her attitudes. PPSK is what gives Maya an epistemic warrant for holding steadfast against his machinations, even if he was already unjustified in challenging her self-ascription challenged her self-ascription in the first place.[13] Moreover, PPSK protects Maya against erroneous self-belief change even if Roman, innocently, fails to recognize that her self-ascription expresses the very attitude that it is about.

[12] See Doyle (2021) and Winokur (2022) for more precise articulations of these claims.

[13] I take expressivism to have brighter prospects than Schwengerer (2021) does, though I also agree with him that not *everything* epistemically interesting about mental state discourse can be explained by first-order phenomena, hence the account given in this paper.

doi: 10.48106/dial.v76.i4.02

The final objection to my argument is that Maya can get away with merely *assuming* that she generally has PPSK, such that she is generally entitled to not defer to interlocutors who challenge her self-beliefs (whether in good or bad faith). More substantively, it might be argued that Maya possesses a distinctively strong *practical* warrant for holding steadfast when faced with accusations of mistaken self-belief, even if she lacks a distinctively strong *epistemic* warrant for doing so.

The trouble with this objection is that it is hard to see what could ground Maya's practical warrant for holding steadfast if it is not really, at bottom, the same as (or itself grounded by) epistemic warrant for doing so. This is because a *merely* practical warrant here would go against her epistemic wellbeing in any number of cases. Specifically, if she did not systematically know herself better than others know her, then she *would* often—perhaps even typically—have an epistemic reason to discourse with others about whether her self-beliefs are true, and this would be in tension with her practical warrant for avoiding such discourse. In other words, it is only if Maya really has PPSK, thus having epistemic warrant for holding on to her self-beliefs, that holding steadfast against her interlocutors' countervailing assertions does not inadvertently prevent her from indulging many epistemically legitimate disagreements about what her attitudes are. It is only if she really has PPSK that *not* entering these disagreements is by and large good for her to do.

It might now be complained, relatedly, that I have merely established the importance of an especially strong epistemic warrant for our self-beliefs, but that this need not amount to PPSK. In other words, Maya might have an especially strong epistemic warrant for her self-beliefs, but these self-beliefs need not be especially reliably *true*.[14] Indeed, such warrant may also suffice for avoiding the efficiency and gridlock concerns described in section 2. But I think a similar response applies here. For, if Maya has especially strong epistemic warrant for her self-beliefs but this warrant does not amount to self-knowledge in at least most of the cases in which she possesses this warrant, then in any number of cases she will still miss out on an epistemic good—that of *true* warranted self-belief—whenever she declines to engage with interlocutors who claim that her self-beliefs are false. Moreover, it is hard to understand how she could possess this special epistemic warrant for her self-beliefs if she did not actually possess self-knowledge in most of those same cases. After all, this would be tantamount to having epistemic

[14] Thanks to an anonymous reviewer for this objection.

warrant for self-beliefs that were not correspondingly likely to be true, and this systematic mismatch between truth and warrant would call the warrant itself into question.[15]

4 Interpersonal Reasoning and Epistemic Control

In section 2-3 I argued that PPSK provides us with widespread (even if not universal) protection against situations in which others provide epistemic reasons for us to change our self-beliefs, whether our interlocutors are operating innocently or in bad faith, and that this helps to ensure the efficiency and reliability of interpersonal reasoning. In this section I show that my account complements another recent account of PPSK's instrumental value.

According to Jared Peterson, PPSK is instrumentally valuable because it facilitates "epistemic control," which is a matter of being able to "keep private or disclose particular facts about one's mind to others" (2021, 368). Take privacy first. If you have PPSK, then you can reliably conceal your attitudes from others. For example, a teacher might fail to motivate a student's learning if the student knows that the teacher is pessimistic about the student's progress. But if the teacher has PPSK of her pessimism, then she has greater epistemic control over whether the student discovers this. Therefore, the teacher has greater control over the student's motivation to continue studying. For an example about disclosing rather than concealing one's mental states, Peterson says that "[a]n estranged lover might want a former partner to know in a highly epistemically secure manner that she still loves him" (2021, 369).

He also says that epistemic control:

> [...] allows societies to function in a much more productive, organized, and amicable way. When we accomplish group objectives in an efficient and peaceful manner we do so in large part by keeping private that which would be counterproductive to the group's efforts, and/or revealing our thoughts, beliefs, desires, etc. that are valuable for other members of a group to know. (2021, 371)

Peterson and I are both happy to emphasize the social importance of PPSK. I am also willing to say that PPSK provides a way to reliably disclose one's

[15] This response is similar in structure to one pursued by Davidson (1991) regarding perceptual belief warrant, though I believe that the strength of our warrants for perceptual beliefs and self-beliefs differ.

doi: 10.48106/dial.v76.i4.02

attitudes to others. However, Peterson does not also acknowledge the additional possibility of, e.g., *expressing one's love itself* as a reliable way of putting one's former partner in a secure position with respect to one's mind, where this expressive capacity may or may not depend on an agent's self-knowledge.

More significantly, though, I submit that PPSK's role as a shield against erroneous self-belief change is independent of its role in enabling one to disclose or conceal one's attitudes from others. To be able to better conceal one's attitudes is to prevent others from discovering what attitudes one has, but this may not matter to interlocutors who do not care (or are simply mistaken) about the facts and, instead, aim to convince you of a certain belief about yourself. Similarly, having an especially epistemic secure way of disclosing your attitudes is something that interlocutors could ignore (as argued in section 3 when discussing first-person authority). Thus, one may be tempted to deny that the instrumental value of PPSK for interpersonal reasoning is a matter of epistemic control.

However, one might be just as easily inclined to regard this as an instance of epistemic control after all, since my account claims that agents with systematically superior knowledge of their self-beliefs thereby exercise greater control over their social-epistemic lives as interpersonal reasoners. Accordingly, the lesson to be drawn may be that we ought to broaden our view of PPSK's contribution to epistemic control, such that epistemic control encompasses (i) control over which attitudes one discloses to others,[16] (ii) control over which attitudes one conceals from others, *and* (iii) control over which attitudes one is able to self-attribute, with especially secure epistemic warrant, in the face of disagreement about one's attitudes, while attempting to reason with others.

5 Conclusion

I have argued that privileged and peculiar self-knowledge contributes to our capacity for interpersonal reasoning about the world around us. To the extent that agents possess PPSK of their attitudes, interpersonal reasoning is a more reliable route to discursively navigating our shared world, and this explains one way in which PPSK is instrumentally valuable.

For the record, I happen to believe that phenomena like epistemic hostage taking are not widespread, and I regard PPSK as at least a partial explanation of this fact. I take myself, therefore, to have contributed to the debate

[16] Again, if this particular capacity requires PPSK at all.

between PPSK-skeptics and PPSK-realists not *merely* by illuminating the debate's stakes, but *also* by taking an anti-skeptical stand within that debate. This being said, I reiterate that others may not be persuaded to go as far as me in this regard, such that the core contribution of this paper is best viewed as an account of how being a PPSK-skeptic or PPSK-realist should affect one's corresponding conception of our interpersonal reasoning competencies.*

Benjamin Winokur
0000-0002-0845-9460
Ashoka University
ben.i.winokur@gmail.com

References

Andreotta, Adam J. 2021. "Confabulation Does Not Undermine Introspection for Propositional Attitudes." *Synthese* 198(5): 4851–4872, doi:10.1007/s11229-019-02373-9 .

—. 2022. "More than Just a Passing Cognitive Show: A Defence of Agentialism About Self-Knowledge ." *Acta Analytica* 37(3): 353–373, doi:10.1007/s12136-021-00492-y.

Bar-On, Dorit. 2004. *Speaking My Mind: Expression and Self-Knowledge*. Oxford: Oxford University Press, doi:10.1093/0199276285.001.0001.

Bilgrami, Akeel. 2006. *Self-Knowledge and Resentment*. Cambridge, Massachusetts: Harvard University Press, doi:10.2307/j.ctv1nzfgcn.

Boyle, Matthew. 2011. "Transparent Self-Knowledge." *Proceedings of the Aristotelian Society, Supplementary Volume* 85: 223–241, doi:10.1111/j.1467-8349.2011.00204.x.

—. 2019. "Transparency and Reflection." *Canadian Journal of Philosophy* 49(7): 1012–1039, doi:10.1080/00455091.2019.1565621 .

Brandom, Robert B. 1994. *Making It Explicit: Reasoning, Representing, and Discursive Commitment*. Cambridge, Massachusetts: Harvard University Press.

Burge, Tyler. 1996. "Our Entitlement to Self-Knowledge." *Proceedings of the Aristotelian Society* 96: 91–116. Reprinted in Burge (2013, 68–87), doi:10.1093/aristotelian/96.1.117.

—. 2013. *Cognition Through Understanding*. Philosophical Essays n. 3. Oxford: Oxford University Press, doi:10.1093/acprof:oso/9780199672028.001.0001.

* Thanks are due to Adam Andreotta for his comments on an earlier version of this paper, and to Rachel Cooper for reading and commenting on a shortened version of it at the 2021 American Philosophical Association Pacific Meeting. Thanks are also due to the other audience members at that meeting, to audience members at the 2021 Canadian Philosophical Association Meeting, and to audience members at the 2022 Meeting of the International Social Ontology Society. Finally, special thanks are due to three anonymous reviewers at *Dialectica* for their many useful comments.

Byrne, Alex. 2018. *Transparency and Self-Knowledge*. Oxford: Oxford University Press, doi:10.1093/oso/9780198821618.001.0001.

Campbell, Lucy. 2018. "Self-Knowledge, Belief, Ability (and Agency?) ." *Philosophical Explorations: An International Journal for the Philosophy of Mind and Action* 21(3): 333–349, doi:10.1080/13869795.2018.1426779.

Carruthers, Peter. 2011. *The Opacity of Mind. An Integrative Theory of Self-Knowledge*. Oxford: Oxford University Press, doi:10.1093/acprof:oso/9780199596195.001.0001.

Cassam, Quassim. 2014. *Self-Knowledge for Humans*. Oxford: Oxford University Press, doi:10.1093/acprof:oso/9780199657575.001.0001.

Coliva, Annalisa. 2016. *The Varieties of Self-Knowledge*. Innovations in Philosophy. London: Palgrave Macmillan, doi:10.1057/978-1-137-32613-3.

Davidson, Donald. 1991. "Epistemology Externalized." *Dialectica* 45(2–3): 191–202. Reprinted in Davidson (2001), doi:10.1111/j.1746-8361.1991.tb00986.x.

—. 2001. *Subjective, Intersubjective, Objective. Philosophical Essays Volume 3*. Oxford: Oxford University Press, doi:10.1093/0198237537.001.0001.

Doyle, Casey. 2021. "There's Something About Authority." *Journal of Philosophical Research* 46: 363–374, doi:10.5840/jpr2021816169.

Fricker, Miranda. 2007. *Epistemic Injustice. Power and the Ethics of Knowing*. Oxford: Oxford University Press, doi:10.1093/acprof:oso/9780198237907.001.0001.

Gopnik, Alison. 1993. "How We Know Our Minds: The Illusions of First-Person Knowledge of Intentionality." *Behavioral and Brain Sciences* 16(1): 1–14, doi:10.1017/S0140525X00028636.

Greco, John. 2019. "The Social Value of Reflection." in *Thinking About Oneself: The Place and Value of Reflection in Philosophy and Psychology*, edited by Waldomiro J. Silva-Filho and Luca Tateo, pp. 45–58. Philosophical Studies Series n. 141. Cham: Springer Verlag, doi:10.1007/978-3-030-18266-3_4.

Keeling, Sophie. 2018. "Confabulation and Rational Obligations for Self-Knowledge." *Philosophical Psychology* 31(8): 1215–1238, doi:10.1080/09515089.2018.1484086.

—. 2019a. "The Transparency Method and Knowing Our Reasons." *Analysis* 79(4): 613–621, doi:10.1093/analys/anz031.

—. 2019b. "Knowing Our Reasons: Distinctive Self-Knowledge of Why We Hold Our Attitudes and Perform Actions." *Philosophy and Phenomenological Research* 102(2): 318–341, doi:10.1111/phpr.12655.

Koreň, Ladislav. 2023. "Have Mercier and Sperber Untied the Knot of Human Reasoning?" *Inquiry* 66(5): 849–862, doi:10.1080/0020174X.2019.1684988.

Levine, Steven M. 2009. "Expressivism and I-Beliefs in Brandom's *Making it Explicit*." *International Journal of Philosophical Studies* 17(1): 95–114, doi:10.1080/09672550802614786.

Marcus, Eric and Schwenkler, John. 2019. "Assertion and Transparent Self-Knowledge." *Canadian Journal of Philosophy* 49(7): 873–889, doi:10.1080/00455091.2018.1519771.

Müller, Jean Moritz. 2019. *The World-Directedness of Emotional Feeling. On Affect and Intentionality*. London: Palgrave Macmillan, doi:10.1007/978-3-030-23820-9.

Musholt, Kristina. 2015. *Thinking About Oneself: From Nonconceptual Content to the Concept of a Self*. Cambridge, Massachusetts: The MIT Press, doi:10.7551/mitpress/9780262029209.001.0001 .

Nguyen, A. Minh. 2015. "What Good is Self-Knowledge? ." *Journal of Philosophical Research* 40: 137–154, doi:10.5840/jpr2015111656.

Parent, Ted. 2017. *Self-Reflection for the Opaque Mind. An Essay in Neo-Sellarsian Philosophy*. London: Routledge, doi:10.4324/9781315618449.

Peterson, Jared. 2021. "The Value of Privileged Access." *European Journal of Philosophy* 29(2): 365–378, doi:10.1111/ejop.12594.

Roelofs, Luke. 2017. "Rational Agency without Self-Knowledge: Could 'We' Replace 'I'?" *Dialectica* 71(1): 3–33, doi:10.1111/1746-8361.12169.

Schwengerer, Lukas. 2021. "Beliefs Over Avowals: Setting Up the Discourse on Self-Knowledge." *Episteme* 18(1): 66–81, doi:10.1017/epi.2018.56.

Shoemaker, Sydney S. 1988. "On Knowing One's Own Mind." in *Philosophical Perspectives 2: Epistemology*, edited by James E. Tomberlin, pp. 189–209. Oxford: Basil Blackwell Publishers. Reprinted in Shoemaker (1996, 25–49), doi:10.2307/2214074.

—. 1996. *The First-Person Perspective and Other Essays*. Cambridge: Cambridge University Press, doi:10.1017/CBO9780511624674.

Siewert, Charles. 2003. "Self-Knowledge and Rationality: Shoemaker on Self-Blindness." in *Privileged Access: Philosophical Accounts of Self-Knowledge*, edited by Brie Gertler, pp. 131–146. Ashgate Epistemology and Mind Series. London: Routledge.

Sorgiovanni, Ben. 2019. "The Agential Point of View." *Pacific Philosophical Quarterly* 100(2): 549–572, doi:10.1111/papq.12263.

Strijbos, Derek W. and Bruin, Leon C. de. 2012. "Making Folk Psychology Explicit: The Relevance of Robert Brandom's Philosophy for the Debate on Social Cognition ." *Philosophia* 40(1): 139–163, doi:10.1007/s11406-010-9288-z.

Valaris, Markos. 2018. "Self-Knowledge ." in *The Philosophy of Knowledge: A History, Volume 4: Knowledge in Contemporary Philosophy*, edited by Stephen Cade Hetherington and Markos Valaris, pp. 155–174. London: Bloomsbury Academic.

Winokur, Benjamin. 2021a. "Critical Reasoning and the Inferential Transparency Method." *Res Philosophica* 98(1): 23–42, doi:10.11612/resphil.1967.

—. 2021b. "Inference and Self-Knowledge." *Logos & Episteme* 12(1): 77–98, doi:10.5840/logos-episteme20211214.

—. 2022. “There is Something to the Authority Thesis.” *Journal of Philosophical Research* 47: 115–132, doi:10.5840/jpr202298189.

The Dual Detector Argument against the Modal Theory

Dan Marshall

The modal theory holds that facts (properties) are identical iff they are necessarily equivalent (coextensive). One of the most prominent arguments against the modal theory is Elliot Sober's dual-detector argument. According to this argument, the fact that some particular thing is a triangle is distinct from the necessarily equivalent fact that it is a trilateral, since it is only the former fact that causes an output of a certain machine. I argue that the dual-detector argument fails, in part because whatever initial plausibility it has relies on the failure to take into consideration a needed relativisation to times and the failure to distinguish between two facts collectively causing a fact and their conjunction singly causing it. I also argue that variants of the argument are equally unsuccessful.

One of the most popular and well known accounts of the identity-conditions of facts and properties is the modal theory.[1] According to this theory: i) two facts are identical iff they are necessarily equivalent to each other; and ii) two properties are identical iff they are necessarily coextensive to each other. That is, the modal theory holds that: i) the fact that ϕ = the fact that ψ iff, necessarily, (ϕ iff ψ); and ii) the property of being F = the property of being G iff, necessarily, for any x, (x is F iff x is G).[2] This theory is prima facie attractive, since it is simple to formulate and provides an account of the

[1] A fact, as I will understand it here, is an obtaining state of affairs, where: i) a state of affairs is either a way things are or a way things aren't, and ii) a state of affairs obtains iff it is a way things are. A fact on this understanding is therefore just as way things are. Proponents of the modal theory include Stalnaker (1984), Lewis (1986a) and Jackso (1998).

[2] For simplicity, I will assume necessitism, according to which, necessarily, for any x, necessarily, for some y, $x = y$. (Without this assumption, "necessarily, for any x" needs to be replaced with "necessarily, for any x, necessarily" in the above characterisation of the modal theory.) I will also assume an abundant theory of facts and properties according to which (except for restrictions needed to avoid paradox), all true sentences express facts, and all predicates that can be used to form true or false sentences express properties.

identity-conditions of facts and properties in terms of (at least relatively) well understood notions. Everything else being equal, the theory is also more parsimonious than rival theories that reject it, since, everything else being equal, there are less facts if the modal theory holds than if it fails to hold and there are distinct facts that are necessarily equivalent to each other.

A prominent argument against the modal theory is the dual-detector argument originally due to Elliot Sober.[3] Briefly, according to this argument, there could be a machine that, as a result of containing detectors measuring different aspects of an input, is causally sensitive to one fact without being causally sensitive to another necessarily equivalent fact. Since, by Leibniz's law, it follows from this that, contra the modal theory, there are distinct facts that are necessarily equivalent to each other, the argument concludes that the modal theory is false. Despite this argument's prominence, discussions of the argument by both its proponents and opponents have been brief and cursory. This paper will provide a more sustained evaluation of the dual-detector argument and will argue that such an evaluation shows that the argument is unsuccessful.[4]

I will proceed as follows. In section 1, I will formulate the dual-detector argument before then arguing in section 2 that it is unsuccessful. In section 3, I will then consider two variants of the argument and I will argue that these variants, and more generally that all variants, are also unsuccessful.

Before proceeding to section 1, it will be useful to briefly discuss another common argument against the modal theory—the constituency argument—in order to set it aside.[5] Suppose (1) and (2) are true, where "*W*" refers to some particular wire.

1. *W* is a closed straight-sided figure that has three *angles*.

2. *W* is a closed straight-sided figure that has three *sides*.

[3] (See Sober 1982). A similar argument is also given by (Enç 1982). A recent proponent of the dual-detector argument, for example, is (Audi 2016). Other philosophers who are sympathetic to the argument include (Miller 1995, 859) and (Molnar 2003, 66). Opponents of the argument include (Jackson 1998, 125–126) and (Armstrong 1997, 145–146).

[4] Two other arguments against the modal theory that appeal to causation have been given by (Achinstein 1974) and (Perry 1989). (Sober 1982, 84–85) gives what I take to be a convincing response to Achinstein's argument. For a response to Perry's argument, (see Marshall 2021).

[5] (See, for example, Audi 2016).

According to the constituency argument, since the fact expressed by (1) has angularity as a constituent while the fact expressed by (2) doesn't have this property as a constituent, the facts expressed by (1) and (2) are not identical to each other. Since the modal theory entails that the facts expressed by (1) and (2) are identical to each other (since they are necessarily equivalent to each other), the constituency argument concludes from this that the modal theory is false.

The constituency argument arguably begs the question against the modal theory by in effect assuming the rival structured theory of facts. According to this rival theory, facts are structured in the same kind of way that sentences are structured. In particular, according to the structured theory, facts are built up out of objects, properties, relations, operators and quantifiers in the same way that sentences are built up out of names, predicates, operator expressions and quantifier expressions.[6] If the structured theory holds so that the facts expressed by (1) and (2) are built up out of objects, properties, relations, operators and quantifiers in the same kind of way that sentences are built up out of names, predicates and other expressions, then it is plausible that the fact expressed by (1) has angularity as a constituent while the fact expressed by (2) doesn't have this constituent. This is much less plausible, however, if the structured theory is false and facts aren't structured like sentences. For example, if facts are instead structured like visual experiences or pictures, then, since it is prima facie plausible to associate (1) and (2) with the same (type) of visual experience or picture, it is prima facie plausible that (1) and (2) express the same fact and hence prima facie plausible that the facts expressed by (1) and (2) don't differ in what constituents they have. (This is because it is at least prima facie plausible that any picture that represents W as being a closed straight-sided figure that has three *angles* also represents W as being a closed straight-sided figure that has three *sides*, and vice versa.) Since the argument from constituency provides no reason to think that facts are structured in the way that the structured theory holds that they are structured, rather than some other way, the argument therefore fails to provide a good reason to think

[6] The structured theory can be formulated more precisely as a thesis endorsing schemas such as (PS) and (OS) (see, for example, Dorr 2016, 58–59).

(PC) For any x and y, if the fact that x is F = the fact that y is G, then: i) $x = y$, and ii) the property of being F = the property of being G.

(OS) If the fact that $\pi_1(\phi_1)$ = the fact that$\pi_2(\phi_2)$, then: i) the operator of π_1 = the operator of π_2, and ii) the state of affairs of it being that ϕ_1 = the state of affairs of it being that ϕ_2.

that (1) and (2) express distinct facts and hence fails to provide a good reason to reject the modal theory.

It is important to appreciate that the structured theory is neither self-evident nor prima facie highly plausible, and hence it cannot simply be assumed to hold in the above argument from constituency without begging the question against the modal theory. Three brief reasons for this are the following: First, prior to investigation and argument, the claim that facts are structured like sentences is no more plausible than the claim that facts have some other type of structure, such as that of visual experiences or pictures. Second, while (1) and (2) arguably differ in their cognitive significance, since a linguistically competent person arguably might endorse one of them while rejecting the other, such a difference in cognitive significance is widely thought to be able to be explained by a difference in what mode of presentation the facts expressed by (1) and (2) have when expressed by these sentences, where this explanation does not require that the facts expressed by these sentences are non-identical.[7] Third, the structured theory conflicts with claims that are widely thought to be at least as prima facie plausible as the structured theory itself, such as the claim made by (3).

3. "*W* is self-identical" expresses the same fact as "*W* is identical to *W*."

(3) conflicts with the structured theory, since, if the structured theory is true, the fact expressed by "*W* is self-identical" has the property of being self-identical as a constituent while the fact expressed by "*W* is identical to *W*" lacks this constituent and instead has the property of being identical to *W* as a constituent.[8] Due to the above difficulty with the constituency argument, and since we cannot simply assume the structured theory in arguing against the modal theory, I will assume in the following that the constituency argument against the modal theory fails.

[7] McKay and Nelson (2010).

[8] A further possible consideration against the structured theory is that, unlike the modal theory, it gives rise to the Myhill-Russell paradox. Goodman (2017). For attempted solutions to the Myhill-Russell paradox that are compatible with the structured theory, see, for example, (Walsh 2016), (Kment 2022) and (Yu 2017). (See Dorr 2016) and (Bjerring and Schwarz 2017) for futher arguments against the structured theory.

1 The Dual-Detector Argument

The dual-detector argument is not meant to rely on the cogency of the constituency argument discussed above, nor is it meant to rely on the truth of the structured theory of facts. Instead, the dual-detector argument is meant to provide a separate reason for rejecting the modal theory. The argument involves a machine *M* that contains two detectors: a closed straight-sided figure detector and a three-angle detector. These detectors are linked in a series in *M*, so that, if a wire (or several wires) are inputted into *M*, they are first inputted into the closed straight-sided figure detector and then, if they are outputted by this first detector, they are inputted into the three-angle detector. If the wire (or wires) are then outputted by the three-angle detector, they are then outputted by *M*. Indeed, I will assume in the following that what it is for something (or some things) to be outputted by *M* is just for it (or them) to be outputted by this second detector.

The closed straight-sided figure detector in *M* works so that "when given a piece of wire as input, it will output the piece of wire if and only if the wire is a closed [plane] figure and all sides of the figure are straight" (Sober 1982, 185). More explicitly, let us say that: i) when given a piece of wire as input that is a closed figure all of whose sides are straight, the closed straight-sided figure detector outputs the wire, and it does this *because* the wire is a closed figure all of whose sides are straight; whereas, ii) when given a piece of wire (or several pieces of wire) as input that is not a single closed figure all of whose sides are straight, the closed-straight-sided figure detector does not output it (or them). The three-angle detector, on the other hand, works so that "when given any number of straight pieces of wire, it outputs them if and only if they have three angles" (Sober 1982, 185). More explicitly: i) when given one or more pieces of wire with straight sides that collectively have three angles, the three-angle detector outputs them and it does this *because* the wire (or wires) collectively have three angles; whereas, ii) when given one or more pieces of wire with straight sides that don't collectively have three angles, the three-angle detector does not output them. The three-angle detector is causally sensitive to whether the input has three angles, and not to whether it has three sides, since, when given a four-sided open figure, it will output the object (since it has three angles), and it will fail to do this if the four-sided figure is closed. In addition, when the three-angle detector is given three unconnected pieces of wire, each containing exactly one angle, the detector will output them, even though it is made up of six straight line segments.

Sober states the dual-detector argument as follows:

> Now consider a particular object—a piece of wire—which is fed into the machine, passes through both [detectors], and is then outputted by the machine. What property of the object *caused* it to be outputted? Given the mechanism at work here, I think that the cause was the object's having the property of being a *closed straight-sided figure having three angles* (i.e., its being a triangle), and not its being a *closed straight-sided figure having three sides* (i.e., its being a trilateral). If this is right, and if a difference in causal efficacy is enough to insure a difference in property, it follows that being a triangle is not the same property as being a trilateral, even though "triangle" and "trilateral" are logically (mathematically) equivalent. (Sober 1982, 185, Author's emphasis)

Let "[ϕ]" abbreviate "the fact that ϕ," and suppose that W is the piece of wire that is fed into M. Let us also suppose that the above process of W being fed into and then being sequentially outputted by the two detectors has occurred. Then, according to Sober's dual-detector argument, (Angle) is true while (Side) is false.

> Angle. [W is a closed straight-sided figure that has three *angles*] causes [M outputs W].

> Side. [W is a closed straight-sided figure that has three *sides*] causes [M outputs W].

The dual-detector argument then employs Leibniz's law to infer from this that, since they differ in what they cause, [W is a closed straight-sided figure having three *angles*] is not identical to the necessarily equivalent [W is a closed straight-sided figure having three *sides*]. The argument then infers from (4) and the non-identity of these facts that the property of being a closed straight-sided figure that has three *angles* (or being triangular) is not identical to the necessarily coextensive property of being a closed straight-sided figure that has three *sides* (or being trilateral).

4. For any x, IF x is F, x is G, and the property of being F = the property of being G, THEN [x is F] = [x is G].

Since these facts and properties are respectively necessarily equivalent to each other and necessarily coextensive with each other (and hence are identical to each other according to the modal theory), the dual-detector argument then concludes from the above results that the modal theory is false.[9]

2 Against the Dual-Detector Argument

One initial problem with the dual-detector argument is that (Angle) is not strictly speaking true, at least if we assume as we did above that the above described process involving W and M has already occurred.

> ANGLE. [W is a closed straight-sided figure that has three *angles*] causes [M outputs W].

To see why this is the case, let us suppose that, after being fed into M and put inside the closed straight-sided figure detector at t_1, W is outputted by the closed straight-sided figure detector so that, at t_2, W is inside the three-angle detector. Let us also suppose that W being inside the three-angle detector at t_2 results in W being outputted by the three-angle detector at t_3, and hence results in W being outputted by M at t_3. Finally, let us also suppose that the times t_1, t_2 and t_3 are all past times. Then the fact that W is a closed straight-sided figure that has three angles (either simpliciter or at the present time)

[9] I am assuming that facts can cause other facts, where this claim is compatible with it also being the case that events can cause other events. If it is instead held that it is only events that can be causal relata, then “fact” can be replaced with “event” in the above presentation of the dual-detector argument to get the conclusion that there are distinct necessarily equivalent events (where two events are necessarily equivalent iff, necessarily, they either both occur or they both fail to occur). This conclusion together with (A) entails that there are distinct necessarily equivalent properties which, given (MF), entails that there are distinct necessarily equivalent states of affairs.

(A) If the property of being F = the property of being G, then, for any x, the event of x having F = the event of x having G.

(MF) The property of being F = the property of being G iff, necessarily, for any x, the state of affairs of x being F = the state of affairs of x being G.

Taking facts to be obtaining states of affairs (as in footnote 1), it follows from this that there are distinct necessarily equivalent facts.

doi: 10.48106/dial.v76.i4.03

does not cause *M* to do anything to *W*, since *W* is no longer interacting with *M*.

The above problem with the dual-detector argument shows that, as it is most charitably understood, it is not (Angle) that is true according to the argument, but is instead either (Angle$_{t1}$) or (Angle$_{t2}$).[10]

ANGLE$_{t1}$. [*At* t$_1$, *W* is a closed straight-sided figure that has three *angles*] causes [*M* outputs *W* at t$_3$].

ANGLE$_{t2}$. [*At* t$_2$, *W* is a closed straight-sided figure that has three *angles*] causes [*M* outputs *W* at t$_3$].

As a result of this need to relativise to either time t_1 or time t_2, we therefore have two versions of the dual-detector argument. The first version—the t_1-version—holds that (Angle$_{t1}$) is true and (Side$_{t1}$) is false, from which it infers that, contra the modal theory, the necessarily equivalent facts [at t_1, *W* is a closed straight-sided figure that has three *angles*] and [at t_1, *W* is a closed straight-sided figure that has three *sides*] are non-identical.

SIDE$_{t1}$. [*At* t$_1$, *W* is a closed straight-sided figure that has three *sides*] causes [*M* outputs *W* at t$_3$].

The second version of the dual-detector argument—the t_2-version—holds instead that (Angle$_{t2}$) is true and (Side$_{t2}$) is false, from which it infers that, contra the modal theory, the necessarily equivalent facts [at t_2, *W* is a closed

[10] In response to the above problem with Sober (1982)'s original formulation of the dual-detector argument, we might modify *M* so that its two detectors act on *W* at the same time rather than sequentially. Such a modified version of the argument faces the same difficulties as the t_1-version of the argument discussed below. First, given this modification, while it is plausible that (Angle*) is true and (Side*) is false (when relativised uniformally to the relevant time), there is an exclusion argument that argues from the truth of (Angle*) to the falsity of (Angle).

(Angle*) [*W* is a closed straight-sided figure] and [*W* has three *angles*] collectively cause [*M* outputs *W*].

(Side*) [*W* is a closed straight-sided figure] and [*W* has three *sides*] collectively cause [*M* outputs *W*].

Second, this modified version of the argument faces the problem that, even if this exclusion argument is rejected, it doesn't seem possible to justify both the truth of (Angle) and the falsity of (Side).

straight-sided figure that has three *angles*] and [at t_2, *W* is a closed straight-sided figure that has three *sides*] are non-identical.

> SIDE$_{t2}$. [*At* t$_2$, *W* is a closed straight-sided figure that has three *sides*] causes [*M* outputs *W* at t$_3$].

As we will see, both these versions of the dual-detector argument have serious problems.[11]

The t_2-version of the dual-detector argument can be quickly seen to fail as follows: It is [*W* has three angles at t_2] that causes *W* to be outputted by the three-angle detector at t_3, rather than say [at t_2, *W* has three angles *and is blue*] that causes this fact (even supposing that *W* is blue at t_2). This is intuitively because [at t_2, *W* has three angles and is blue] goes beyond what is causally relevant to whether *W* is outputted by the three-angle detector at t_3. Similarly, it is [*W* has three angles at t_2] that causes *W* to be outputted by the three-angle detector at t_3 rather than [at t_2, *W is a closed straight-sided figure* that has three angles] that causes this fact. This is because the latter fact also goes beyond what is causally relevant to whether *W* gets outputted by the three-angle detector at t_3. Since *W* getting outputted by the three-angle detector just is what it is for *M* to be outputted by *W*, it follows that (Angle$_{t2}$) is false.

> ANGLE$_{t2}$. [*At* t$_2$, *W* is a closed straight-sided figure that has three *angles*] causes [*M* outputs *W* at t$_3$].

Since the falsity of (Angle$_{t2}$) conflicts with the t_2-version of the dual-detector argument, this version of the argument fails.

[11] There is also a temporally mixed version of the dual-detector argument that holds that (Angle$_{t1,t2}$) is true and (Side$_{t1,t2}$) is false.

(Angle$_{t1,t2}$) [*W* is a closed straight-sided figure at t_1 that has three *angles* at t_2] causes [*M* outputs *W* at t_3].

(Side$_{t1,t2}$) [*W* is a closed straight-sided figure at t_1 that has three *sides* at t_2] causes [*M* outputs *W* at t_3].

This version of the argument at best only shows that [*W* is a closed straight-sided figure at t_1 that has three *angles* at t_2] is not identical to [*W* is a closed straight-sided figure at t_1 that has three *sides* at t_2], which does not conflict with the modal theory since these facts are not necessarily equivalent to each other.

doi: 10.48106/dial.v76.i4.03

I will now argue that the t_1-version of dual-detector argument is also unsuccessful and hence that both versions of the dual-detector argument fail. I will do this by first giving an argument from causal exclusion that, contrary to the dual-detector argument, (Angle$_{t1}$) is false. I will then argue that, even if this causal exclusion argument is rejected, it is not possible to justify both the truth of (Angle$_{t1}$) and the falsity of (Side$_{t1}$), the justification of both of which is required for the t_1-version of the argument to be successful. (Or at least, I will argue that one cannot justify the truth of (Angle$_{t1}$) and the falsity of (Side$_{t1}$) without appealing to some other argument against the modal theory that, if successful, would refute the modal theory by itself and hence would render the dual-detector argument superfluous.)

To set up the needed background for the argument from causal exclusion against (Angle$_{t1}$), note that, in the case of M processing W, [W is a closed straight-sided figure at t_1] causes W to be outputted by the closed straight-sided figure detector, and so causes W to be in the three-angle detector at t_2. Hence we have (5)

5. [W is a closed straight-sided figure at t_1] causes [W is in the three-angle detector at t_2].

Since [W is in the three-angle detector at t_2] and [W has three angles at t_2] collectively cause W to be outputted by the three-angle detector at t_3, which is what it is to be outputted by M at t_3, we also have (6).

6. [W is in the three-angle detector at t_2] and [W has three angles at t_2] collectively cause [M outputs W at t_3].

Since plausibly one of the causes of W having three angles at t_2 is that it had three angles at previous times before t_2, (7) plausibly also holds.

7. [W has three angles at t_1] causes [W has three angles at t_2].

Assuming, as is plausible, that the causal transitivity principle (T) holds in this causal situation, (5-7) then entail (Angle*$_{t1}$).[12]

[12] While causation is plausibly transitive in many typical cases, such as in the case above, many philosophers hold that causation is not unrestrictedly transitive. For alleged counterexamples to transitivity, see, for example, (Kvart 1991) and (McDermott 1995). For a defense of transitivity unrestrictedly holding, (see Hall 2000).

T. IF the members of Φ_1 collectively cause r_1, the members of Φ_2 collectively cause r_2 … and r_1, r_2 … collectively cause r; THEN the members of $\Phi_1 \cup \Phi_2 \cup$ … collectively cause r.

ANGLE*$_{t1}$. [W is a closed straight-sided figure at t_1] and [W has three angles at t_1] collectively cause [M outputs W at t_3].

With the above background in place, it might seem like it should now be easy to derive (Angle$_{t1}$) from (Angle*$_{t1}$), and hence establish that (Angle$_{t1}$) holds.

ANGLE$_{t1}$. [*At* t$_1$, W is a closed straight-sided figure that has three *angles*] causes [M outputs W at t$_3$].

However using the above background, we can now give the following argument from causal exclusion that (Angle$_{t1}$) is instead false: Just as [at t_2, W is a closed straight-sided figure that has three angles] fails to cause the closed straight-sided figure detector to output W at t_3 (since the former fact goes beyond what is causally relevant), [at t_1, W is a closed straight-sided figure that has three angles] fails to cause the closed straight-sided figure detector to output W (since this fact also goes beyond what is causally relevant) and hence this fact fails to cause [W is in the three-angle detector at t_2]. Hence we have (8).

8. [at t_1, W is a closed straight-sided figure that has three angles] does not cause [W is in the three angle detector at t_2].

Similarly, while [W has three angles at t_1] is a cause of [W has three angles at t_2], it is not the case that [at t_1, W is a closed straight-sided figure that has three angles] causes this fact, since it goes beyond what is causally relevant. Hence we have (9).

9. [at t_1, W is a closed straight-sided figure that has three angles] does not cause [W has three angles at t_2].

Since [at t_1, W is a closed straight-sided figure that has three angles] is also not caused by either [W is in the three-angle detector at t_2] or [W has three angles at t_2], and the causal chain that leads up to [M outputs W at t_3] goes through [W is in the three-angle detector at t_2] and [W has three angles at t_2], it therefore follows from (8) and (9) that [at t_1, W is a closed straight-sided

doi: 10.48106/dial.v76.i4.03

figure that has three angles] isn't part of the causal chain that leads to [M outputs W at t_3] and hence does not cause it. Hence (Angle$_{t1}$) is false.

> Angle$_{t1}$. [*At* t$_1$, W is a closed straight-sided figure that has three *angles*] causes [M outputs W at t$_3$].

A more rigorous version of the above argument against (Angle$_{t1}$) can be given by appealing to the version of the principle of causal exclusion given by (PCE).[13]

> PCE. In cases where there is no genuine causal overdetermination, if S is a set of facts that occur at a time t whose members collectively completely cause f, then S is the unique set of facts that occur at t and collectively completely cause f.

In (PCE), a fact is said to occur at a certain time iff the fact only concerns how things are at that time. Genuine causal overdeterminism, on the other hand, occurs when two independent causal processes converge on the same effect, such as when a house burns down because a lit match starts a fire in the garbage at the same time as lightning strikes the house.

Since there is no genuine causal overdetermination in the case of W being outputted by M, (PCE) can be used to argue that (Angle$_{t1}$) is false as follows: Suppose, for reductio, that (Angle$_{t1}$) is true. Then [W is a closed straight-sided figure that has three angles at t_1] together with the members of some possibly empty set Ψ_1 completely cause [M outputs W at t_3]. Since (Angle*$_{t1}$) holds, it is also true that [W is a closed straight-sided figure at t_1], [W has three angles at t_1] together with the members of some possibly empty set Ψ_2 collectively completely cause [M outputs W at t_3]. Since the relevant facts occur at the same time, these two consequences together with (PCE) then entail (10).

> 10. [W is a closed straight-sided figure at t_1],[W has three angles at t_1], [W is a closed straight-sided figure that has three angles] and the members of some possibly empty set Ψ collectively completely cause [M outputs W at t_3].

[13] For discussion of the principle of causal exclusion, see, for example (Kim 2005) and (Moore 2018).

If (10) is true, then [W is a closed straight-sided figure at t_1], [W has three angles at t_1] and the members of Ψ by themselves collectively completely cause [M outputs W at t_3], since [at t_1, W is closed straight-sided figure that has three angles] is superfluous given the presence of [W is a closed straight-sided figure at t_1] and [W has three angles at t_1]. Given (PCE), however, this consequence conflicts with (10). Hence, the reductio assumption (Angle$_{t1}$) is false.

The above argument shows that (Angle$_{t1}$) fails to hold if (PCE) holds. Not all philosophers, however, accept (PCE), and these philosophers will not be convinced by the above argument from causal exclusion that the dual-detector argument fails. For example, some philosophers reject (PCE) on the grounds that it conflicts with the popular counterfactual dependency thesis (Dep).[14]

> Dep. Suppose that f and g obtain, and that, had f failed to obtain, it would have been that g failed to obtain. Then, f causes g.

Other philosophers reject (PCE) because they hold that, in cases where there is no genuine causal overdetermination of a fact, there can still be multiple complete causal chains that converge on that fact, provided these chains are systematically related to each other in the right way. In particular, some philosophers hold that there can be multiple such causal chains provided that, for each such chain, either that chain generates all the other chains, or that chain is generated by at least one other such chain. Someone who endorses this view, for example, might endorse (Conj).[15]

> conj. If f_1 and f_2 together with the members of a set Φ collectively completely cause f, then the conjunction of f_1 and f_2 together with the members of Φ collectively completely cause f.

It follows from (Conj) that, contra (PCE), if there is one causal chain leading to f that contains the facts f_1 and f_2 occurring at a time t, then there is a further causal chain which is systematically related to it by virtue of containing the

[14] (See, for example, Loewer 2007). Proponents of (Dep) typically place certain restrictions on (Dep), such as requiring that the counterfactual is to be read in a suitable non-backtracking sense (see Lewis 1973), that the facts (or events, when (Dep) is applied to events) that stand in the causation relation are "sufficiently distinct" (so that, for example, we don't have the consequence that each fact causes itself) Lewis (1986c), and that these facts (or events) are non-disjunctive (see Lewis 1986c).

[15] Φ in (Conj) can be the empty set.

conjunction of f_1 and f_2 instead of f_1 and f_2 themselves. Given (Conj), it is natural to hold that this further causal chain containing the conjunction of f_1 and f_2 is generated by the former chain containing its conjuncts.

In light of the above views, the argument from causal exclusion does not by itself decisively refute the t_1-version of the dual-detector argument. In addition to facing the argument from causal exclusion, however, the t_1-version of the dual-detector argument faces the problem that, even if the caual exclusion argument fails, it doesn't appear possible to justify the truth of (Angle$_{t1}$) while also justifying the falsity of (Side$_{t1}$). (Or at least, it doesn't seem possible to do this without relying on some other argument against the modal theory which, if successful, would by itself refute the modal theory. I will discuss two attempts to give such a justification, and I will argue that both these attempts fail. The failure of these two attempts will give us reason to think that no such justification is possible, and hence reason to think that, even if (PCE) and the argument from causal exclusion fail, the t_1-version of the dual-detector argument is still unsuccessful.

The first attempt to justify the truth of (Angle$_{t1}$) (while also justifying the falsehood of (Side$_{t1}$) appeals to (Conj) above. This first attempt accepts (Angle*$_{t1}$) on the basis of the transitivity reasoning given for it above. It then infers from (Angle*$_{t1}$) and (Conj) that the conjunction of [*W* is a closed straight-sided figure at t_1] and [*W* has three angles at t_1] collectively (partially) cause *M* to output *W* at t_3. Assuming (as I will from now on) that this conjunction is the fact [at t_1, *W* is a closed straight-sided figure that has three angles], it follows from this that (Angle$_{t1}$) is true.

> Angle$_{t1}$. [*At* t_1, *W* is a closed straight-sided figure that has three *angles*] causes [*M* outputs *W* at t_3].

Let us assume that the above justification of (Angle$_{t1}$) is successful. The question that now needs to be addressed is whether we can go on to justify the falsehood of (Side$_{t1}$).

> Side$_{t1}$. [*At* t_1, *W* is a closed straight-sided figure that has three *sides*] causes [*M* outputs *W* at t_3].

One argument that tries to justify the falsehood of (Side$_{t1}$) is the following: Unlike (Angle$_{t1}$), (Side$_{t1}$) cannot be generated from the causal facts given to us in the description of *M* processing *W* given in the dual-detector argument using causal generational principles such as (T) and (Conj). As a result, the truth of (Side$_{t1}$) would require some additional primitive causal fact to hold in the case of *M* processing *W*, which would be unparsimonious. Moreover, since any such additional primitive causal fact would only contingently hold, the possibility of such a fact holding can be removed by simply stipulating that no such additional primitive causal fact holds in the possible case of *M* processing *W* that we are concerned with. Hence, according to this argument, the truth of (Side$_{t1}$) can be ruled out either on parsimony grounds or by stipulation.

The problem with this argument for the falsity of (Side$_{t1}$) is that it begs the question against the modal theory. It does this because, if the modal theory is true, then, contra the above argument, (Side$_{t1}$) can be generated from the causal facts given to us in the description of the case of *M* processing *W* in the dual-detector argument and the generational principles (T) and (Conj) in the same way that (Angle$_{t1}$) can be so generated. This is because, if the modal theory is true, then [at t_1, *W* is a closed straight-sided figure that has three *sides*] is the conjunction of [*W* is a closed straight-sided figure at t_1] and [*W* has three angles at t_1], just as much as [at t_1, *W* is a closed straight-sided figure that has three *angles*] is. Hence, if the modal theory is true, then (Side$_{t1}$) can be derived from (Angle*$_{t1}$) and (Conj) in the same way that (Angle$_{t1}$) can.

An alternative way of trying to justify the falsehood of (Side$_{t1}$) appeals to (Conj*).[16]

> Conj*. If the conjunction of f_1 and f_2 partially causes f, then f_1 and f_2 collectively partially cause f.

We can give the same kind of argument from parsimony and contingency for the falsity of (Side*$_{t1}$) as was given above for the falsity of (Side$_{t1}$), with the

[16] (Conj) and (Conj*) are in the vicinity of two principles, (A) and (B), that Sober appeals to when defending the dual-detector argument.

(A) If two devices, "which are linked in series in the [machine], are sensitive just to properties *P* and *Q*, respectively, then the [machine] itself is sensitive to the conjunctive property *P*-and-*Q*." (Sober 1982, 186)

(B) If "two devices which are linked in series are such that the first is sensitive to *P* and the second is *not* sensitive to *R* (where $P \neq R$, and neither implies the other), then the [machine] is *not* sensitive to the conjunctive property *P*-and-*R*." (Sober 1982, 186)

As argued below in the case of (Conj*), (B) immediately conflicts with the modal theory and is hard to justify.

doi: 10.48106/dial.v76.i4.03

difference that this argument for the falsity of (Side*$_{t1}$), unlike the argument for the falsity of (Side$_{t1}$), does not beg the question against the modal theory.

> Side*$_{t1}$. [W is a closed straight-sided figure at t_1] and [W has three *sides* at t_1] collectively cause [M outputs W at t_3].

Indeed, plausibly both opponents and proponents of the modal theory should reject (Side*$_{t1}$). Given the falsity of (Side*$_{t1}$), however, the falsity of (Side$_{t1}$) follows from (Conj*).[17] If we are justified in endorsing (Conj*), then, we can use it to justify the falsehood of (Side$_{t1}$).

One problem with (Conj*) is that the principle directly conflicts with the modal theory. This is because, if the modal theory holds, then (Conj*) has the absurd consequence that, if f partially causes g, then any fact h that is necessitated by f also causes g. (This is because, according to the modal theory, if a fact f necessitates a fact h, then f is the conjunction of f and h.) If [Suzy throws a rock] causes [the window breaks], for example, then, if the modal theory holds, (Conj*) entails that [Suzy throws a rock or Suzy does not throw a rock] (which is necessitated by [Suzy throws a rock]) also causes [the window breaks], which is absurd. In light of this, one problem with (Conj*) is that, if it is accepted, then we don't need the dual-detector argument to refute the modal theory, since (Conj*) by itself achieves this task. If the dual-detector argument needs to rely on (Conj*) in order to be successful, then, the argument is superfluous.

A second (more serious) problem with (Conj*) is that it is not clear why we should believe it. A proponent of (Conj*) might attempt to justify the principle by arguing that, in ordinary language, sentences of the form (11) are equivalent to sentences of the form (12).

11. φ because ϕ and ψ.

12. φ because ϕ and because ψ.

Such a proponent might then argue that (on its relevant causal use) (11) is equivalent to (11*) and (12) is equivalent to (12*).

11*. [ϕ and φ] causes [φ].

[17] I am assuming that [at t_1, W is a closed straight-sided figure that has three sides] is the conjunction of [W is a closed straight-sided figure at t_1] and [W has three sides at t_1].

12*. [ϕ and φ] collectively cause [φ].

Assuming that these equivalences all hold, it follows that (11*) entails (12*), from which it follows that (Conj*) holds.

A problem with this attempted justification for (Conj*) is that (12) is plausibly ambiguous between a conjunctive reading and a non-conjunctive reading, just like (13) is.[18]

13. Jane wants to go swimming and go hiking.

(13) has a non-conjunctive reading on which the proposition Jane is described as desiring is the proposition that Jane goes swimming and hiking. On this reading, (13) is true iff (13n) is true.

13*n*. Jane wants to go (swimming and hiking).

(13) also has a conjunctive reading on which (13) is true iff (13c) is true.

13*c*. Jane wants to go swimming and Jane wants to go hiking.

(11) is plausibly similarly ambiguous between a non-conjunctive reading on which it is equivalent to (11n) and a conjunctive reading on which it is equivalent to (11c).

11*n*. φ because (ϕ and ψ).

11*c* (φ because ϕ). and (φ because ψ).

On its conjunctive reading, while (11) is equivalent to ((12) (on its causal use), there is no reason to think that (on its causal use) (11) is equivalent to (11*) (or at least no such reason has yet been provided).[19] On its non-conjunctive reading, on the other hand, there is no reason to think that (11) is equivalent to (12). As a result, appealing to natural language does not appear to help a proponent of the dual-detector argument justify (Conj*). In light of this, it

[18] Cf. (Marshall 2021, 8035).

[19] The claim that (12) is equivalent to (12*) can also be resisted, since it might be denied that "f causes h" and "g causes h" entails "f and g collectively cause h." For example, this inference might be thought to fail if f and g are individually complete causes of h that concern different times.

doi: 10.48106/dial.v76.i4.03

is not clear how (Conj*) might be justified.[20] As a result, it does not appear possible to justify the truth of (Angle$_{t1}$) by appealing to (Conj) while also justifying the falsity of (Side$_{t1}$).

I will discuss one further attempt to justify both the truth of (Angle$_{t1}$) and the falsity of (Side$_{t1}$). Instead of appealing to (Conj), this second attempt appeals to the popular counterfactual dependency thesis (Dep) stated above.[21]

> Dep. Suppose that *f* and g obtain, and that, had *f* failed to obtain, it would have been that g failed to obtain. Then, *f* causes g.

Assuming that (Dep) holds, we can derive (Angle$_{t1}$) as follows: In the case of *M* outputting *W*, had it not been that, at t_1, *W* was a closed straight-sided figure that had three *angles*, then either: i) *W* would not have been a closed straight-sided figure at t_1; or ii) *W* would not have had three angles at t_1, in which case *W* would also not have had three angles at t_2. If *W* had failed to be a closed straight-sided figure at t_1, *W* would not have been outputted by the closed straight-sided figure detector at t_2, and hence *W* would not have been outputted by *M* at t_3. On the other hand, if *W* had failed to have three angles at t_2, it would not have been outputted by the three-angle detector at t_3, and hence would also not have been outputted by *M* at t_3. Hence, had it not been that, at t_1, *W* was a closed straight-sided figure that had three *angles*, *M* would not have outputted *W* at t_3. It therefore follows from (Dep) that (Angle$_{t1}$) is true.

> Angle$_{t1}$. [*At* t$_1$, *W* is a closed straight-sided figure that has three *angles*] causes [*M* outputs *W* at t$_3$].

Assuming that (Dep) holds, then, a proponent of the dual-detector argument can use (Dep) to justify (Angle$_{t1}$). Unfortunately for proponents of the dual-detector argument, however, if (Dep) holds it can also be used to justify the truth of (Side$_{t1}$). To see why, note that, had it not been that, at t_1, *W* was a closed straight-sided figure that had three *sides*, then *W* would also either:

[20] Or at least, it is not clear how (Conj*) might be justified without begging the question against the modal theory. It might perhaps be possible to justify (Conj*) if we assume the structured theory and give a general account of how less fundamental facts get to have their causal features in terms of the causal features of more fundamental facts that involves principles like (Conj).

[21] Related principles we might try to appeal to in order to simultaneously justify the truth of (Angle) and the falsity of (Side) (which have similar problems to (Dep) are difference-making principles, such as those proposed by (Sartorio 2005) and (List and Menzies 2009).

i) not have been a closed straight-sided figure at t_1 or ii) not have had three angles at t_1, in which case it would not have had three angles at t_2. Hence, had it not been that, at t_1, W was a closed straight-sided figure having three *sides*, at least one of the detectors would not have outputted W, and so M would not have outputted W at t_3. Hence, it also follows from (Dep) that (Side$_{t1}$) is true. Hence, a proponent of the dual-detector argument cannot use (Dep) to justify the combination of (Angle$_{t1}$) being true and (Side$_{t1}$) being false. This second attempt at justifying the truth of (Angle$_{t1}$) and the falsehood of (Side$_{t1}$) therefore fails.

I have now discussed two attempts to justify the truth of (Angle$_{t1}$) and the falsity of (Side$_{t1}$), and I have argued that both of these attempts fail. As far as I can see, other attempts to do this are equally unsuccessful. If this is the case, then both the t_1-version and the t_2-version of the dual-detector argument fail.

3 Variants of the Dual-Detector Argument

In the face of the failure of the original version of Sober's dual-detector argument, it might be thought that the argument can be modified so that it evades the problems discussed in section 2. In particular, it might be thought that these problems can be evaded by replacing the necessarily equivalent facts expressed by ($1t_1$) and ($2t_1$) with some other necessarily equivalent facts and describing a machine that is causally sensitive to one of these facts but not the other.

$1t_1$. W is a closed straight-sided figure that has three *angles* at t_1.

$2t_1$. W is a closed straight-sided figure that has three *sides* at t_1.

As far as I can see, however, this cannot be done.

To illustrate the difficulty involved in successfully modifying the dual-detector argument in the above manner, I will briefly consider two attempts to do this that replace the facts expressed by ($1t_1$) and ($2t_1$) with the facts expressed by (14) and (15), where W^* is a circular wire and where the facts expressed by (14) and (15) are both necessarily equivalent to the fact that W^* is a circle.[22]

[22] This variant was suggested by a referee.

doi: 10.48106/dial.v76.i4.03

14. W^* is a closed (plane) figure all of whose points are equidistant from a point.

15. W^* is a closed (plane) figure of constant curvature.

For the first attempt, consider a machine M_1^* that, when given a closed (plane) figure as an input, scans that figure by having a distinct curvature detector for each point of the figure. Suppose that each of these detectors measures the curvature of their associated point in the figure and sends the result of this measurement in the form of a signal to the CPU of M_1^*. Further, suppose that, if all the signals the CPU receives are of the same value, then the fact that the signals it receives have the same value causes the figure to be outputted by M_1^*. Finally, suppose that the circular wire W^* is inputted into this machine M_1^*, is scanned by it, and is then outputted by it. It might then be claimed that, in this case, (Curv) is true while (Dist) is false, and that, due to Leibniz's law, this difference in truth-value entails that the modal theory is false.

Curv. [W^* is a closed figure with constant curvature] causes [M_1^* outputs W^*].

Dist. [W^* is a closed figure all of whose points are equidistant from a point] causes [M_1^* outputs W^*].

A problem with this first attempt at finding a successful variant of the dual-detector argument is that it is no more obvious that (Curv) holds than it is that (Angle$_{t1}$) holds in Sober's original case.

Angle$_{t1}$. [*At* t_1, W is a closed straight-sided figure that has three *angles*] causes [M outputs W at t_3].

Instead, using transitivity reasoning, what can be uncontroversially established in the variant case of machine M_1^* is a claim along the lines of (Curv*), just as what can be uncontroversially established using such reasoning in Sober's original case of machine M is Angle$^*_{t_1}$.

Curv*. [Point p_1 of W^* has curvature C], [point p_2 of W^* has curvature C]... collectively cause [M_1^* outputs W^*].

> ANGLE*$_{t1}$. [W is a closed straight-sided figure at t_1] and [W has three angles at t_1] collectively cause [M outputs W at t_3].

Moreover, an opponent of the modal theory who wishes to show that (Curv) and (Dist) differ in their truth-value faces the same challenges that a proponent of Sober's original version of the dual-detector argument faces in showing that (Angle$_{t1}$) and (Side$_{t1}$) differ in their truth-value. First, they need to resist an argument from causal exclusion that (Curv*) entails the falsehood of (Curv). And second, they need to find some way of justifying the truth of (Curv) while also justifying the falsehood of (Dist), a task that appears to be just as difficult as finding a way of justifying the truth of (Angle$_{t1}$) while also justifying the falsehood of (Side$_{t1}$). Hence, this first attempt at describing a machine that is differentially sensitive to the facts expressed by (14) and (15) results in a variant of the dual-detector argument that is no more successful than Sober's original argument.

For a second attempt to show that there could be a machine that is causally sensitive to one of the facts expressed by (14) and (15) but not the other, consider a machine M_2^* that contains an extendable straight rod that rotates around one of its endpoints. When given a closed figure as input, M_2^* works by placing this rod inside the inputted closed figure, fixing the location of one of the rod's endpoints, extending the length of the rod until its other endpoint touches the inputted figure, and then rotating the rod around its fixed endpoint while keeping the length of the rod fixed. If the rod does a full rotation without moving the inputted figure or losing touch with it, then the fact that it does this causes M_2^* to output the figure. Suppose now that the circular wire W^* is inputted into M_2^* and that the rod of M_2^* is placed inside of W^* and does a full rotation meeting the above conditions, so that W^* gets outputted by M_2^*. It might then be claimed that, in this case, (Dist) is true and (Curv) is false, and hence that the modal theory is false.

The problem with this second variant of Sober's version of the dual-detector argument is that, if W^* is a circle that is inputted into and then outputted by M_2^*, then there is no reason to think that (Curv) and (Dist) differ in their truth-value. In particular, if W^* is so inputted and outputted, it is equally plausible to say that the machine measures the curvature of the points of W^* as it is to say that it measures the equidistance of those points from a common point. After all, the rod would fail to do its full rotation (while touching but not moving W^*) if the points of W^* didn't have constant curvature, just as it would fail to do this if the points of W^* weren't equally distant from some common point. There is therefore no grounds for thinking that W^* being outputted by M_2^* is due to one of these facts rather than the other. Hence, M_2^* also fails to be a demonstrable case of a machine that is causally sensitive to one of the facts expressed by (14) and (15) and not the other.

Other variations of Sober's original version of the dual-detector argument face similar problems to those described above. Indeed, the above two attempts to construct a successful variant of Sober's original version of the argument arguably illustrate a dilemma facing any such attempt. This dilemma is the following: Suppose we have a machine whose output is intended to be caused by the fact f_1 and not by the necessarily equivalent fact f_2. Then the machine will either contain multiple detectors that differ in what aspects of the input they measure (as in the cases of M and M_1^*), or the machine will only contain detectors (or a single detector) that don't so differ (as in the case of M_2^*). If the machine contains multiple detectors that differ in what aspects of the input they measure, then the argument against the modal theory based on this machine will arguably face the same challenges facing Sober's original argument and the first variant of it discussed above. In particular, the argument will need to resist an argument from causal exclusion and will face the same difficulties in justifying the claim that the input being outputted is caused by f_1 and not by f_2 that Sober's original dual-detector argument faces in justifying the truth of (Angle$_{t1}$) and the falsity of (Side$_{t1}$). On the other hand, if the machine contains only a single detector (or multiple detectors that don't differ in what aspects of the input they measure), then it will arguably fail to be even initially plausible that f_1 and f_2 differ in whether they cause the input to be outputted just as there is no even initial plausibility for thinking that the facts expressed by (14) and (15) differ in whether they cause W^* to be outputted by machine M_2^*. Hence, whether or not we have a machine that contains detectors that differ in what aspects they measure, the argument against the modal theory based on this machine will arguably fail. In light of this, it is reasonable to conclude that, not only does Sober's original version of the dual-detector argument fail, but it is not possible to modify the argument so that it is successful. If this is correct, then all variants of the dual-detector argument fail and some other kind of argument will be needed if we are to have reason to reject the modal theory of facts and properties.*

* Research in this paper was supported by an Early Career Scheme grant from the Research Grants Council of Hong Kong SAR,China (LU23607616). Thanks to Andrew Brenner, Daniel Waxman and three anonymous referees for their valuable comments on this paper.

Dan Marshall
0000-0002-5763-3875
Lingnan University
Danmarshall@ln.edu.hk

References

Achinstein, Peter. 1974. “The Identity of Properties.” *American Philosophical Quarterly* 11(4): 257–275.

Armstrong, David M. 1997. *A World of States of Affairs*. Cambridge: Cambridge University Press, doi:10.1017/cbo9780511583308.

Audi, Paul. 2016. “Property Identity.” *Philosophy Compass* 11(12): 829–840, doi:10.1111/phc3.12380.

Bjerring, Jens Christian and Schwarz, Wolfgang. 2017. “Granularity Problems.” *The Philosophical Quarterly* 67(266): 22–37, doi:10.1093/pq/pqw028.

Braun, David. 1998. “Understanding Belief Reports.” *The Philosophical Review* 107(4): 555–595, doi:10.2307/2998375.

Dorr, Cian. 2016. “To Be *F* Is To Be *G*.” in *Philosophical Perspectives 30: Metaphysics*, edited by John Hawthorne, pp. 39–134. Hoboken, New Jersey: John Wiley; Sons, Inc., doi:10.1111/phpe.12079.

Enç, Berent. 1982. “Intentional States of Mechanical Devices.” *Mind* 91(362): 161–182, doi:10.1093/mind/xci.362.161.

Goodman, Jeremy. 2017. “Reality is Not Structured.” *Analysis* 77(1): 43–53, doi:10.1093/analys/anw002.

Hall, Ned. 2000. “Causation and the Price of Transitivity.” *The Journal of Philosophy* 97(4): 198–222, doi:10.2307/2678390.

Jackson, Frank. 1998. *From Metaphysics to Ethics: A Defense of Conceptual Analysis*. Oxford: Oxford University Press, doi:10.1093/0198250614.001.0001.

Kim, Jaegwon. 1973. “Causes and Counterfactuals.” *The Journal of Philosophy* 70(17): 570–572, doi:10.2307/2025312.

—. 2005. *Physicalism, or Something Near Enough*. Princeton Monographs in Philosophy. Princeton, New Jersey: Princeton University Press, doi:10.1515/9781400840847.

Kment, Boris. 2022. “Russell-Myhill and Grounding.” *Analysis* 82(1): 49–60, doi:10.1093/analys/anab028.

Kvart, Igal. 1991. “Transitivity and Preemption of Causal Relevance.” *Philosophical Studies* 64(2): 125–160, doi:10.1007/bf00404826.

Lewis, David. 1973. “Causation.” *The Journal of Philosophy* 70(17): 556–567. Reprinted, with a postscript (Lewis 1986d), in Lewis (1986b, 159–213), doi:10.2307/2025310.

—. 1986a. *On the Plurality of Worlds*. Oxford: Basil Blackwell Publishers.

—. 1986b. *Philosophical Papers, Volume 2*. Oxford: Oxford University Press, doi:10.1093/0195036468.001.0001.

—. 1986c. "Events." in *Philosophical Papers, Volume 2*, pp. 242–169. Oxford: Oxford University Press, doi:10.1093/0195036468.001.0001.

—. 1986d. "Postscript to Lewis (1973)." in *Philosophical Papers, Volume 2*, pp. 172–213. Oxford: Oxford University Press, doi:10.1093/0195036468.001.0001.

List, Christian and Menzies, Peter. 2009. "Nonreductive Physicalism and the Limits of the Exclusion Principle." *The Journal of Philosophy* 106(9): 475–502, doi:10.5840/jphil2009106936.

Loewer, Barry C. 2007. "Mental Causation, or Something Near Enough." in *Contemporary Debates in Philosophy of Mind*, edited by Brian P. McLaughlin and Jonathan Cohen, pp. 243–264. Contemporary Debates in Philosophy n. 8. Malden, Massachusetts: Basil Blackwell Publishers, doi:10.1093/acprof:oso/9780199580781.003.0015.

Marshall, Dan. 2021. "Causation and Fact Granularity." *Synthese* 199(3-4): 8029–8045, doi:10.1007/s11229-021-03151-2.

McDermott, Michael. 1995. "Redundant Causation." *The British Journal for the Philosophy of Science* 46(4): 523–544, doi:10.1093/bjps/46.4.523.

McKay, Thomas J. and Nelson, Michael. 2010. "Propositional Attitude Reports." in *The Stanford Encyclopedia of Philosophy*. Stanford, California: The Metaphysics Research Lab, Center for the Study of Language; Information. Revision, October 5, 2010, of the version of February 16, 2000, https://plato.stanford.edu/archives/win2010/entries/prop-attitude-reports/.

Miller, Alexander. 1995. "Objectivity Disfigured: Mark Johnston's Missing-Explanation Argument." *Philosophy and Phenomenological Research* 55(4): 857–868, doi:10.2307/2108336.

Molnar, George. 2003. *Powers: A Study in Metaphysics*. Oxford: Oxford University Press. Edited by Stephen Mumford, doi:10.1093/acprof:oso/9780199204175.001.0001.

Moore, Dwayne. 2018. "Mind and the Causal Exclusion Problem." in *Internet Encyclopedia of Philosophy*. University of Tennessee at Martin, https://iep.utm.edu/mind-and-the-causal-exclusion-problem/.

Partee, Barbara Hall. 1989. "Possible Worlds in Model-Theoretic Semantics: A Linguistic Perspective." in *Possible Worlds in Humanities, Arts and Sciences*, edited by Sture Allén, pp. 93–123. Research in Text Theory n. 14. Berlin: Walter de Gruyter. Proceedings of Nobel Symposium 65, doi:10.1515/9783110866858.93.

Perry, John R. 1989. "Possible Worlds and Subject Matter. Discussion of Partee (1989)." in *Possible Worlds in Humanities, Arts and Sciences*, edited by Sture Allén, pp. 124–137. Research in Text Theory n. 14. Berlin: Walter de Gruyter.

Reprinted with a postscript in Perry (1993, 173–192) and in Perry (2000, 145–160), doi:10.1093/0so/9780195049992.003.0008.

—. 1993. *The Problem of the Essential Indexical and Other Essays*. New York: Oxford University Press, doi:10.1093/0so/9780195049992.001.0001.

—. 2000. *The Problem of the Essential Indexical and Other Essays*. Stanford, California: CSLI Publications. Enlarged edition of Perry (1993).

Sartorio, Carolina. 2005. "Causes as Difference-Makers." *Philosophical Studies* 123(1–2): 71–96, doi:10.1007/s11098-004-5217-y.

Sober, Elliott R. 1982. "Why Logically Equivalent Predicates May Pick Out Different Properties." *American Philosophical Quarterly* 19(2): 183–189.

Stalnaker, Robert C. 1984. *Inquiry*. Cambridge, Massachusetts: The MIT Press.

Walsh, Sean Drysdale. 2016. "Predicativity, the Russell-Myhill Paradox, and Church's Intensional Logic." *The Journal of Philosophical Logic* 45(3): 277–326, doi:10.1007/s10992-015-9375-5.

Yu, Andy Demfree. 2017. "A Modal Account of Propositions." *Dialectica* 71(4): 463–488, doi:10.1111/1746-8361.12193.

Lewisian Worlds and Buridanian *Possibilia*

Boaz Faraday Schuman

Many things can be other than they are. Many other things cannot. We talk about such things all the time. But what is this talk about? One answer, presently dominant in analytical philosophy, is that we are speaking of possible worlds: if something can be other than it is, then it actually is that way in some (other) world. If something cannot be otherwise, it is not otherwise in any world whatsoever. But what are these worlds? David Lewis famously claims that every world exists, just like ours does. In contrast, the medieval thinker John Buridan understands modal logic in terms of objects and causal powers: if something can be other than it is, then there is a causal power that can make it that way. If it cannot, then no causal power—not even God—can make it otherwise. As we'll see, (i) the Lewisian plurality is not possible on Buridan's account, and accordingly (ii) a basic tenet of classical theism is untenable on Lewis's metaphysics. In short, either the Lewisian plurality is incoherent, or a core monotheistic tenet is impossible.

Modal sentences deal with things that can or must or cannot be. For example, we say that a triangle *can* be drawn, *must* be three-sided, and *cannot* be round. What makes a modal sentence modal? Short answer: its inclusion of a modal term like *can* (*possibly*), *must* (*necessarily*), and so forth. Such terms register that a claim is being qualified in such a way that the conditions of its truth are not limited to the way things actually are. But what is this modal talk about? Over the past two and a half millennia, answers have varied. Relatively recently, we have come to think of modes in terms of quantification over worlds: what is possible is true in at least one world, and what is necessary is true in all. Call this the *worlds-reading* (WR) of modal sentences. David Lewis (1941–2001) famously understands WR ontologically: these worlds really exist as spatiotemporal isolates, and are every bit as real as our own.

Contrast WR with a much older—and for a long time prominent—understanding of what modes are: terms whose operation on sentences expands (or *ampliates*) the extension of their terms, so that the terms range over possible objects, including non-existent ones. The modal properties of these objects are grounded in the causal powers of existing things: a triangle can be drawn because you or I can draw one; it is necessarily three-sided because there is no causal power (not even God) capable of making a triangle to be otherwise—at least, not without depriving it of its triangularity. Call this the objects reading (OR) of modal sentences. This is the view of John Buridan (c.1300–1361).[1] A careful examination of these views reveals that (i) they are incompatible, so that the Lewisian plurality is not a possible object or collection of objects; and accordingly that (ii) the worlds-reading, at least in its Lewisian form, is incompatible with a basic tenet of classical theism.

Why compare Buridan and Lewis? I have three reasons. First, Lewisian modal realism is well-known, and therefore provides a convenient off-the-shelf foil for Buridan's modal ontology. Second, Lewis has clear ontological commitments, and so he is easy to pin down. Compare the ontologically agnostic Kripkean modal semantics and syntax: you and I may have very different views on what worlds are, but nevertheless agree on a Kripkean reading of the claims of WR. So the Kripkean account does not provide a clear and illuminating contrast for Buridan's modal ontology, the way Lewis's approach does. Third, contrasting Lewis and Buridan illuminates latent aspects of both. It gives us an insight into Lewis, hitherto unrecognised in the literature; and it reveals Buridan's own views on the limitations on divine power—limitations he does not explicitly discuss at length. After all, placing restrictions on God's power would have been a hazardous thing to do at the fourteenth-century University of Paris.[2] All the more so for an Arts Master who, as he explicitly acknowledges, is not qualified to teach theology.[3] All the same, we can tease out the consequences of the views Buridan does express. And there is more here than meets the eye.

[1] For a discussion of earlier debates about causal powers in the twelfth and thirteenth centuries, see Peter King (2021).

[2] In particular the infamous Condemnations of 1277 insisted on the boundlessness of divine power. For a discussion, see Grant (1979), and more recently Thijssen (2018).

[3] That Buridan never advanced beyond the post of arts master, and so—in spite of his evident brilliance—never taught at the higher and more prestigious Faculty of Theology, is remarkable. In modern terms, this would be a bit like deciding to remain an assistant professor for life, even when promotion was available. For a discussion, see Jack Zupko (2003, xi–xii).

Let's begin with WR, which is relatively familiar, and has two important shortcomings that point to two strengths of OR.

1 Possible Worlds

Nowadays, we tend to think of modality in quantificational terms: a modal is a sentence with a modal operator like "$\Box$" or "$\Diamond$," for necessity and possibility, respectively. Such operators quantify across possible worlds. On these lights, $\Box\varphi$ just says that φ holds in all possible worlds, and $\Diamond\varphi$ says that φ holds in at least one. The parallel, then, is with the ordinary first-order quantifiers: ($\Box$-like) "$\forall$," and ($\Diamond$-like) "$\exists$."[4]

There is much to be said for WR, but here I will limit myself to two points. First, it's versatile: we can use the apparatus of worlds to construct a wide variety of systems of alethic modal logic—that is, modal systems dealing with necessary truths, possible truths, and so on. We can characterise an astonishing number of systems in this way, and haggle about which one is best (or best for what). We can also characterise non-alethic systems to model knowledge and belief (epistemic logic), past, present, and future time (tense logic), and morality (deontic logic). WR, then, is extremely fruitful.[5]

Second, the WR is precise: can we give clear quantificational definitions of terms like *necessarily* and *possibly*, which might otherwise seem qualitative and murky. And, using Kripke's apparatus of frames, we can characterise our systems with mathematical precision. But beyond all this, we might wonder: what are these worlds, anyway?

1.1 *Lewisian Worlds*

David Lewis's answer to this question is famous and bold: all possible worlds exist, and they are just as real as ours. As he tells us (1986, 2):

[4] One need not, however, be committed to a semantics of possible worlds in order to think of modal terms quantificationally: already in 1924, well before the possible-worlds innovations of Kripke, Otto Jesperson pointed out that "necessity means that *all* possibilities are comprised, just as impossibility means the exclusion of all possibilities" (1924, emphasis original, 325).

[5] As Graham Priest (2016, 2653) puts it, "the clarity of the mathematics involved, and their usefulness in an analysis of many things other than modality—such as conditionals, meaning, knowledge and belief—meant that they [i.e., possible worlds] soon became part of the intellectual landscape."

doi: 10.48106/dial.v76.i4.04

> The other worlds are of a kind with this world of ours. To be sure, there are differences of kind between things that are parts of different worlds […] but […] the difference between this and the other worlds is not a categorical difference. Nor does this world differ from the others in its manner of existing.

According to Lewis, there are many worlds—as many, in fact, as there are ways things can be. This ontological account of WR prompts two questions: how are these worlds externally distinct from each other, and how are they internally unified? Answers to both questions turn on spatiotemporal relations. To the former, Lewis tells us (1986, 3):

> There are no spatiotemporal relations at all between things that belong to different worlds. Nor does anything that happens at one world cause anything to happen at another. Nor do they overlap; they have no parts in common.

Lewis frequently treats causation as the paradigmatic spatiotemporal relation. Since the worlds have no spatiotemporal relations to one another, there can be no causal interactions between them. They are therefore not like planets that are too far removed to interact with each other. They are, rather, spatiotemporal isolates. Call this Lewis's *isolation doctrine*.

Importantly, Lewis does not say that different worlds *cannot* interact, as if blocked from doing so. Rather, they just *do* not: the notion of interaction between different worlds makes no sense within his theory. This requirement has a stipulative flavour—and, indeed, it is precisely that: a stipulation. This point is important, and we will return to it in section 3.

In like manner, Lewis accounts for the unity of worlds in terms of spatiotemporal relations (1986, 71):

> If two things are spatiotemporally related, they are worldmates […] things are worldmates iff they are spatiotemporally related. A world is unified, then, by the spatiotemporal interrelation of its parts.

Again, this is presented in a stipulative way, though it is a corollary of the doctrine of isolation: worlds are spatiotemporally isolated, and therefore any spatiotemporally related things belong, *eo ipso*, to the same world. Here, whether or not causal interaction *actually* occurs is less important than immediately above: there does not need to be any obvious causal relation between two things for them to belong to the same world. A long-dead star too distant from Earth to interact with it nevertheless has spatiotemporal relations to us: it is some distance away in time and space, and it came into being at some time relative to us. It is, therefore, our worldmate.

The foregoing considerations can be distilled into a precise account of Lewisian worlds or *possibilia*, to wit:

> POSSIBILIA$_\mathrm{L}$. A world *w* is an isolated unity of spatiotemporally interrelated parts. If *x* and *y* have any spatiotemporal relations, they are members of the same world.

The spatiotemporal relation is, in its most general sense, Euclidean. Let R be the spatiotemporal relation, so that R*xy* says that *x* is spatiotemporally (though not necessarily causally) related to *y*. Then, by POSSIBILIA$_\mathrm{L}$,

$$\forall xyz(\mathrm{R}xy \wedge \mathrm{R}xz \rightarrow \mathrm{R}yz)$$

For clarity, we can also represent this diagrammatically, as follows:

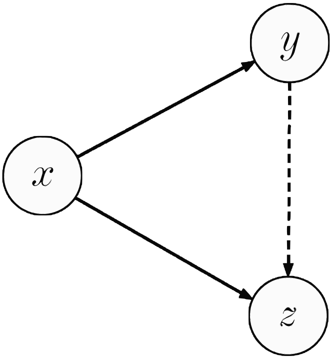

Figure 1: Euclidean R

Here, R is represented by arrows; if the relation represented by the solid arrows between *x* and *y*, *x* and *z* hold, then the relation represented by the dotted arrow between *y* and *z* also holds.

This fact makes the case that the Lewisian plurality is impossible (set out in section 3) much easier to make, so let's linger on it for a moment. Let R*xy*

doi: 10.48106/dial.v76.i4.04

and Rxz. It follows that Ryz. If it didn't, then x would be worldmates with two objects that are not themselves worldmates with each other. So there would be partial but incomplete overlap among at least two worlds. And this goes against both *possibilia*$_{\mathrm{L}}$, and against commonsense thinking about spatiotemporal relations: if, for example, x is some spatial or temporal distance from both y and z, then there must be some distance, however great, between y and z themselves. Therefore, the spatiotemporal relation R is Euclidean.

At the beginning of this section, I noted two significant advantages to the WR of ordinary modal language: WR is precise, and fruitful. Before we turn to the possible objects of Buridan, it's worth asking whether WR has any drawbacks. For present purposes, I want to highlight two: WR does not represent what is going on in ordinary modal language, and taken on its own, it is uninformative about what grounds the modal properties of things.

To begin with the latter: the extensional account furnished by WR does not capture the ordinary notion of necessity *for* or *as*. For example, triangles are necessarily three-sided; three-sidedness is necessary *for* triangle-hood. Whereas you can paint a triangular object blue without removing its triangularity, you cannot, say, rearrange its parts in such a way that it gains (or loses) a side, and yet remains a triangle. This fact is not directly expressible on WR; all it can tell us about this (or any other) necessary claim is that it is true in every world. Fair enough, but such claims do not account for the inseparability of three-sidedness and triangularity.

Probably for this reason, most ordinary modal talk is not about worlds at all, but rather about things, and the ways they can be in *this* world. Scott Soames gives some remarks that support this point in his discussion of reference to non-existent objects (2010, 128):

> Although this is controversial, the idea that we can refer to, and quantify over, only things that exist is, I believe, an unfounded philosophical prejudice at variance with our ordinary thought and talk. For instance, imagine that I have all the materials to build a doghouse, plus a plan specifying every detail of the design and construction, including how each of the materials will be used. From studying the plan and materials, I know exactly which structure I intend to create. Having identified it uniquely, I can refer to it, predicate properties of it, and even name it.

Soames's dog house is a possible, non-existent object. What makes it possible is what *he* can do with materials and plans in *this* world. A lot of our day-to-day modal talk is like this: when, for example, someone says they can paint their house green, they are talking about *themselves*, and what they can do with *their house*—not about their counterpart, in a relevantly similar world in which their counterpart's house is green.

Thus for all its versatility and precision, WR does not provide a full and accurate report of what is going on in ordinary modal language. Such language, judging by Soames's example, is about possible things, at least some of which do not exist, whose modal properties are grounded in existing causal powers. I have called this the objects reading (OR) of modal language; it is the approach taken by John Buridan. It turns out that objects like Soames's doghouse are precisely what Buridan has in mind in his analysis of *possibilia*.

2 Possible Objects

In the WR of modal language, modes operate on whole sentences, quantifying over possible worlds. In contrast, Buridan's modal logic is not propositional but *terminist*; he thinks of modes as acting on sentences' terms.[6] Hence in his treatment of modal semantics in *Tractatus de Consequentiis* (2.4), he tells us that:

> A sentence (*propositio*) [...] about possibility has a subject term that is ampliated (*ampliatum*) by the modal term that follows it, so that it stands (*ad supponendum*) not only for those things which exist, but also for those things which *can* exist even though they do not. Hence in this way it is true that air can come from water, although this is not true of any air that presently exists.[7]

[6] While Buridan's *possibilia* have not received much attention, a good deal has been said already about Buridan's modal syntax and semantics. To date, the most thorough treatment of his syntax is chapter 9 of Paul Thom's (Thom 2003). And, following the concluding suggestions in G.E. Hughes' (Hughes 1989), Catarina Dutilh Novaes (Dutilh-Novaes 2007, 79–114) and Spencer Johnston (2015, 2–12; 2017, 41–43) have given detailed analyses of Buridan's logic in terms of possible worlds. Gyula Klima, too, has remarked in his monumental translation of Buridan's *Summulae de Dialectica* that Buridan's modal semantics contains "effectively the gist of the idea of modern possible-worlds semantics" (2004, 82, n.123).

[7] "Propositio [...] de possibili habet subiectum ampliatum per modum sequentem ipsum ad supponendum non solum pro his quae sunt sed etiam pro his quae possum esse quamvis non sint. Unde sic est verum quod aer potest fieri ex aqua, licet hoc non sit verum de aliquo aere qui est." (Unless otherwise stated, all translations here are mine.) Note that Buridan is here talking about *divided* (roughly, *de re*) modals; he deals with *composite* (roughly, *de dicto*) modals elsewhere. Now, immediately below this passage, Buridan tells us that a modal sentence "B is

doi: 10.48106/dial.v76.i4.04

Air from water is, as Paul Thom (Thom 2003, 170) has observed, a simple account of boiling. The water in this pot could boil; but since it is not boiling, it is not true of any actual air that it came from this water. Hence this water is possible—but not actual—air. Elsewhere, Buridan gives the example of vinegar that could be produced from this wine, but will not, simply because I am going to drink the wine first (*de Caelo*, 1.25).[8] These are the non-existent possible objects—or *possibilia*—to which the modal terms expand—or ampliate—the terms of a sentence.[9]

What are these non-existent *possibilia*?[10] Buridan deals with *possibilia* obliquely in his logic and metaphysics, and so we will have to reconstruct his view from these discussions. Here, I present three key passages: one dealing with necessity, one with impossibility, and the last with possibility. Approaching Buridan's account of the *possibilia* from these three angles will allow us to build up a consistent and robust picture of his views on what they are.

2.1 *Necessity in the* Prior Analytics

If S is necessarily P, then (by modal duality) it is not possible for S not to be P. Yet this analysis faces a problem. As Buridan asks in his *Quaestiones super libros "Analyticorum Priorum"* (*QAPr* 1.25), what is the modal status of the following sentence?

(1) Humans are animals.

Is (1) necessarily true? In *Prior Analytics* 1.9 (310^{a}31), Aristotle clearly thinks so. And indeed, (1) serves as a stock example of a necessary truth in medieval

possibly A" is equivalent to "What is or can be B can be A." An anonymous reviewer for this journal has remarked on the connection with Williamson's (2013, sec. 1.3) distinction between two readings of "possible stick:" the *predicative* reading ("x is a stick and x could have existed"), and the *attributive* reading ("x could have been a stick"). Buridan's own account looks, *prima facie*, more like the predicative reading; but perhaps the two are not equivalent. At any rate, this question could form the basis of a stand-alone paper.

[8] Cf. Aristotle's cloak in *Peri Hermeneias* 9, which can be cut up, but may also simply wear out first (19^{a}12–16).

[9] For an overview of Buridan's semantic doctrine of modal ampliation, and a case for it as one of his most significant contributions to the development of logic, see Zupko (Zupko 2003, 67–70), & (Zupko 2018, sec. 4).

[10] An anonymous reviewer for this journal has remarked that the common use of the term *possibilia* is for non-existent (possible) things, and does not extend to existing things as well. This is how I use it here, though it should be borne in mind that all *actualia* are, for Buridan, *possibilia* as well. After all, everything actual is possible.

logic.[11] Yet (1) is falsifiable, since God could annihilate all human beings. As Buridan tells us (*QAPr* 1.25, arg. 3):

> If it were supposed that (1) were not necessary, it would be because God is capable of annihilating every human being. And in such a case, no human would exist, and so no human would be an animal.[12]

For Buridan, all affirmative sentences, including universals, have existential import, in contrast with negative sentences (both universal and particular), which do not. Thus Buridan would reject the reading of (1) given by classical FOL ($\forall x[\text{Human}(x) \rightarrow \text{Animal}(x)]$), which is capable of vacuous truth.[13] Since there is no vacuous truth for affirmatives, (1) can be rendered false by the annihilation of its subject matter. Therefore, since (1) is falsifiable, it expresses a contingent truth.

Nor is this sort of contingency limited to sentences which, like (1), are taken from the natural sciences. It is also a problem for geometry:

> If this were so, then no claim of geometry would be necessary either, since God can just as well annihilate all magnitudes as all human beings. And then it would follow that geometry would not be a science, which everyone would regard as false and unsuitable. (*QAPr* 1.25, arg. 3).[14]

God can annihilate everything with magnitude, and therefore magnitude itself. If God were to do that, then all the affirmative claims of geometry would be false, since the things they deal with would not exist. This is a consequence of Buridan's anti-realism, which extends even to the objects of mathematics and geometry: if it so happened that there were no triangular arrangements of matter, then there would be no triangles (though it would still be possible

[11] Along with "God exists" and "No human is a donkey." Modern logical textbooks prefer mathematically-flavoured examples like "The set of primes is denumerable" and "$a = a$." The conventionalised role of these stock examples is clear.

[12] "Item, si poneretur quod non esset necessaria, hoc esset pro tanto quia deus posset annihilare omnem hominem; ideo nullus homo esset, et sic nullis homo esset animal."

[13] I have discussed this aspect of Buridan's logic, in connection with the traditional Square of Opposition, in Schuman (2022), 205–208.

[14] "Si hoc obstaret, nulla propositio geometrica esset necessaria, cum deus ita possit annihilare omnes magnitudines, sicut omnes homines. Et tunc ultra sequeretur quod geometria non esset scientia, quod reputatur ab omnibus falsum et inconveniens."

to think and talk about them, like the roses of yesteryear). The same holds for all other geometric and mathematical objects.

Worse, even if God never gets that destructive, a crisis remains: the mere fact that geometric claims *could* be falsified by an act of divine will entails that these claims are contingent. If the truth of any claim is contingent, so is its subject matter. Since the subject matter of any science (*scientia*) must be necessary, it follows that even geometry is not a science. We can expect the other sciences—with the obvious exception of theology—to fare no better, given that God could annihilate their subject matter, too. So can there be any science (apart from theology) at all?

Buridan's answer is *yes*: the claims of geometry (and of the other sciences) are necessary, but their necessity is attenuated: they are not necessarily true *simpliciter*. Rather, they are true "so long as" or "just when" (*de quando*) the things their subject and predicate terms stand for exist. Assuming no annihilation of their subject matter occurs, they will remain true—indeed, *necessarily* true:

> Necessity "just when" (*de quando*) comes about from the fact that, whenever the subject and predicate terms do stand for anything, they stand for the same thing (I am here speaking of affirmative sentences). And in this way I say that the following are necessary: "Humans are animals," or also "Horses are animals." Indeed, even "A rose is a flower" is necessary in this way, even if there are no roses now. And although there is not a lunar eclipse happening right now, still the following is necessary: "An eclipse is an obstruction of the moon by the sun." (*QAPr* 1.25, co).[15]

So a sentence like (1) is necessarily true, assuming the existence of the things it deals with, namely humans. Likewise, the claims of astronomy are true even when the events they describe are not presently occurring, since any time they *do* occur, the sentences are true. Thus, according to the account set out by Buridan in *QAPr* 1.25, a sentence like (1) can only be falsified by

[15] "Necessitas de quando ex hoc provenit quod oportet subiectum et praedicatum quandocumque supponunt pro aliquo supponere pro eodem; et hoc dico in affirmativis. Et sic dico quod haec est necessaria 'homo est animal,' vel etiam 'equus est animal.' Immo etiam haec est necessaria 'rosa est flos,' licet modo nulla sit rosa. Et quamvis non sit eclipsis lunae, tamen haec est necessaria 'eclipsis lunae est defectus luminis a sole.' Sed isto modo haec non est necessaria 'uacuum est locus' si ponamus cum Aristotele quod impossibile est uacuum esse."

the *annihilation* of the things it deals with. There is no way to falsify (1) that leaves humans intact. So whenever humans exist, (1) is true.

Thus the contrast between necessity and contingency in terms of modality simply construed (*simpliciter*) is the contrast between unfalsifiability and falsifiability. The contrast between necessity and contingency in terms of *de quando* modality is the contrast between falsifiability only by annihilation (*de quando* necessity) and falsifiability by alteration (*de quando* contingency). That humans are animals is *de quando* necessary, because it can only be rendered false by the removal of its subject matter. On the other hand, the fact that some humans are bearded is *de quando* contingent, since shaving them alters the fact, but leaves the subjects essentially intact.

From these observations, we can give the following Buridanian definition of necessity:

> BURIDANIAN NECESSITY. S is necessarily P just in case S can only be made to be not-P by annihilating S.

This provides a good starting point for Buridanian modality; however there are crucial ambiguities that must be sorted out, if the above definition is to be consistent with the others we will look at below. Its adoption here is, therefore, tentative.

2.2 *Impossibility in the* Peri Hermeneias

In *Peri Hermeneias* 2 (16^a19), Aristotle tells us that nouns (ὀνόματα; Aristoteles Latinus: *nomina*) have signification. But Buridan asks, what about nouns like *chimera*, which do not signify anything at all?

> We ask: does every noun (*nomen*) signify something?
>
> Objection: it does not, because the term *chimera* signifies nothing apart from a chimera. And yet a chimera is nothing. Therefore, it signifies nothing whatsoever.[16]

A chimera not only does not exist, like the roses of yesteryear; it is, in fact, impossible. Buridan makes this point several times: the chimera is made of

[16] “Queritur utrum omne nomen significat aliquid. Arguitur quod non, quia iste terminus ‘chimaera’ nihil significat aliud a chimaera. Et tamen nihil est chimaera. Ergo nihil omnino significat” (*Peri. Herm.* 1.2, arg. 1).

doi: 10.48106/dial.v76.i4.04

incompossible parts.[17] In this respect, we may take it to be just like Schopenhauer's wooden iron or Frege's square circle (Schopenhauer (1819)), vol.1, §53; Frege (1884), §74). Because the chimera cannot exist, it cannot be signified. And this seems to present a semantic counterexample to the *Peri Hermeneias* definition of nouns, even though syntactically, *chimera* functions like any other noun.

Buridan's solution here is to treat *chimera* as equivalent with the phrase "animal made up of parts that cannot be combined," and to note that, although this whole phrase does not signify anything, it has significative parts (namely *animal* and *part*). The details of this solution need not detain us here. What is significant for our purposes is the role of the chimera as an impossible object, whose impossibility is a function of its putative combination of incompossible parts. We can use such *impossibilia* for our next definition:

> Buridanian Impossibility. S is not possibly P if S and P cannot be combined.

This relatively straightforward definition will figure prominently in an important exegetical problem in section 2.4.

2.3 *Potency in the* Metaphysics

Buridan's most detailed discussion of modal properties of *possibilia* is in his *Questions on the "Metaphysics" of Aristotle* (*QM*) 9.5. There, Buridan asks whether everything that something *will* do can be said to be what it is *able* to do. If so, we get some strange results, as Buridan points out:

> A horse can come from wool. For earth comes from wool [by decomposition], and grass come from the earth, and from those grass which perhaps a horse will eat there can come horse semen, and, at length, another horse. And so even a horse can come from wool. And the same holds for all other modes of transmutation.[18]

[17] "Chimaera est animal compositum ex membris ex quibus impossibile est aliquod animal componi." (*De Demonstrationibus* 8.2.3). For a lively discussion of the role of the chimaera in the history of philosophy, see Ebbesen (1986).

[18] "Similiter ex eadem lana potest fieri equus, quia ex lana fiet terra, de inde herba, et ex illa herba forte quam equus comedet poterit fieri sperma equi et tandem equus. Et ita etiam ex lana potest fieri equus. Et sic de omnibus aliis modis transmutandi." (*QM* IX, 5, fol. 58rb). Among the other

Here the problem is whether or not the relation between S and P expressed by "S is possibly P" is transitive: if S can be P, and P can be Q, does it follow that S can be Q?

No, says Buridan: when we say that S can be P, we are generally speaking in terms of a *proximate* potency, rather than a remote one: S is proximately possibly P if S can become P in no more than one transmutation. In this way, wool is possibly earth, because it can become earth in one transmutation (i.e., decay); similarly, earth can become grass, and so on. Any other potencies that require multiple transmutations are remote—as is, for instance, the potency of wool to become a horse. Hence Buridan tells us that:

> Aristotle concludes the opposite. For he asks, when should something be said to be in potency, and when should it not? And he says that something should not be said to be in potency with respect to some form, except when only one transmutation is required, by which that form may be imparted on it.[19]

So although remote potencies can be discussed transitively, proximate potencies cannot. If the two are conflated, as in the wool-into-horse example, then, according to Buridan, the result is an equivocation.[20] Thus, although wool can decompose into earth, grass can grow from earth, and so forth, it does not follow that wool can become grass—much less a horse. Hence in speaking of possible horses, we are not speaking of all the things that, through multiple transmutations, could become a horse. If we were, then everything would be a possible horse, since, as Buridan observes, "anything can come from anything—albeit through several transmutations."[21]

So much for *possibilia* arising from natural causes, like possible dirt that can be generated from wool. But a problem remains: why couldn't God just rearrange the matter in a horse, say, to make it into a pile of dirt? So then a

modes of transmutation Buridan discusses here are "Wool can become a hatchet" (wool > earth > stone > iron > hatchet), and "An infant can build a house" (infant > adult human > carpenter).

[19] "Oppositum determinat Aristoteles. Querit enim quando aliquid debeat dici in potentia et quando non. Et dicit quod aliquid non debet dici in potentia ad aliquam formam, nisi quando sola transmutatio requiritur per quam illa forma perducatur" (*QM* 9.5, fol. 58rb). Buridan seems to have in mind Aristotle's *Physics* 1.4 (188^{a}32–^{b}3).

[20] "Modo in proposito est bene aqeuivocatio de potentia propinqua et remota" (*QM* 9.5, fol. 58va).

[21] "Quia ex quolibet potest fieri quodlibet—licet per multas transmutationes" (*QM* 9.5, fol. 58rb).

doi: 10.48106/dial.v76.i4.04

horse is possibly dirt (and vice-versa).[22] And if so, then our main problem comes roaring back: everything is possibly everything.

Buridan himself does not consider this problem, but there is indirect textual evidence that he would reject such a claim: after all, he frequently tells us that the following is impossible:

(2) A human is a donkey.

Granted, it is not beyond divine power to transform the matter of a human being into a donkey by imparting on it the appropriate form. But again, (2) is impossible. How?

The solution is to appeal to the notion of change entailing annihilation (or destruction—more on this in a moment), which we saw above in connection with *de quando* necessity. For example, consider the following sentence:

(3) Socrates is a human.

Any formulation of (3) is true whenever Socrates exists. And while (3) can be rendered false, this can only happen by the destruction of Socrates. Similarly if, instead of being served a hemlock cocktail, Socrates met his demise by having his matter suddenly morphed into the form of a donkey, (3) would become false. But so would the claim that Socrates himself is a donkey, since Socrates himself would no longer exist. So Socrates is not possibly a donkey.

We have limited ourselves to transmutation in talking about things-possibly-being-other-things, and to one transmutation at that. Granted, then, God can morph Socrates' matter into a donkey. But this morphing does not count as a transmutation in the natural sense, nor is it a potency belonging to Socrates. And so this fact no more entails that Socrates is a possible donkey than does the fact that Socrates can die and decay into soil, which then nourishes a plant, which a donkey eats, etc.

Here, then, we return to the original claim that *impossibilia* are incompossible combinations: donkey-Socrates, chimaeras—anything, in short, made up of parts that cannot be combined. Soon, we will see that Lewisian possible worlds, too, are Buridanian *impossibilia*. But first, we have to find a way of making the foregoing definitions consistent.

[22] I'm aware I am treading dangerously close to an old problem at which even young Socrates is reported to have balked: does dirt have an essence? (*Parmenides* 130c–d). I wish to remain neutral on this point: for my purposes, the only concession I have to make is that whatever makes horses horsey is essentially different from whatever makes dirt dirty. Maybe I beg the question on this. But I invite you to beg it with me. After all, we're in good company, historically speaking.

2.4 *What are Buridanian* Possibilia*?*

In a seminal (1989) paper, G.E. Hughes raises several questions about Buridan's modal logic and its underlying ontology. Concerning the latter, he tells us (1989, 97):

> For a long time I was puzzled about what Buridan could mean by talking about possible but non-actual things of a certain kind. Did he mean by a "possible A," I wondered, an actual object which is not in fact A but might have been, or might become, A? My house, e.g., is in this sense a possible green thing because, although it is not in fact green, it could become green by being painted. But this interpretation won't do; for Buridan wants to talk, e.g., about possible horses; and it seems quite clear that he does not believe that there are, or even could be, things which are not in fact horses but which might become horses.

Here Hughes makes no mention of the *Metaphysics* discussion—about horses, too!—which we just considered. This comes as no great surprise: that text is, to this day, neither edited nor translated.[23]

Here, Hughes's initial proposal is quite close to Buridan's own account: a house is a possible green thing, because there are powers in the world capable of making it so. The issue of substantial change—things becoming horses—is somewhat more thorny, since it seems odd to speak of things which are not horses, but which could become horses, as Hughes observes. And yet this is precisely what we are warranted to do, as Buridan explicitly tells us, provided we limit ourselves to at most one transmutation: horse semen is not a horse, but it is a possible horse.

Frustrated by his version of the horse puzzle, and unaware of Buridan's *QM* discussion, Hughes falls back on the familiar framework of possible worlds:

> What I want to suggest here, very briefly, is that we might understand what he says in terms of modern "possible world semantics." Possible world theorists are quite accustomed to talking about possible worlds in which there are more horses than there are in the actual world. And then, if Buridan assures us that by "Every horse can sleep" he means "Everything that is or can be a horse

[23] Granted, Hughes himself did know Latin, and was experienced in palaeography. He even edited a portion of the *Logica Magna* of Paul of Venice (ca. 1369–1429). Still, one can't read everything.

> can sleep," we could understand this to mean that for everything that is a horse in any possible world, there is a (perhaps other) possible world in which it is asleep. It seems to me, in fact, that in his modal logic he is implicitly working with a kind of possible worlds semantics throughout.

Here, Hughes first claims that Buridan's modal logic can be understood using the modern apparatus of possible worlds semantics. But then he strengthens that claim: Buridan *is* in fact working with possible-worlds semantics, however implicitly.

From what we've seen of Buridan so far, we can see that at least the latter claim is mistaken. Buridan's view of modality is grounded in *causation*: if there exists no power to make S to be not P (at least without annihilating S), then S is necessarily P. Likewise, if S can be made to be P (through at most one transmutation), then S is possibly P. Thus something's modal properties are grounded in the powers that exist *in this world*, which are capable of making it to be this or that way. In other words, Buridanian *possibilia* are, in general terms, objects, some of them nonexistent, whose modality depends on the causal powers of actually existing things. Since one of these existing things is the Almighty, and since the Almighty exists by simple (which is to say strictly unalterable) necessity, the modal properties of the *possibilia* are stable. There are no other worlds in the picture.

So much for what Buridan's view is not. But the definitions we've distilled from the texts face an important exegetical problem: both necessity, on one hand, and possibility, on the other, are each in their own way inconsistent with the account of impossibility as sketched above. Impossibility, unlike necessity, does not turn on annihilation: a chimaera is made up of incompossible parts, not parts that would be literally reduced to nothing if they were combined. Moreover, there are diachronic possibilities, such as a human turning into a corpse, which are not synchronically possible: a human cannot be inanimate and rational at the same time. Just like *chimera*, *inanimate rational animal* therefore picks out an impossible object. The language of transmutations is therefore not applicable to synchronic incompossibilities. These facts call for a re-examination of necessity and of possibility as set out above. We will soon see that (i) these accounts can, happily, be made consistent, and (ii) that the consistent account that emerges gives us a straightforward definition of Buridanian *possibilia*.

First, the account of necessity, which turns on annihilation (rather than destruction) of the subject is too strong: for there is more than one way to make Socrates not a human: through (divine) annihilation—literal reduction to nothing—or through (divine or natural) destruction—undergoing a change that entails removal of his (human) essence. After all, following his death, Socrates is no longer a human, but this fact does not turn on any annihilation of Socrates.

Why then does Buridan discuss necessity in terms of annihilation at all? Recall that, in the *QAPr*, Buridan is (*inter alia*) worried about the falsification of geometry: if *all* magnitudes were annihilated, then the propositions of geometry would be rendered false. But this would not follow if everything with mass were simply destroyed—that is, if everything now existing were reduced to an undifferentiated soup. Even in that soup, there would be at least some dimension, surface, and so on. Conversely, the claim that humans are animals *would* be falsified if all humans were destroyed—that is, if everyone died all at once. Hence it seems that the reliance on annihilation is stronger than it needs to be for the definition of humans as animals, though perhaps not for the propositions of geometry taken collectively. I therefore propose a weakening of this requirement, at least for our definition of *possibilia*: S is necessarily P, just in case S cannot be made other than P without *destroying* S.

The second exegetical problem is that the definition of possibility is quite weak: supposing that S is possibly P just in case S can become P through at most one transmutation, it follows that Socrates, while still alive, is possibly a corpse. Fair enough; but, as we observed, the combination of Socrates, *qua* rational animal, and corpse, *qua* inanimate object, is impossible.[24] Therefore, the most straightforward reading of impossibility, set out in section 2.2, clashes with the weak sort of possibility set out in section 2.3. What do we do?

It is true that Socrates is possibly a corpse. And it is also true that Socrates, while alive and barbate, is possibly clean-shaven. In the former case, Socrates loses his essence; in the latter he does not. We should therefore distinguish two kinds of change: one which involves loss of essence, but only through one transmutation; and another which leaves the subject intact.

Which kind of possibility is relevant to our purposes? *Impossibilia* are incompossible combinations; *possibilia* then should be possible ones. Since at least some transmutations involve change into something incompossible with

[24] For a discussion of related problems in the logic and semantics of the twelfth century, see Cameron (Cameron, M. A. 2015).

the essence of the subject, as our example of *rational animal* and *inanimate object* shows, *possibilia* cannot comprise contrary diachronic states considered synchronically. We should, therefore, take the stronger reading of possibility, suggested by the account of impossibility: S is possibly P iff S can be P in a way that does not entail the destruction of S.

From these considerations, we can give the following definition of *possibilia*, which balances out the accounts in Buridan's texts:

> POSSIBILIA$_B$. S is possibly P just in case there is a power to make S to be P without destroying the essence of S.[25]

This definition casts a pretty wide net: *possibilia* will include not just the various natural kinds and subkinds we see in the world, but also anything else which could be produced by any power—including God—without destruction of the subject. So horses larger than planets are, presumably, (divinely) possible; as are humans capable of walking on water, virgin mothers, and so on. But conspicuously absent from this jungle of *possibilia* is the Lewisian plurality of worlds with which we began.

3 Are Lewisian Possible Worlds Possible?

—Or, to put the question in Buridanian terms: can God create a Lewisian plurality of worlds? First, the argument pro: it seems that God can indeed create as many worlds as God pleases. Recall our account of the unity of Lewisian worlds, set out above (section 1.1). So long as we conceive of a world as just a cluster of spatiotemporally interrelated *possibilia*, there seems to be no barrier in principle to clustering them. Here is why: some—and probably most—possible objects are made up of interrelated possible parts. Consider, for example, a possible watch that does not now exist. Such a possible watch will

[25] As an anonymous reviewer for this journal has pointed out, this definition, and the intuitions that motivate it, rest on essentialist assumptions. That is true, but the assumptions are weak ones: we need not assume that we have correctly identified the essence of S; we need only say that as a member of a natural kind, S *has* an essence—whether or not we know what it is. Still, one might worry about possibilities for houses and other artifacts, since (at least in Aristotelian metaphysics) artifacts do not have essences. A house, then, is possibly green, and also possibly a heap of rubble, and neither of these changes involves a loss of essence. Perhaps we could appeal to the house's function, which is preserved in the case of painting, but lost when it is reduced to rubble. But I leave that for another day.

not be undifferentiated all the way through, like *pâté*, but will have interrelated possible parts—possible gears, possible springs, etc.

Now it would be arbitrary and just plain wrong to place a limit on how large such a possible object could be, at least in terms of what God can create: if a watch can be made the size of a tower clock, why not a watch the size of Manhattan? Likewise, it would be arbitrary to place a limit on their complexity: if a watch the size of Manhattan is permissible, why not a huge and complex astronomical horologium—one as large and complex as our universe, even?

From these considerations, we can distill two principles, namely:

(i) *possibilia* can be internally complex, comprising interrelated possible parts; and
(ii) there is no limit in principle to the size or complexity of such *possibilia*.

From (i) and (ii)—so the argument runs—it follows that God could make worlds, roughly construed as manifolds of interrelated objects.

In fact, we can strengthen this claim: the *possibilia* just *have* to be in some possible world. Consider a possible object, say a fork: can such an object exist outside a world or manifold? Or must any such possible object exist within some kind of manifold? The existence of a fork outside some spatiotemporal manifold seems, if not impossible, then at least a little weird. A fork in the absence of other objects is one thing, but a fork in the absence of space-time is quite another. And so, it seems, possible objects only ever inhabit worlds. Thus a metaphysics of possible objects must, if it is to be coherent, collapse into a metaphysics of possible worlds.[26]

So much for the argument pro; now for the argument contra. These worlds are either actual, in the sense that God has made them, or they are possible but non-existent, in the sense that God has not made them, but could. In either case, the question is: could God make an actual plurality of worlds? If so, then the Lewisian plurality is possible; if not, then it is impossible.

Following Lewisian doctrine, these worlds will have to be isolated: if they are not, they no more count as distinct possible worlds than do planets in different galaxies or cities in different epochs. They must not be at any spatiotemporal distance from each other. So can God create worlds that are not worldmates in this way?

Suppose God made these worlds. What does it mean to say such worlds are causal isolates—i.e., that they cannot interact? Distance will not do the

[26] I owe the gist of this argument to Douglas Campbell.

doi: 10.48106/dial.v76.i4.04

trick: worlds are not causally isolated by any spatiotemporal distance, the way you and I are isolated from a long-dead star in Andromeda. Space is not what separates the worlds. Nor is time. Lewis has been clear.

Perhaps we can say that God stipulates that the worlds cannot interact: there is just an impermeable barrier between the worlds, analogous to the glass plates separating different tanks in a divided aquarium, or the walls splitting off different theaters in a cineplex. Perhaps it is physical, perhaps it is by divine *fiat*. Either way, we face three problems.

First, what happens when two things in different worlds interact with the dividing barrier or *fiat* that separates them? Suppose, for instance, that there is a barrier between worlds A and B; and *a* and *b*, which are possible objects in A and B respectively, are blocked from interacting by the barrier/*fiat* (imagine fish bumping into the opposite sides of a glass aquarium divider). Then a barrier that prohibits causal interaction between the two worlds, A and B, nevertheless causally interacts with both of them. Therefore, that barrier will be a member of both worlds, according to Lewis's definition: it has worldmates on both sides. But preventing such world-straddling was precisely what the barrier was supposed to do. We can try adding barriers so that the two barriers on the A and B sides are separated, a bit like parallel sheets of glass in a double-paned window. But then we get a regress: what keeps the barriers themselves apart? What would happen if one barrier collided with whatever separates it from the other? In any case, the barriers must both interact with whatever separates them.

Second, even if God could somehow separate A and B causally from each other, it would still make sense to think of them as related temporally: just as we can speak of one movie in a cineplex starting at the midpoint of another, so we can speak of a universe being half as old as another—that is, as being created midway along the life cycle of another universe. For instance, we could reasonably ask whether, from God's perspective, the timeline of B is half as long as that of A, whether B already existed when A was created, and so on.

Third, and most importantly, even if such worlds could be isolated from each other in a way that circumvents the foregoing two problems, they will still still be causally related via their causal dependence on God. Recall, from section 1, that the general spatiotemporal relation (though not necessarily causation) is Euclidean: if $x\mathrm{R}y$ and $x\mathrm{R}z$, then $z\mathrm{R}y$. Thus although two worlds may not causally interact, they are not spatiotemporally independent, since they have the same cause. They are, then, causal siblings, even if they never interact. And if they are produced by the same cause, then they are causally related, if only in virtue of being created by the same God.

Lewis considers pseudo-pluralities like these (1986, 72), which, according to him, are not made up of truly isolated worlds. Their constituents are, rather, worldmates, even if locally they look like isolated worlds. Here is the one our cineplex and aquarium examples most closely resemble:

> The spacetime of the big world might have an extra dimension. The world-like parts might then be spread out along this extra dimension, like a stack of flatlands in three-space.

But, as Lewis is quick to point out, this is not a true plurality. Thus there is no way, on Lewis's account, to speak of temporal relations across truly isolated worlds: if there is anything like a God's eye view of the sort we have been discussing, then the worlds belong to the same manifold. And if they belong to the same manifold, they are not truly isolated.[27]

Here is the most common objection I have faced to this line of reasoning: it is not that Lewisian worlds *cannot* interact, in the sense that there is some mechanism keeping them apart. Instead, they just *do not*. We already noticed (in section 1.1, above) that the isolation doctrine is not a conclusion Lewis reaches by argument. It is, rather, a stipulation. And in fact, this is how Lewis presents it: right up front, on the second page of his (1986) exposition. It is thus more a starting point than a destination.

Accordingly, no criticism of this doctrine can address Lewis's arguments for it, since he does not give us any. All that can be asked is whether it makes any sense. The answer, on Buridan's metaphysics (or any metaphysics that posits one First Cause), is *no*. To anyone who espouses such a metaphysics, then, a Lewisian plurality of worlds must be something like Naive Set Theory: plausible on the face of it, but deep down self-contradictory. Lewis's worlds simply do not work on Buridan's framework. And, we might think, so much the better for Buridan.

I am not, by the way, the first to make any claims about the (in)compatibility of Lewisian worlds with classical theism, though the causal one I have been elaborating here is novel. Paul Sheehy (Sheehy 2006) sets out a number of

[27] Something similar could be said for the synchronic contrary possibilities of Scotus' (much discussed) *Lectura* I, dist. 39, q.1–5. Since these possibilities are rooted in the causal powers of a (single) will, they are worldmates. Therefore, these synchronic contrary possibilities are not true worlds in the Lewisian sense. For a discussion of Scotus in terms of possible worlds, see Wyatt (2000).

doi: 10.48106/dial.v76.i4.04

problems for the classical theistic conception of God on Lewisian modal metaphysics. The most significant of these is his argument, suggested by Richard Davis (Davis 2008), that Lewisian possible worlds effectively chop God up, making each counterpart God a world-bound entity—an understanding that runs contrary to classical theism's commitment to divine unity. Ross Cameron (2009) disagrees: Lewisian metaphysics can countenance *abstracta* existing outside of any world, as numbers do, so long as these *abstracta* are pure sets—that is, sets which contain only sets in their transitive closure (sets, sets of sets, sets of sets of sets, and so on, but no elements anywhere but sets, including the empty set). God, it seems, could be such a set—even if it's doubtful whether such a set is what God's believers believe in (or, anyway, believe they believe in). Subsequent debate Collier (2021) has dealt with this problem of divine (unitary) existence and world-boundedness, and whether, in these ways, God can be countenanced on Lewisian worlds. Brian Leftow (2012, 541–545) has, moreover, criticised Lewis on the grounds that positing one God is more economical than positing several (more on this in a moment).

For my part, I agree with Cameron and Collier that a Lewisian ontology can indeed countenance an abstract, un-world-bound Necessary Being of sorts. And I agree with Sheehy and Vance that Lewisian worlds are incompatible with classical theism, albeit for reasons different from the ones they examine. After all, it is integral to classical theism that God has a creative—which is to say causal—role to play as well: God "created the heavens and the earth" (Genesis 1:1), is the One without Whom "nothing was made that was made" (John 1:2), the Originator, "Who commands only"Be!" and it is" (*Al Baqarah* "The Heifer," 117), and so on. (Countless other sources could be cited to this effect, but you get the idea). This central aspect of God's activity is incompatible with Lewis's doctrines about the plurality of worlds. Accordingly, possible worlds of the sort we have considered here will be deeply incompatible with (monotheistic) medieval philosophy in general—even if certain aspects of a given thinker's modal logic or ontology might remind us of this (by now quite familiar) framework.[28]

[28] This will be true even when philosophical discussion centers on the notion of multiple worlds, e.g., in the claim of Al Ghazali and the Ashʿarite theologians that God could have made other worlds than this one. Here, too, the worlds that could exist are referred back to a single unified power to bring them into existence, and so there is a similar problem for Lewis's separation doctrine to the one discussed above. For a lively and interesting overview of this aspect of Al Ghazali's thought, see Taneli Kukkonen (2000). (I am grateful to Silvia Di Vincenzo for bringing this to my attention).

What about Lewisian metaphysics considered in its own right? Even though a unified First Cause is not available on this framework, it does not follow that Lewis and his followers have to be atheists; if there is plurality in the worlds, there can also be a plurality of first causes. There is textual evidence that Lewis recognises this implication of his theory: in the introduction to the first volume of his (Lewis 1983) *Philosophical Papers*, he remarks in passing that his view is consistent with the claim that "there are countless gods but none of them are our worldmates" (xi). Since the worlds are, ontologically speaking, just like ours, it follows that our worldmates could include a local deity, and Lewis could merely be mistaken about the constituents of our actual world. So the Lewisian can still opt for a kind of polytheism, or mono-poly-theism, to adapt a term coined by Hart (2013, 127). But even basic classical monotheism is, on these lights, impossible. For Lewisian ontology is a jealous god.[*]

Boaz Faraday Schuman
0000-0001-5763-8628
University of Copenhagen
boaz.schuman@hum.ku.dk

References

Cameron, Margaret Anne. 2015. "The Logic of Dead Humans. Abelard and the Transformation of the Porphyrian Tree." in *Oxford Studies in Medieval Philosophy*, volume III, pp. 32–63. Oxford: Oxford University Press, doi:10.1093/acprof:oso/9780198743798.003.0002.

Cameron, Ross P. 2009. "God Exists at Every (Modal Realist) World: Response to Sheehy (2006)." *Religious Studies* 45(1): 95–100, doi:10.1017/S0034412508009827.

Collier, Matthew James. 2019. "God's Necessity on Anselmian Theistic Genuine Modal Realism." *Sophia. International Journal of Philosophy and Traditions* 58(3): 331–348, doi:10.1007/s11841-018-0659-4.

—. 2021. "God's Place in the World." *International Journal for Philosophy of Religion* 89(1): 43–65, doi:10.1007/s11153-020-09764-w.

[*] I am grateful for the insightful comments and criticisms of Irene Binini, Jon Bornholdt, Douglas Campbell, Nate Charlow, Jack Copeland, Graziana Ciola, Silvia Di Vincenzo, Peter King, Gyula Klima, Fred Kroon, Chris Martin, Calvin Normore, Diane Proudfoot, Jeremy Seligman, Trevor Teitel, and Andrew Withy. Thanks are also due to audiences at the University of Auckland, and the University of Canterbury, Christchurch, as well as audiences at meetings of the American Philosophical Association–Eastern, and the European Symposium of Medieval Logic and Semantics. The paper has, finally, benefitted from the feedback of two anonymous reviewers for *Dialectica*. Thanks, everyone!

Davis, Richard Brian. 2008. “God and Modal Concretism.” *Philosophia Christi* 10(1): 57–74, doi:10.5840/pc20081014.

Dutilh-Novaes, Catarina. 2007. *Formalizing Medieval Logical Theories: suppositio, consequentiae and obligationes*. Logic, Epistemology, and the Unity of Science n. 7. Dordrecht: Springer Verlag, doi:10.1007/978-1-4020-5853-0.

Ebbesen, Sten. 1986. “The Chimera’s Diary.” in *The Logic of Being – Historical Studies*, edited by Simo Knuuttila and Jaakko Hintikka, pp. 115–144. Synthese Historical Library n. 28. Dordrecht: Kluwer Academic Publishers.

Frege, Gottlob. 1884. *Die Grundlagen der Arithmetik: eine logisch-mathematische Untersuchung über den Begriff der Zahl*. Breslau: Wilhelm Koebner. Reissued as Frege (1961).

—. 1961. *Die Grundlagen der Arithmetik*. Hildesheim: Georg Olms Verlagsbuchhandlung.

Grant, Edward. 1979. “The Condemnation of 1277, God’s Absolute Power, and Physical Thought in the Late Middle Ages.” *Viator* 10: 211–249, doi:10.1484/J.VIATOR.2.301526.

Hart, David Bentley. 2013. *The Experience of God: Being, Consciousness, Bliss*. New Haven, Connecticut: Yale University Press, doi:10.12987/9780300167337.

Hughes, George E. 1989. “The Modal Logic of John Buridan.” in *Atti del convegno internazionale di storia della logica: le teorie delle modalità: San Gimignano, 5-8 dicembre 1987*, edited by Giovanna Corsi, Corrado Mangione, and Massimo Mugnai, pp. 93–111. Bologna: Cooperativa Libraria Universitaria Editrice (CLUEB).

Jespersen, Otto. 1924. *The Philosophy of Grammar*. London: George Allen & Unwin.

Johnston, Spencer. 2015. “A Formal Reconstruction of Buridan’s Modal Syllogism.” *History and Philosophy of Logic* 36(1): 2–17, doi:10.1080/01445340.2014.934090.

—. 2017. “The Modal Octagon and John Buridan’s Modal Ontology.” in, pp. 35–52, doi:10.1007/978-3-319-45062-9_4.

King, Peter O. 2021. “Causal Powers in the Latin Christian West.” in *Powers. A History*, edited by Julia Jórati, pp. 112–142. Oxford Philosophical Concepts. Oxford: Oxford University Press, doi:10.1093/oso/9780190925512.003.0008.

Klima, Gyula. 2004. “Consequences of a Closed, Token-Based Semantics: the Case of John Buridan.” *History and Philosophy of Logic* 25(2): 95–110, doi:10.1080/01445340310001610944.

Kukkonen, Taneli. 2000. “Possible Worlds in the *Tahâfut al-Falâsifa*: Al-Ghazâlî on Creation and Contingency.” *Journal of the History of Philosophy* 38(4): 479–502, doi:10.1353/hph.2005.0033.

Leftow, Brian. 2012. *God and Necessity*. Oxford: Oxford University Press, doi:10.1093/acprof:oso/9780199263356.001.0001.

Lewis, David. 1983. *Philosophical Papers, Volume 1*. Oxford: Oxford University Press, doi:10.1093/0195032047.001.0001.

—. 1986. *On the Plurality of Worlds*. Oxford: Basil Blackwell Publishers.

Priest, Graham. 2005. *Towards Non-Being. The Logic and Metaphysics of Intentionality*. Oxford: Oxford University Press, doi:10.1093/0199262543.001.0001.

—. 2016. *Towards Non-Being. The Logic and Metaphysics of Intentionality*. 2nd ed. Oxford: Oxford University Press. First edition: Priest (2005), doi:10.1093/acprof:oso/9780198783596.001.0001.

Schopenhauer, Arthur. 1819. *Die Welt als Wille und Vorstellung*. Leipzig: Bibliographisches Institut, F.A. Brockhaus.

Sheehy, Paul. 2006. "Theism and Modal Realism ." *Religious Studies* 42(3): 315–328, doi:10.1017/s0034412506008419.

—. 2009. "Reply to Cameron, R. P. (2009)." *Religious Studies* 45(1): 101–104, doi:10.1017/S0034412506008419.

Soames, Scott. 2010. *Philosophy of Language*. Princeton, New Jersey: Princeton University Press, doi:10.23943/princeton/9780691138664.001.0001.

Thijssen, Hans. 2018. "Condemnation of 1277." in *The Stanford Encyclopedia of Philosophy*. Stanford, California: The Metaphysics Research Lab, Center for the Study of Language; Information. Revision, November 13, 2018, of the version of January 30, 2003, https://plato.stanford.edu/entries/condemnation/.

Thom, Paul. 2003. *Medieval Modal Systems: Problems and Concepts*. Aldershot, Hampshire: Ashgate Publishing Limited.

Vance, Chad. 2016. "Classical Theism and Modal Realism are Incompatible." *Religious Studies* 52(4): 561–572, doi:10.1017/S003441251600010X.

Williamson, Timothy. 2013. *Modal Logic as Metaphysics*. Oxford: Oxford University Press, doi:10.1093/acprof:oso/9780199552078.001.0001.

Wyatt, Nicole. 2000. "Did Duns Scotus Invent Possible Worlds Semantics?" *Australasian Journal of Philosophy* 78(2): 196–212, doi:10.1080/00048400012349481.

Zupko, John Alexander [Jack]. 2003. *John Buridan: Portrait of a Fourteenth-Century Arts Master*. Notre Dame, Indiana: University of Notre Dame Press.

—. 2018. "John Buridan." in *The Stanford Encyclopedia of Philosophy*. Stanford, California: The Metaphysics Research Lab, Center for the Study of Language; Information. Revision, July 3, 2018, of the version of May 13, 2002, https://plato.stanford.edu/entries/buridan/.

Avner Baz's Ordinary Language Challenge to the Philosophical Method of Cases

Paul O. Irikefe

Avner Baz argues that the philosophical method of cases presupposes a problematic view of language and linguistic competence, namely what he calls “the atomistic-compositional view”. Combining key elements of social pragmatism and contextualism, Baz presents a view of language and linguistic competence, which he takes to be more sensitive to the open-endedness of human language. On this view, there are conditions for the “normal” and “felicitous” use of human words, conditions that Baz thinks are lacking in the context of the philosophical method of cases, and which make the question that philosophers are prone to ask in that context and the answers they give to that question to be pointless. However, in this paper, I argue as follows. First, Baz's conditions for the “normal” and “felicitous” use of human words are in tension with the open-endedness of human language and the use of human words. Second, it is not even clear that those conditions are really missing in the context of the philosophical method of cases. And third, even if we grant that those conditions are missing in that context, this does not licence his damning conclusion on the philosophical method of cases since we are not forced to embrace the view of language and linguistic competence on which that damning conclusion is plausible. This last move is secured by advancing and defending a skill or virtue-based view of language and linguistic competence inspired by the later work of Donald Davidson.

The philosophical method of cases (henceforth, PMOC) arguably plays some role in how philosophers investigate issues of great philosophical interest like knowledge, free will, and reference.[1] In this practice, a philosopher would

[1] There is evidence of the use of the method outside the Western tradition of philosophy (Boh 1985).

describe a certain scenario, whether real or hypothetical, and invite us to say whether the case so described would count as falling under the relevant property or term or concept under investigation. The judgement formed on the described scenario is then enlisted in arguing for or against certain philosophical views.[2]

The question then is, what linguistic competence guides this practice? In some very illuminating works, Avner Baz (2016, 2017) argues that the PMOC presupposes a problematic view of language and linguistic competence, what he calls the atomistic-compositional view. The atomistic-compositional view as he presents it is presupposed by defenders of the method in mainstream analytic philosophy and critics of the method, including experimental philosophers. Combining key elements of social pragmatism and contextualism, Baz presents what we might call a "social pragmatic view of language", a view he thinks enjoys better empirical support and is more sensitive to the open-endedness of human language. On this view, there are "normal" and "felicitous" conditions for the use of words and human language, conditions he takes to be lacking in the context of the PMOC and the questions philosophers are prone to ask in that context such as: "Does *X* know *Y*?"

However, in this paper, I argue as follows. First, Baz's conditions for the normal and felicitous use of words and language stand in tension with the open-endedness of words and language. Second, it is not even clear that those conditions are really missing in the context of the PMOC. And third, even if we grant that those conditions are missing in that context, this does not licence any damning conclusion on the PMOC since we are not forced to embrace the view of language and linguistic competence on which that conclusion seems plausible. This last move is secured by advancing and defending a skill-based view of language and linguistic competence inspired by Donald Davidson (1986).

The paper proceeds as follows. In section 1, I discuss what Baz calls the "minimal assumption" about language which he says is presupposed by both armchair philosophers and their experimental counterparts. I show that the assumption expresses two worries. The first is the correctness worry and the second is about the kind of linguistic competence we rely on in the PMOC,

[2] In recent times, there have been serious discussions about the evidential status of these judgements, in particular, whether this status is due to their being intuitive (Cappelen 2012, 2014; Deutsch 2015; Earlenbaugh and Molyneux 2009; Ichikawa and Jarvis 2009; Irikefe 2020; Williamson 2007). I would set aside this issue in this paper by staying neutral about the evidential nature of these judgements since nothing here hangs on it.

which he calls the "atomistic-compositional" view. I briefly respond to the first worry, and I indicate that the second worry is more pressing and would therefore be of present concern. In section 2, I discuss Baz's social pragmatic view of language and linguistic competence, which he takes to have better empirical support than the atomistic compositional view. I explore some of the ingredients of the social pragmatic view, its negative implications for the PMOC and why we might worry that some aspects of the view do not seem consistent with recognisable features of the PMOC and the nature of human language itself. In section 3, I explore how we might look to defend or rely on the PMOC without any problematic assumptions about language and linguistic competence and without either the atomistic compositional view or Baz's social pragmatic view. I end the paper by showing how the present defence of the PMOC meshes with a broader trend in the epistemology of philosophy and lends independent support to it.

1 The Atomistic-Compositional View of Language and the Philosophical Method of Cases

The philosophical method of cases is a standard practice in analytic philosophy. A philosopher wants to argue for or against certain views about knowledge, causation, free will or moral permissibility. An imaginary scenario is described, and we are asked whether or not a certain property, term or concept obtains in the described scenario. For example, in Gettier's 10-coin case, we are asked the question whether the protagonist in the described scenario knows some particular proposition, that is, whether the protagonist knows that the man who will get the job has ten coins in his pocket (Gettier 1963).

According to Baz (2016, 2017), the method depends on a "minimal assumption" about language to get off the ground, namely, the assumption that questions like that as presented in the context of the PMOC are

> in principle, in order—in the simple sense that they are clear enough and may be answered correctly or incorrectly—and that, as competent speakers, we ought to understand those questions and be able to answer them correctly, just on the basis of the descriptions of the cases and our mastery of the words in which the questions are couched. (Baz 2017, 6)

doi: 10.48106/dial.v76.i4.03

We can distinguish two kinds of worries in the minimal assumption. The first one is the correctness worry, namely, the worry whether the questions at stake in the method of cases can be answered correctly or incorrectly, rightly or wrongly, and what the ontological status of such answers might be like, precisely whether these answers would be about concepts or the world independent of concepts (Baz 2017, 6). Baz links this worry with what he calls the "representational-referential" view of language and traces it to Timothy Williamson (2007), Herman Cappelen (2012) and Frank Jackson (2011). On this view, *the primary function* of language at any given moment or as he puts it "the fundamental aim of (all?!) discourses" (Baz 2017, 74, fn. 6) is to say true or false things about the world. Although this is not the worry I intend to address in this paper, I believe that friends of the PMOC do not need to commit themselves to any problematic assumption here. On the contrary, I think *pace* Baz, what they need to hold is that *among other things that human language is for*, human language is used to say true or false things about the world (I would return to this in section 3). In the same vein, friends of the method may not need to settle the issue of what the answers to the questions at stake in the method of cases would be true of, whether they would be true of our concepts or items in the world existing independently of our conception of them. As Ernest Sosa noted: "We can conduct our controversies, for example, just in terms of where the truth lies with regard to them, leaving aside questions of objectual ontology" (Sosa 2007, 100–101).

The second worry in the minimal assumption is the more pressing one. And it is the one I wish to address in this paper. It says that

> as competent speakers, we ought to understand those questions [i.e., the questions at stake in the method of cases] and be able to answer them correctly, just on the basis of the descriptions of the cases and our mastery of the words in which the questions are couched. (Baz 2017, 6)

Baz notes that this is an assumption about language that derives from and is dependent on the atomistic-compositional view of language. In this view, the meaning of the whole of an utterance comes from the fixed meaning of the parts of that utterance. Baz traces the atomistic-compositional view of language to Jackson (2011), who presents it as the linguistic competence that the method depends on. Jackson says that how a sentence like "it is raining outside" represents things is a

> function of the representational contents of its parts and how they are combined.[3] Moreover, we have a grasp of the representational contents of these parts, and of the way various modes of combination into sentences generate representational structures whose contents are a function of the contents of their parts and the way the parts are put together. [Jackson (2011), 472][4]

In Jackson (2011), this view of language and linguistic competence goes side by side with a view of conceptual competence. On this view, in learning philosophically significant terms like "knowledge" we are latching onto the pattern or rule or categorisation of "knowledge." Thus, he says:

> How did we acquire the word "knowledge"? We came across lots of examples. We were told a bit about what mattered. Perhaps, we were simply instructed that if it is false, it cannot be knowledge. At some point we latched onto *the* pattern. (Jackson 2011, 474)

This rule or pattern on Jackson's view in turn guides our knowledge ascriptions, that is, it enables us to say whether or not the protagonist in a Gettier text knows or does not know a given proposition.

In the next section, I consider Baz's argument that the atomistic-compositional view of language is problematic and his argument that in the context of the PMOC the conditions for the normal and felicitous use of words and language are lacking. As we shall see too, Baz takes himself to be establishing a demarcation of the boundary of linguistic sense, one that makes clear that the PMOC is outside that boundary and that the questions philosophers are prone to ask in that context are fundamentally problematic.

[3] Compare the atomistic-compositional view with the view of Paul Elbourne: "Suppose you are interpreting an uttered sentence. In a series of extremely intricate processes that are largely subconscious, you access the sentence's words in your mental lexicon and find their meanings; you work out the intended sense of any ambiguous words it might contain; you work out the references of indexicals in the sentence; you work out the sentence's syntactic structure and resolve any structural ambiguities there may be; and you combine the contents of the words in the compositional semantics… If implicit content is not mediated by means of covert indexicals (and thus covered by the second step mentioned above), you add some of this too. Finally, you have worked out the content of the sentence, as uttered on that occasion". (Elbourne 2011, 131), cited in (Baz 2017).

[4] Although differently expressed, Baz identifies Williamson as holding this view as well: "[E]xpressions refer to items in the mostly non-linguistic world, *the reference of complex expressions is a function of the reference of its constituents*, and the reference of a sentence determines its truth value". (Williamson 2007, 281, emphasis mine)

doi: 10.48106/dial.v76.i4.03

2 Baz's Social Pragmatic View of Language and Linguistic Competence

The way Baz shows the atomistic-compositional view of language to be problematic is by presenting and defending an alternative view of language that he takes to enjoy better empirical support. He finds support from a scientific study of how children acquire their first natural language (Bartsch and Wellman 1995). But Karen Bartsch and Henry Wellman were not interested in natural language acquisition for its own sake. More specifically, they were tracking the natural development in the use of belief-desire terms in children between ages one and a half to six years. Six of these children are boys and four are girls. One of them is African American and the others are not. Because of their interest, Bartsch and Wellman were necessarily selective. They were coding only for terms expressing genuine psychological reference, where this is judged so if with respect to a suitable context it referred to psychological states like desire, belief, or knowledge. As a result, they discounted conversational use of belief-desire terms like when a child says "you know what?" when seeking to get someone's attention; repetition of phrases uttered by someone else, for example, a mother saying "tell him you know where it is," to which the child responds "I know where it is", and so on.

For present purposes, let us focus on what the study uncovered about the term "knows" and its cognates. The authors found (as Baz pointed out) that the word "knows" and its cognates do not admit of a simple formula. More specifically, they found that children use "knows" and its cognates to refer to instances of belief "felt to be justified, assumed to be true, or that enjoys markedly higher conviction than one described by *think*" (Bartsch and Wellman 1995, 40). Later on in their development, they use it to refer to "situations involving successful actions or to correct statements" (Bartsch and Wellman 1995, 60). In other words, there is no single pattern that a child is trying to master in being a competent user of "knows" and its cognates.

What is interesting about this study as Baz rightly observed is that it is one of the few scientific studies that have focussed on philosophically interesting terms like "knows" and its cognates. Most scientific studies about words and concepts are usually too broad in their scope and coverage to tell us what we need to know in doing philosophy. This is important because although the empirical result is not yet conclusive, it indicates that ordinary words like "table" are not just like philosophically interesting words like "knows"; the latter is more complex and traces no single or simple pattern *pace* Jackson.[5] It also indicates, as Baz argued, that human language is open-ended, that is, capable of being used to make completely new moves not just at the level of the whole of an utterance but at the level of the individual parts or words in a way that is problematic for the atomistic-compositional view of language and linguistic competence. For present purposes, we can take the current empirical evidence for granted, and inquire into how to make sense of it.

Baz thinks that the best way to make sense of the data is a view that combines contextualism and social-pragmatism, a view whose central ingredients come from Wittgenstein's (1953) *Philosophical Investigations* and Merleau-Ponty's (2002). Following Wittgenstein, Baz argues that we need to think of meaning as use in the sense that the significance of words depends not on their referring to items but "on whether and how we use the words, on our *meaning* them in one way or another, in a context suitable for meaning them in *that* way" (Baz 2017, 130). The advantage of the usage view in Baz's opinion is that it shows clearly that our words need not be representational and need not be thought of as naming items in the non-linguistic world to be suitable for different uses.[6]

Following Merleau-Ponty, Baz argues that we need to reclaim the place of the *actual speaker* in the speech act, "the person who finding herself in some particular situation or other, may find herself moved, motivated, to speak (or think)" (Baz 2017, 131). This means that understanding the speech of another is not merely the putting together of the already fixed meaning of her words, but "coming to see her point," meaning coming to see her cares, her commitments, her history, how she sees the situation, and so on. In a significant sense therefore, the view reverses the direction of linguistic meaning implied in the atomistic compositional view: we understand the parts of speech by first understanding the whole of it, and that requires understanding the point of the actual speaker. In this connection, Baz notes that:

[5] In fact, we do not need the study of how children acquire "knows" and its cognates to realise that words like "knows" do not trace a simple pattern that can be framed in terms of necessary and sufficient conditions for all instances of knowledge. We already have reasons to suspect that this is so from the failure to produce a simple account of necessary and sufficient conditions for knowledge in analytic epistemology (Shope 1983).

[6] Baz says too that Wittgenstein's comparison of words to game pieces also lends credence to this idea of language.

doi: 10.48106/dial.v76.i4.03

> The notion of "motive" is very important to Merleau-Ponty's avoidance of both mechanistic and intellectualist approaches to the understanding of behavior in general and linguistic expression in particular (see (Merleau-Ponty 2002, 48–50)). On Merleau-Ponty's way of looking at things, our speech (and behavior more generally) is normally *motivated*, in the sense that we are not merely *caused* mechanically to speak, and in the sense that our behavior manifests an *understanding* of the phenomenal world to which we respond. (Baz 2017, 131, fn. 14)

Baz argues that this view of language and linguistic competence gives support to a social-pragmatic account of conceptual competence inspired by Michael Tomasello (2003, 2008). On this view, in being a competent employer of "knows" and its cognates, what the child learns is different actual constructions of speech and their communicative functions, or more plainly, "stored exemplars of utterances" (Baz 2017, 162) "and what commitments (liabilities, risks) one takes upon oneself when using the words in one way or another, and in responding in one way or another to other people's use of them" (Baz 2017, 169).

Furthermore, Baz thinks that if we accept this way of thinking about language, linguistic competence, and conceptual competence, the PMOC would be found to be seriously defective. How so? Well, if understanding the speech of another is coming to see the point of an *actual* speaker, which means coming to see her *cares,* her *commitments* vis-à-vis the question, and what *risks* and *liabilities* she may assume in answering the question one way or the other, and what empirical options we might explore to investigate whether things are thus and so, and what practical interest makes that question intelligible either to us or to the speaker, and how what is said in that context may influence what we do after; it seems clear that these conditions are lacking in the context of the PMOC. And it is because Baz thinks these conditions—let us call them "social-pragmatic conditions"—are not so realised in the PMOC that he takes the PMOC to be deeply defective and the questions asked in that context to be pointless as well. Put more generally, the view is the following:

> The Social Pragmatic View of Language and Linguistic Competence. If Hearer *H* in a context *C* understands the speech of a speaker *S*, *H* does so only if the social pragmatic conditions are realised in context *C*.

Notice that the view is silent as to the further question of whether the social pragmatic conditions are the only conditions required for linguistic understanding to be possible or for words to be meaningfully used. It merely says that the social-pragmatic conditions are essential or necessary for words to do their work and for questions to have intelligibility.

One urgent question is, why commit Baz to the broader goal of demarcating the region of the meaningful use of words rather than the more modest view that the questions asked by the practitioners of the PMOC are problematic or pointless?[7] Or put differently, why think that Baz's criticism concerns the descriptions of the PMOC rather than the questions themselves and whether or not the questions are pointless? Well, the short answer is that the questions themselves are pointless precisely because the social pragmatic conditions for the felicitous use of words by both hearers and speakers are lacking in the description of the case. Baz says this precisely when he tries to show how his project fits within a broader demarcation argument that goes back to champions of experimental philosophy such as Jonathan Weinberg (2007), and more recently Edouard Machery (2017). This kind of argument relies on showing that there is a discontinuity between the scenarios described in the PMOC and the scenarios that we regularly encounter in everyday situations in a way that makes the former bad and the latter good. However, doing that often requires coming up with a set of properties defining one context but not the other context.[8] Here is textual evidence that lends support to construing Baz in this way.

> The argument of this book is meant to show that the discontinuity is primarily a matter, not of *the sorts of cases* theorists have tended to focus on, as Weinberg has suggested, but of *the peculiar context* in which we attend to those cases and try to answer the theorist's questions. (Baz 2017, 33, fn. 33)

And again:

> [But] if as I will argue, the ordinary and normal *conditions* for the felicitous use of the word (or concept) under investigation are lacking in the theoretical context—and, again, lacking by design—

[7] An anonymous reviewer for this journal pressed me on this objection.

[8] For replies to Weinberg's claim of discontinuity, see Cappelen (2012) and Nado (2015); and for a reply to Machery, see Nado (2022).

doi: 10.48106/dial.v76.i4.03

> then there is good reason to worry that the theorizing is bound to distort what it aims to clarify. (Baz 2017, 3)

Notice that the theoretical context is also the peculiar context. Notice too that if we seek to restrict Baz's demarcation only to occasions of speech when terms like "knows" and "cause" are featured, this would be *ad hoc*. The reason these terms retain philosophical interest is due to their everyday provenance. Indeed, 'knows' and its cognates are some of the most ubiquitous terms in human language.

There are two worries I would like to point out here. The first is this. Baz's claim of discontinuity implies that in the peculiar context of the PMOC, some essential conditions for the felicitous use of human words are lacking in a way that problematises the kind of questions philosophers are prone to ask in that context, as well as the answers they give. But this stands in tension with the open-endedness of human language. How so? The idea that language is open-ended, if it means anything really, means that whatever set of conditions we can identify and establish as part of the normal and felicitous use of language and words, there would always be occasions where those conditions are unmet, and yet a speaker with some ingenuity employs it in a meaningful way; a way that transmits knowledge or understanding or that serves other useful functions. Of course, language is not a human practice where anything goes. However, the thought is that given proper context, speakers and hearers can always tell the difference between what is meaningful and what is not without any predetermined criteria. Further, the thought is that these criteria, if any, would not be something that can be captured in any principled way and articulatable as something like some social pragmatic conditions. Moreover, the realm of meaning and meaningful questions and answers involving terms like "knows" and "cause" is not correctly restricted to the realm of the pragmatic or the practical for creatures like us. And that is because human beings have a capacity to engage meaningfully in things that transcend their self-interest. It seems that for evolutionary reasons, this would be a good thing. Information that has no pragmatic import for a hearer in a given context and at a particular time can have life-saving significance for that agent in a different context at a future time or perhaps for close kin. Edward Craig has a similar story of how our practically oriented concept of knowledge evolved into a more objectivised and demanding standard, where a high degree of reliability even in an improbable world is built into it. Thus, he says:

> In saying that someone knows whether *p* we are certifying him as an informant on that question, and we have no idea of the practical needs of the many people who may want to take him up on it; hence a practice develops of setting the standard very high, so that whatever turns, for them, on getting the truth about, we need not fear reproach if they follow our recommendation. (Craig 1990, 94)

Perhaps it is also why "knows" and its cognates have some exceptional qualities such as being lexical universals, with the rare quality of being in the core vocabulary of all known human languages (Haspelmath and Tadmor 2009), and having a one-word equivalent in all natural languages (Goddard 2010).

The second worry: Baz is assuming that in the theoretical or peculiar context of the PMOC, nothing hangs for the hearers and speakers, or the thought experimenter and his or her audience except for a theoretical interest, namely, the affirmation or the refutation of a view. But can we take that assumption for granted? I think not. For very often, the success of counterexamples or more generally, philosophical cases is decisive for the dominance of a particular theory and field of research. Think about the debate between compatibilism and incompatibilism, internalism and externalism, physicalism and anti-physicalism and the decisive role that thought experiments played in those debates like Mary the colour scientist case (Jackson 1982), Gettier cases (Gettier 1963) and Truetemp case (Lehrer 1990). True enough, we only care about the truth or facts that obtain or do not obtain in those cases rather than their instrumental value. And yet because of the role those cases play in the rise and fall of certain fields of research and research prospects, it is fair to say that the facts that obtain or fail to obtain in those cases make those cases stand in the same relation to real or actual situations that are of interest to Baz: They are not idle issues to which we feel unconcerned and to which our interests, cares, and commitments are unrelated.[9]

In the next section, I discuss a further challenge for Baz's account, namely, the problem of malapropism, which shows that sometimes the conditions for the ordinary use of our words are violated, and yet linguistic understanding

[9] In the same vein, it is not clear that there is nothing we can do to find out whether the verdict in the cases is correct or incorrect. Indeed, this is what experimental philosophers have been doing. Although one might argue that consensus or corroboration is not correctness of intuitional judgements. But so too are perceptual judgements.

is possible. This then sets the stage for presenting and developing a Davidson-inspired alternative view of language and linguistic competence.

3 The Skill or Virtue-Based Account of Language and Linguistic Competence

In his later writings, Davidson found the problem of "malapropism" very perplexing. Dealing with this problem led him to a view of language that affirms a continuity between linguistic competence and intellectual abilities more generally. To be sure, malapropism is a ubiquitous phenomenon in human language and registers

> our ability to perceive a well-formed sentence when the actual utterance was incomplete or grammatically garbled, our ability to interpret words we have never heard before, to correct slips of the tongue, or to cope with new idiolects. (Davidson 1986, 95)

On the standard view of language and linguistic competence, what a hearer needs to be able to interpret a speaker is something like a complex theory or rule plus the ability to use this rule or theory or generalisation in a systematic way to make sense of novel situations. Further, because this capacity is taken as a learned convention, one that is shared between hearers and speakers, it is something that the hearer has in advance of the occasion of linguistic exchange. Notice that this standard view is also the view defended by Jackson as previously presented and discussed (Jackson 2011). Recall that on that view, namely, the atomistic-compositional view, language is like the numbering system where there are finite numerals that can be used to generate complex ones infinitely. Speakers and hearers have this system in advance of particular linguistic exchanges.

However, the phenomenon of malapropism challenges this notion because the competence (or capacity) that it calls for from the hearer is not part of what normally constitutes one's basic linguistic competence, mastered in advance of the occasion of linguistic exchange. Indeed, as Davidson points out, the fact that makes the theory or rule general equally makes it unsuitable to cope with the particular linguistic habits of different individuals, say that of Mrs. Malaprop's "nice derangement of epitaphs" being "nice arrangement of epithets".[10] More generally, the theory or rule is unhelpful in coping with a

[10] Malaprop was a character famous for her verbal blunders in Richard Sheridan's play *The Rivals*.

particular speaker at a particular time in a particular occasion. This applies to Baz's account too since for him there are "ordinary and normal conditions for the felicitous use" of human words or concepts (Baz 2017, 3), conditions which he thinks are lacking in the context of the PMOC. But then, in malapropism such as grammatically garbled utterances and slips of tongues, those normal conditions for the felicitous use of words and for their "functioning as they do" in ordinary discourse (Baz 2017, 22) are violated. Further, it is not the case that for Baz there is one generic condition, namely, that one's utterance has a point. On the contrary, that one's utterance has a point is fixed by it satisfying "the ordinary and normal conditions" for the felicitous use of human words and for meaning words one way or the other. For he says:

> And the basic problem with so much philosophizing, both traditional and contemporary—the basic problem with the method of cases as commonly practiced, for example—is that the philosopher either takes his words to mean something clear even apart from *his* meaning something clear by means of them, or else takes himself to be able to mean his words in some determinate way, *even though the conditions for thus meaning his words are missing in his particular context* and cannot be created by a sheer act of will, or by concentrating one's mind in some special way. (Baz 2017, 141, italics mine)

Here is an additional challenge from malapropism to any generic view of language and linguistic competence. Sometimes in linguistic exchange, linguistic understanding is transmitted despite the hearer completely mistaking the speaker's verbal communication and vice versa. Davidson gives an example of such a case:

> When I first read Singer's piece on Goodman Ace, I thought that the word 'malaprop', though the name of Sheridan's character, was not a common noun that could be used in place of 'malapropism'. It turned out to be my mistake. Not that it mattered: I knew what Singer meant, even though I was in error about the word; I would have taken his meaning in the same way if he had been in error instead of me. We could both have been wrong, and things would have gone as smoothly. (Davidson 1986, 90)

Here as elsewhere, learned convention breaks down and the conditions for the normal and felicitous use of words are violated and yet linguistic under-

doi: 10.48106/dial.v76.i4.03

standing is transmitted or made possible. The question is, how is this possible? What capacity does the hearer (and speaker) depend on? Davidson makes the following suggestion:

> This characterisation of linguistic ability is so nearly circular that it cannot be wrong: it comes to saying that the ability to communicate by speech consists in the ability to make oneself understood, and to understand. It is only when we look at the structure of this ability that we realise how far we have drifted from standard ideas of language mastery. For we have discovered no learnable common core of consistent behaviour, no shared grammar or rules, no *portable* interpreting machine set to grind out the meaning of an arbitrary utterance. We may say that linguistic ability is the ability to converge on a passing theory from time to time—this is what I have suggested, and I have no better proposal. But if we do say this, then we should realise that we have abandoned not only the ordinary notion of a language, but we have erased the boundary between knowing a language and knowing our way around in the world generally. (Davidson 1986, 445–446, italics mine)

We can summarise the import of this account as follows:

> The skill or virtue-based account of language and linguistic competence. If Hearer *H* in a context *C* understands the speech of a speaker *S*, *H* does so in virtue of her skills or virtues.

The rationale for speaking of skills or virtues here is two-fold. First, it is to pick up on a suggestion by Davidson when he talks about the skillful hearer (and speaker) as being one that can get along well in linguistic exchanges and performances without needing mastery or knowledge of Gricean principles, because these general principles "are a kind of *skill* we expect of an interpreter and without which communication would be greatly impoverished" (Davidson 1986, 437). Relatedly, he talks about virtues such as practical wisdom, intelligence, and wit as the non-linguistic competencies we rely on in getting things right from time to time, occasion to occasion (Davidson 1986, 446). Davidson also mentions luck. But here luck is not a capacity of speakers or hearers. Rather, it merely refers to their being in a favourable environment such that under normal circumstances, when they attempt to understand one another in linguistic exchange, they achieve that aim. Further, I persist in speaking of "skills and virtues" because although all skills can be classified as virtues of agents, not all virtues can be classified as skills. One particular exception to this is practical wisdom (Stichter 2018). Let us take these points in turn. First, virtues are skills because acting well is much like working well (Annas 1995) and both involve practices of self-regulation to achieve a goal: in one case, the goal of acting well, and in the other case, the goal of working well (Stichter 2018). And second, although practical wisdom involves some elements of skills, namely, making good judgements in particular situations, it also involves other dimensions, namely, considering how one's action fits into an overall conception of the good life (Stichter 2018). So, while it might be true that agents rely on some aspects of practical wisdom in order to act well in particular situations and to get along in a linguistic exchange, practical wisdom in itself is too broad and varied to be classified merely as a set of skills.

Furthermore, the competent hearer (and speaker) would also recruit other capacities of the virtuous agents. Of particular importance in the present context would be "sensibility." In her discussion of the virtues (and the vices of the mind), Alessandra Tanesini defines sensibility as a disposition to "use one's perceptual capacities in distinctive ways in the service of epistemic activities" (Tanesini 2021, 27). The example she gives is the observant person:

> The person who is observant has reliable vision but he also experiences as salient those features of the visual field that are relevant to his epistemic aims. He directs visual attention to these aspects of the environment. By directing attention to them, and thus putting them at the centre of his visual field, he is able to take in more detail about these items since foveal vision has a higher degree of resolution than peripheral vision. Had those items remained at the periphery of his vision, many of their features would have remained undetected. If this is right, being observant is the complex disposition to detect the salient aspects of the environment by experiencing feelings that direct one's attention towards these features. (Tanesini 2021, 27–28)

Applied as a competence essential to linguistic understanding, sensibility is an auxiliary competence, an enabler of visual and auditory competencies of agents. And what that means precisely is that it makes it possible for one to put to use those primary competencies in picking up what is being passed across, verbally and non-verbally, where this is something that can be missed easily if one is not attentive to another's peculiar linguistic habits in the context of linguistic exchange.

The second rationale for the skill, or virtue-based model, is that it allows us to cash out the Davidson-inspired view in a way that makes the relevant competence an instance of a more general and familiar kind of know-how. One difficulty that we can resolve in Davidson's account if we take seriously the virtue or skill-based model is how to understand a practice that is non-rule-based and yet rational and well-ordered. And the thing to say is that in both virtue and skills, we already have human practices that are well-regulated without the agents relying on rules. Take the skill-based model. Following this model, I am suggesting that knowing a language is much like knowing how to drive a car. In the beginning, the driver learns rules of thumb such as "shift up when the motor sounds like it is racing and down when it sounds like it is straining."[11] As Dreyfus and Dreyfus who have studied human skills in various domains of performance argued:

> It seems that beginners make judgements using strict rules and features, but that with talent and a great deal of involved experience the beginner develops into an expert who sees intuitively what to do without applying rules and making judgements at all. (Dreyfus and Dreyfus 1991, 235)

On this thinking, if one is following rules in a practice, that just shows one is not yet proficient in that practice. The same story applies to the virtuous agent. As Linda Zagzebski puts it: "Persons with practical wisdom learn how and when to trust certain feelings, and they develop habits of attitude and feeling that enable them to reliably make good judgments without being aware of following a procedure" (Zagzebski 1996, 226). Notice too the role of the virtues and skills here: they are dispositions that allow agents to act in a systematic and organised way and to do so well in a context where the relevant practice is not rule-governed. Plausibly, the reason this is so is because both skill and virtues have a kind of *logos*, in the sense that they have an intrinsic intellectual structure built into them (Bloomfield 2000). Mastering a skill, including language, is mastering this *logos*; and thus, possessing the practical

[11] Such rules of thumb are just heuristics or generalisations about language that hold for the most part.

intelligence to act and to sensibly follow the actions of others and to solve problems in the relevant domain or activity.

From this standpoint, we can appreciate another respect in which the skill or virtue-based account and Baz's view diverge. On Baz's account, the motive of the speaker plays an essential role in coming to see the point of the speaker. Notice that "motive" here does not mean intention. It means rather the "motivating factors", which are internal to the perspective of the speaker, namely, the cares, the commitments, the risks and the liabilities of the speaker. On the other hand, for the skill or virtue-based account, that component is not always essential even though it sometimes can form a part of the process of coming to see the point of the speaker's utterance. Indeed, I believe that that form of internalism about linguistic sense, or meaning, was part of the tradition of thought that Gilbert Ryle tries to wean analytic philosophy from (see also, Putnam (1975b)) when he argued that we should think of understanding as *knowing how* and linguistic understanding including, as an exercise of that *knowing how.* He writes:

> Understanding a person's deeds and words is not, therefore, any kind of problematic divination of occult processes. For this divination does not and cannot occur, whereas understanding does occur. Of course, it is part of my general thesis that the supposed occult processes are themselves mythical; there exists nothing to be the object of the postulated diagnoses. But for the present purpose it is enough to prove that, if there were such inner states and operations, one person would not be able to make probable inferences to their occurrence in the inner life of another. (Ryle 2009, 41)

Let me elaborate more on what this rejection of the internalistic picture in the motivating sense means by commenting on what Ryle is getting at here. Suppose I am playing chess with Magnus Carlsen, the Norwegian grandmaster. He makes a particular opening move that seems initially surprising to me. But as a fellow grandmaster who is equally skilful or competent in the game and who has sufficient experience dealing with a move like that, I can know what that move is about without caring about what has made Carlsen make this move. I can know that a move like that in a context like this means that a particular form of attack on my king is imminent and that moving my pieces in a specified way is the best way to counter it. The same is true of "moves" in linguistic performances, as Baz would like to call human utterances or the use of words in language. Hearers can tell that an utterance like this in a context like that means so and so without caring about what has moved the speaker to say so and so.

With this view of language and linguistic competence in mind, let us address two challenges in connection with the PMOC. The first challenge here is to explain how, as competent speakers, we are able to understand and answer the questions that philosophers often ask in the context of the PMOC, such as, does the protagonist in that scenario know so and so? And the second challenge is how to make the aim of using the PMOC intelligible in the light of the complexity of human language, that is, without glossing over that very complexity. I take each in turn.

On the skill or virtue-based view, competent speakers can understand and answer the questions of the sort "does *X* know *Y*?" not because they have latched onto the pattern of "knows" *pace* Jackson or because they possess stored exemplars of utterances and knowledge of the communicative motives of speakers *pace* Baz. On the contrary—when they do, that is in virtue of their having mastered a technique in the use of "knows" and its cognates. In fact, this suggestion finds its earliest expression in the later Wittgenstein when he says:

> The grammar of the word "know" is evidently closely related to the grammar of the words "can", "is able to." But also closely related to that of the word "understand" (*To have 'mastered' a technique*). [Wittgenstein (1953), § 150-151, italics mine][12]

[12] Should we read Wittgenstein's suggestion as the mastery of grammatical rules or relationships? There is abundant evidence in the text and elsewhere that that is not what Wittgenstein had in mind. To start with, in the paragraphs that followed this statement (i.e., Wittgenstein 1953, sec. 151–152), he says that it is conceivable that the relevant formula (or rule or grammatical relationship) might occur to the speaker and yet the speaker fails to understand. Further, in an unpublished manuscript, translated by Norman Malcolm (1989), Wittgenstein writes: "Often one can say: this pattern looked at so, must have this continuation. I want, however, to stipulate an 'interpretation' [*Auffassung*], (something like the old 'Proposition'), which determines the series like an infallible machine through which a conveyor belt runs. So that only this continuation fits this interpretation. In reality, however, there are not two things that here fit together. But one can say: You are by your training, so adjusted [*eingestellt*], that always, without reflection, you declare some definite thing to be that which fits. Something that agrees with what others declare to be what fits" (Wittgenstein, Unpublished manuscript, 86-87; cited in Malcolm (1989)). On this view, it is by one's training as a member of a shared community and practice that one is able to reliably employ terms like "knows" and extend the practice in similar situations. For Wittgenstein, that

Such skills or techniques are suitably grounded in experience in such a way that the agents exercising them can always be counted upon to answer such questions in a range of situations, not only in actual ones but in possible ones that bear similarity to the actual ones, where what is "similar" cannot be established in any rigid way, for example, through the claim of discontinuity between the context of the PMOC and everyday contexts. Indeed, as argued earlier, being competent users of "knows" and answering questions such as "does X know Y?" in a range of situations might be part of our evolutionary heritage. Also, a recent trend in cognitive science seems to lend support to this skill-based suggestion. Here is Lawrence Barsalou and colleagues summarising the emerging consensus here:

> [C]onceptual knowledge is not a global description of a category that functions as a detached database about its instances. Instead, conceptual knowledge is the *ability* to construct situated conceptualizations of the category that serves agents in particular situations. [Barsalou et al. (2003), 89][13]

picture of a skill or technique grounded in training replaces the picture of the grammatical rule acting like an infallible conveyor belt that determines its extension in novel situations.

[13] As previously pointed out, Baz argues that the atomistic-compositional view seems to go side by side with the assumption that the primary purpose of language is to transmit information, that is, it seems to go side by side with the representational-referential view of language. Again, there is no need to hold on to that problematic assumption. All that is necessary for the philosophical method of cases to get off the ground once the atomistic-compositional view is set aside and the skill or virtue-based view is assumed is that among other things, language can be used to transmit information, where again given appropriate context agents can tell when this is the case. In fact, the empirical study that Baz analyses in support of his view does not presuppose otherwise. To see this, notice that although in Baz's discussion of this study, he cites the frequency with which children refer to their own mental states as clear vindication of his view of language, the data also show that this frequency diminishes as the children grow older. Bartsch and Wellman also note that "our data provide no evidence that a representational understanding of beliefs is a significantly later achievement, following only on the heels of an earlier 'connections' misconstrual of beliefs" (Bartsch and Wellman 1995, 57). Further, even in their first-person reference to mental states, the data do not contradict representational presuppositions. As the authors put it "[W]hen children first use *know* to refer to people's knowledge in our data, in their utterances coded as genuine psychological references, they primarily refer either to situations involving successful actions or to correct statements" (Bartsch and Wellman 1995, 60). And lastly, in an earlier study of our everyday conception of knowledge as manifested in words like "knows" and "knew", Perner (1991) shows that knowledge is associated with success and successful actions, with factual states of affairs and is formed by exposure to the relevant information or experience.

Now the second challenge. Using the PMOC, Edmund Gettier drew the attention of the philosophical community to an aspect of knowledge, namely, that the term is a success notion; the term does not apply to someone whose belief is chancy or accidental. Does that gloss over the complexities in our use of "knows" and its cognates? Baz thinks so (see Baz (2017, 122)). But there are good reasons to doubt that conclusion. To start with, notice that the idea that knowledge is a success term is implied in the result of the study of Bartsch and Wellman (1995). Further, imagine as we do in the analysis of knowing that we highlight "success" or "achievement" as a salient feature of the term "knows" and explain knowledge in terms of these notions (Greco 2010). I argue that doing so does not obscure the subject matter of philosophy as Baz implies. On the contrary, doing so advances our understanding of the subject matter. Indeed, this is closely related to scientific practice. Biologists know that the term "fish" picks out various kinds of properties such as having fins, having scales, having a tail, breathing underwater, being oviparous, not suckling one's young, and being cold-blooded. But from the point of view of understanding, and classifying future unknown cases, they merely highlight a fewer set of properties rather than all of the above, especially those that are natural and explanatory so that the term "fish" is used to refer to a completely aquatic, water-breathing, cold-blooded craniate vertebrate (Slote 1966). I believe the same story applies here to the PMOC in the analysis of knowing. In highlighting the fact that knowledge is a success term, we are able to track something important, deep and explanatory about this phenomenon, something we can also use to understand other terms or concepts or issues. For example, knowledge firsters use the suggestion that knowledge is a success term to understand the notion of intellectual ability or competence (Kelp 2021).

Let us conclude this section by noting how the skill or virtue-based model of language and linguistic competence shares something positive with Baz's social pragmatic account. Clearly, both recover the place of the speaking subject and reject the idea implied in the atomistic-compositional view that human words can speak for themselves, "over our heads as it were—and of language as a system of significant signs that does not depend on speakers (and listeners) for its ongoing maintenance" (Baz 2017, 96). Indeed, in evaluating Gettier cases, for example, we often need to tell whether or not and in what relevant sense the cases we are evaluating resemble clear instances where the property or term is clearly instantiated in a case. And "which way one goes depends on what one finds normal or natural, which partly depends on the past course of one's sense experience" (Williamson 2007, 190). Notice that the capacity to tell that something is "normal or natural" is much in line with the capacity that comes with practical wisdom, which is shaped by experience, including sense experience, and expressed in habits of attitude and feeling that enable one to reliably make good judgements without being aware of following any rule. Moreover, in a non-actual instance of a Gettier case, readers often need to follow "in their own imaginative construction the lead of the author of the examples" (Sosa 2009, 107), and they have to fill out the details of the stories, which are often partial and incomplete. Here as elsewhere too, one needs to tell whether or not and in what relevant sense the case one is evaluating resembles clear instances where the property or term is clearly instantiated. Moreover, which way one goes depends on what one finds normal or natural. Notice also that if the kind of story that particularists such as Jonathan Dancy tell about the use of thought experiments in moral philosophy is true, namely, that no suitable supply of general principles can help the moral agent in picking out what is morally salient about a case (Dancy 1985), then we have good reason to believe that even here what the agent does is to recruit the kind of capacities that the skill or virtue-based model highlights. In any case, a theory of language and linguistic competence begins from the correct assumption that ordinary speakers already do well in linguistic performances and presents an explanation of how speakers are able to so perform. I have argued that once we reject the atomistic compositional view, it does not follow that we must embrace the social pragmatic story and all the problems it poses for the PMOC.[14]

[14] An anonymous reviewer for this journal pressed the following worries. The first worry is that "the proposed virtue-based account of linguistic understanding is perfectly compatible with there being cases/situations in which it doesn't make sense to ask about a certain subject and a certain fact 'Does *S* know that *p*?' Hence, it seems to me that further argument is needed in order to make the case for the meaningfulness of the theorist's questions about the philosophical thought experiments discussed in Baz." Reply: The worry that there are particular cases, say some very outlandish cases, where it does not make sense to ask about a certain subject and a certain fact 'Does *S* know that *p*?' does not licence the general or global worry about the PMOC as discussed in Baz. Even mainstream philosophers themselves have expressed concern that some cases are so outlandish that they are not theoretically useful because they do not resemble cases we face in everyday life (Weatherson 2003, 8). Here is another related worry pressed by the reviewer: "Davidson will also need some distinction (or demarcation) between situations in which the utterances of a certain sentence, e.g., of the form '*x* knows that *p*' makes sense and situations in which it doesn't (because obviously, you cannot meaningfully utter just any sentence in any context). And it is not obvious to me that according to Davidson the first kind of situations won't be exactly the ones in which the relevant utterance has a point." Reply: It is not exactly clear why

4 Conclusion

In this paper, I have argued essentially that the philosophical method of cases does not need to presuppose the problematic view of language and linguistic competence Baz attributes to its practitioners or defenders—the atomistic compositional view. And neither do friends of the PMOC need to embrace the social pragmatic view that Baz presents with all its negative consequences for the PMOC. Let me end with where the Davidson-inspired skill or virtue-based view leaves us in terms of the epistemology of philosophy. In my opinion, it lends independent support to the view, now current in the epistemology of philosophy that the epistemology of philosophy is an application of social epistemology. Williamson (2007); Nagel (2012) and more recently Irikefe (2022) champion this epistemological thesis and it seems to me the right way to explain how philosophical knowledge is possible and how it can be defended against various challenges posed against it.[*]

Paul O. Irikefe
0000-0003-4086-8229
University of California, Irvine
Irikefep@uci.edu

References

Annas, Julia. 1995. "Virtue as a Skill." *International Journal of Philosophical Studies* 3(2): 227–243, doi:10.1080/09672559508570812.

Barsalou, Lawrence W., Simmons, W. Kyle, Barbey, Aron K. and Wilson, Christine D. 2003. "Grounding Conceptual Knowledge in Modality-Specific Systems." *Trends in Cognitive Science* 7(2): 84–91, doi:10.1016/S1364-6613(02)00029-3.

we need such a demarcation to start with since we can get along without it, and if a theory of language is an attempt to model what people already do, that is exactly the kind of story our theory should also be telling. Further, one might wonder whether such a demarcation does not imply by its very existence that there is a rigid boundary of what counts as meaningful linguistic occasions and what does not count as so. It does in my opinion. And it is why in the history of philosophy these kinds of projects, which seek to demarcate some regions of language as linguistically acceptable and others that are not on the basis of some criteria have had little or no success. In any case, the Davidson-inspired view shows us a way to proceed without it.

[*] I am grateful to two anonymous reviewers for this journal for their invaluable comments that helped shape the development of the paper to its current form. In the same vein, I thank Alessandra Tanesini and Mona Simion. Their comments were helpful in refining the paper's central claims.

Bartsch, Karen and Wellman, Henry M. 1995. *Children Talk About the Mind*. Oxford: Oxford University Press, doi:10.1093/oso/9780195080056.001.0001.

Baz, Avner. 2016. "Recent Attempts to Defend the Philosophical Method of Cases and the Linguistic (Re)turn." *Philosophy and Phenomenological Research* 92(1): 105–130, doi:10.1111/phpr.12106.

—. 2017. *The Crisis of Method in Contemporary Analytic Philosophy*. Oxford: Oxford University Press, doi:10.1093/oso/9780198801887.001.0001.

Bloomfield, Paul. 2000. "Virtue Epistemology and the Epistemology of Virtue." *Philosophy and Phenomenological Research* 60(1): 23–43, doi:10.2307/2653426.

Boh, Ivan. 1985. "Belief, Justification, and Knowledge–Some Late-Medieval Epistemic Concerns." *Quidditas: Online Journal of the Rocky Mountain Medieval and Renaissance Association* 6(1): 8, https://scholarsarchive.byu.edu/rmmra/vol6/iss1/8.

Cappelen, Herman. 2012. *Philosophy Without Intuitions*. Oxford: Oxford University Press, doi:10.1093/acprof:oso/9780199644865.001.0001.

—. 2014. "X-Phi Without Intuitions?" in *Intuitions*, edited by Anthony Robert Booth and Darrell P. Rowbottom, pp. 269–286. Oxford: Oxford University Press, doi:10.1093/acprof:oso/9780199609192.003.0015.

Craig, Edward J. 1990. *Knowledge and the State of Nature: An Essay in Conceptual Analysis*. Oxford: Oxford University Press, doi:10.1093/0198238797.001.0001.

Dancy, Jonathan. 1985. "The Role of Imaginary Cases in Ethics." *Pacific Philosophical Quarterly* 66(1–2): 141–153, doi:10.1111/j.1468-0114.1985.tb00246.x.

Davidson, Donald. 1986. "A Nice Derangement of Epitaphs." in *Truth and Interpretation: Perspectives on the Philosophy of Donald Davidson*, edited by Ernest LePore, pp. 433–446. Oxford: Basil Blackwell Publishers. Also published in Grandy and Warner (1986, 157–174), doi:10.7146/sl.v0i25.103806.

Deutsch, Max Emil. 2015. *The Myth of the Intuitive. Experimental Philosophy and Philosophical Method*. Cambridge, Massachusetts: The MIT Press, doi:10.7551/mitpress/9780262028950.001.0001.

Dreyfus, Hubert L. and Dreyfus, Stuart E. 1991. "Towards a Phenomenology of Ethical Expertise." *Human Studies* 14(4): 229–250, doi:10.1007/bf02205607.

Earlenbaugh, Joshua and Molyneux, Bernard. 2009. "If Intuitions Must Be Evidential then Philosophy is in Big Trouble." *Studia Philosophica Estonica* 2(2): 35–53, doi:10.12697/spe.2009.2.2.03.

Elbourne, Paul. 2011. *Meaning: A Slim Guide to Semantics*. Oxford: Oxford University Press.

Gettier, Edmund L., III. 1963. "Is Justified True Belief Knowledge?" *Analysis* 23(6): 121–123, doi:10.1093/analys/23.6.121.

Goddard, Cliff. 2010. "Universals and Variation in the Lexicon of Mental State Concepts." in *Words and the Mind: How Words Capture Human Experience*, edited by Barbara C. Malt and Phillip Wolff, pp. 72–92. New York: Oxford University Press, doi:10.1093/acprof:oso/9780195311129.003.0005.

Grandy, Richard E. and Warner, Richard, eds. 1986. *Philosophical Grounds of Rationality: Intentions, Categories, Ends*. Oxford: Oxford University Press.

Greco, John. 2010. *Achieving Knowledge. A Virtue-Theoretic Account of Epistemic Normativity*. Cambridge: Cambridge University Press, doi:10.1017/CBO9780511844645.

Haspelmath, Martin and Tadmor, Uri. 2009. "The Loanword Typology Project and the World Loanword Database." in *Loanwords in the World's Languages. A Comparative Handbook*, edited by Martin Haspelmath and Uri Tadmor, pp. 1–34. Berlin: De Gruyter Mouton, doi:10.1515/9783110218442.1.

Ichikawa, Jonathan Jenkins and Jarvis, Benjamin W. 2009. "Thought-Experiment Intuitions and Truth in Fiction." *Philosophical Studies* 142(2): 221–246, doi:10.1007/s11098-007-9184-y.

Irikefe, Paul Oghenovo. 2020. "A Fresh Look at the Expertise Reply to the Variation Problem." *Philosophical Psychology* 33(6): 840–867, doi:10.1080/09515089.2020.1761541.

—. 2022. "The Epistemology of Thought Experiments without Exceptionalist Ingredients." *Synthese* 200(3), doi:10.1007/s11229-022-03690-2.

Jackson, Frank. 1982. "Ephiphenomenal Qualia." *The Philosophical Quarterly* 32(127): 127–136. Reprinted in Jackson (1998, 57–69), doi:10.2307/2960077.

—. 1998. *Mind, Method, and Conditionals: Selected Essays*. London: Routledge, doi:10.4324/9780203271308.

—. 2011. "On Gettier Holdouts." *Mind and Language* 26(4): 129–157, doi:10.1111/j.1468-0017.2011.01427.x.

Kelp, Christoph. 2021. "Theory of Inquiry." *Philosophy and Phenomenological Research* 103(2): 359–384, doi:10.1111/phpr.12719.

Lehrer, Keith. 1990. *Theory of Knowledge*. 1st ed. Dimensions of Philosophy Series. Boulder, Colorado: Westview Press. Second edition: Lehrer (2000).

—. 2000. *Theory of Knowledge*. 2nd ed. Dimensions of Philosophy Series. Boulder, Colorado: Westview Press. First edition: Lehrer (1990).

Machery, Edouard. 2017. *Philosophy Within Its Proper Bounds*. Oxford: Oxford University Press, doi:10.1093/oso/9780198807520.001.0001.

Malcolm, Norman. 1989. "Wittgenstein on Language and Rules." *Philosophy* 64(247): 5–28, doi:10.1017/s0031819100044004.

Merleau-Ponty, Maurice. 1945. *Phénoménologie de la perception*. Tel. Paris: Gallimard.

—. 2002. *Phenomenology of Perception*. London: Routledge. Translation of Merleau-Ponty (1945) by Colin Smith, doi:10.4324/9780203994610.

Nado, Jennifer. 2015. "Intuition, Philosophical Theorizing, and the Threat of Skepticism." in *Experimental Philosophy, Rationalism, and Naturalism. Rethinking*

Philosophical Method, edited by Eugen FISCHER and John COLLINS, pp. 204–221. London: Routledge.

—. 2022. "Philosophizing Out of Bounds." *Philosophical Studies* 179(1): 319–327, doi:10.1007/s11098-020-01582-0.

NAGEL, Jennifer. 2012. "Intuitions and Experiments: A Defense of the Case Method in Epistemology." *Philosophy and Phenomenological Research* 85(3): 495–527, doi:10.1111/j.1933-1592.2012.00634.x.

PERNER, Josef. 1991. *Understanding the Representational Mind*. Cambridge, Massachusetts: The MIT Press, doi:10.7551/mitpress/6988.001.0001.

PUTNAM, Hilary. 1975a. *Mind, Language and Reality: Philosophical Papers, Volume 2*. Cambridge: Cambridge University Press, doi:10.1017/cbo9780511625251.

—. 1975b. "The Meaning of 'Meaning' ." in *Minnesota Studies in the Philosophy of Science, Volume VII: Language, Mind, and Knowledge*, edited by Keith GUNDERSON, pp. 131–193. Minneapolis, Minnesota: University of Minnesota Press. Reprinted in Putnam (1975a, 215–271).

RYLE, Gilbert. 2009. *The Concept of Mind*. London: Routledge, doi:10.4324/9780203875858.

SHOPE, Robert K. 1983. *The Analysis of Knowing. A Decade of Research*. Princeton, New Jersey: Princeton University Press.

SLOTE, Michael Anthony. 1966. "The Theory of Important Criteria." *The Journal of Philosophy* 63(8): 211–224, doi:10.2307/2023977.

SOSA, Ernest. 2007. "Experimental Philosophy and Philosophical Intuition." *Philosophical Studies* 132(1): 99–107, doi:10.1007/s11098-006-9050-3.

—. 2009. "A Defense of the Use of Intuitions in Philosophy." in *Stich and His Critics*, edited by Dominic MURPHY and Michael A. BISHOP, pp. 101–112. Philosophers and Their Critics. Oxford: Wiley-Blackwell, doi:10.1002/9781444308709.ch6.

STICHTER, Matt. 2018. *The Skillfulness of Virtue. Improving our Moral and Epistemic Lives*. Cambridge: Cambridge University Press, doi:10.1017/9781108691970.

TANESINI, Alessandra. 2021. *The Mismeasure of the Self: A Study in Vice Epistemology*. Oxford: Oxford University Press, doi:10.1093/oso/9780198858836.001.0001.

TOMASELLO, Michael. 2003. *Constructing a Language: A Usage-Based Theory of Language Acquisition*. Cambridge, Massachusetts: Harvard University Press, doi:10.2307/j.ctv26070v8.

—. 2008. *Origins of Human Communication*. Cambridge, Massachusetts: The MIT Press.

WEATHERSON, Brian. 2003. "What Good Are Counterexamples?" *Philosophical Studies* 115(1): 1–31, doi:10.1023/a:1024961917413.

WEINBERG, Jonathan M. 2007. "How to Challenge Intuitions Empirically Without Risking Skepticism." in *Midwest Studies in Philosophy 31: Philosophy and the Empirical*, edited by Peter A. FRENCH and Howard K. WETTSTEIN, pp. 318–343.

Boston, Massachusetts: Basil Blackwell Publishers, doi:10.1111/j.1475-4975.2007.00157.x.

WILLIAMSON, Timothy. 2007. *The Philosophy of Philosophy*. Oxford: Basil Blackwell Publishers, doi:10.1002/9780470696675.

WITTGENSTEIN, Ludwig. 1953. *Philosophical Investigations / Philosophische Untersuchungen*. Oxford: Basil Blackwell Publishers. Edited by G.E.M. Anscombe and R. Rhees, translation from the German by G.E.M. Anscombe.

ZAGZEBSKI, Linda Trinkaus. 1996. *Virtues of the Mind. An Inquiry into the Nature of Virtue and the Ethical Foundations of Knowledge*. Cambridge: Cambridge University Press.

Weakly Discerning Vertices in a Plenitude of Graphs

Eric E. Sheng

De Clercq (2012) proposes a strategy for denying purported graph-theoretic counterexamples to the Principle of the Identity of Indiscernibles (PII), by assuming that any vertex is contained by multiple graphs. Duguid (2016) objects that De Clercq fails to show that the relevant vertices are discernible. Duguid is right, but De Clercq's strategy can be rescued. This note clarifies what assumptions about graph ontology are needed by De Clercq, and shows that, given those assumptions, any two vertices are weakly discernible, and so are not counterexamples to the version of PII that requires only weak discernibility.

The Principle of the Identity of Indiscernibles (hereafter PII) states that there are no *solo numero* differences. In other words, between any two things that differ numerically (i.e., differ in identity), there is a non-numerical difference (a difference that is not merely a difference in identity). Various purported counterexamples to PII have been proposed, among them Black's (1952) two intrinsically identical spheres located two miles apart in empty space. Saunders (2003) and Ladyman (2005) point out that Black's spheres and similar examples do not violate the version of PII whereby only *weak discernibility* is necessary for non-identity. A relation *R* weakly discerns objects *a* and *b* if and only if *Rab*&*Rba*&¬*Raa*&¬*Rbb* (Caulton and Butterfield 2012, 50). Black's spheres are weakly discerned by the relation *being two miles from*. Leitgeb and Ladyman (2008) propose cases drawn from graph theory in which, they claim, two distinct objects are not even weakly discernible. Leitgeb and Ladyman claim that—whereas the two vertices in the graph consisting of two vertices and an edge connecting them are weakly discernible—the two vertices in the graph consisting of two vertices and no edges are not in any way discernible. De Clercq (2012) argues that Leitgeb and Ladyman's counterexample rests on a controversial view about the ontology of graphs, namely one that rejects assumption (i) below; and that on another plausible view about the ontology of graphs, which De Clercq favours, the case that Leitgeb and Ladyman propose is not a counterexample to PII, because the two vertices are discernible in virtue of the relations in which they stand in other graphs that contain them. Duguid (2016) objects that the two vertices are not discernible in virtue of such relations, so that, even granting De Clercq's favoured view about the ontology of graphs, Leitgeb and Ladyman's case is a counterexample to PII, even the version of PII that requires only weak discernibility.

In this note, I clarify what assumptions about the ontology of graphs are needed by De Clercq, and show that De Clercq's strategy can be rescued from Duguid's rejoinder insofar as it can be shown that, granted De Clercq's assumptions about the ontology of graphs, any two vertices are weakly discernible. I give an example of a relation that weakly discerns vertices: *x has greater degree in some graph than y*. If De Clercq is correct about the ontology of graphs, therefore, the purported graph-theoretic counterexamples that have been proposed do not falsify the version of PII that requires only weak discernibility and thus do not, in this respect, improve on Black's spheres.

1 De Clercq's Strategy

Graphs are arrangements of vertices and edges connecting vertices such that edges do not have a direction and any two vertices in a graph are either connected by one edge or not connected by any edge.[1] The *degree* of a vertex in a graph is the number of edges that connect it with other vertices in the graph. A vertex is *isolated* in a graph if and only if it has degree 0 in the graph. Two graphs are *isomorphic* if and only if (regardless of the identities of their vertices) they have the same structure of vertices and edges; that is, two graphs are isomorphic if and only if there is a bijection from the set of the vertices of the first graph to the set of the vertices of the second graph such that any two vertices are connected by an edge in the first graph if and only if their images under the bijection are connected by an edge in the second graph.

More formally, graphs are commonly defined set-theoretically, so that a *graph* G is an ordered pair (V, E) where V is a set of *vertices* and E is a (possibly empty) set of subsets of V that have two members, and any distinct vertices v and w in V are said to be connected in (V, E) by an *edge* if and only if $\{v, w\}$

[1] In *directed graphs*, edges have a direction. In *multigraphs*, vertices may be connected by more than one edge. Directed graphs and multigraphs are not considered in this note.

is a member of E.[2] Let us call the identification of graphs with ordered pairs of vertex and edge sets *Identity*. Leitgeb and Ladyman do not accept *Identity*, while De Clercq does. Other graph-theoretic terms can also be defined set-theoretically.

De Clercq defends PII against Leitgeb and Ladyman's purported counterexample by arguing that, in a graph G_0 that consists of two vertices a and b and no edges, a and b are discernible in virtue of the relations in which they stand in other graphs: "vertices in labeled graphs are always distinguishable, not just because they bear different labels, but also because they feature in (structurally!) different ways in different graphs" (2012, 670). The distinctness of a and b, for example, is, according to De Clercq, not a *solo numero* difference, because there is a graph G_2, consisting of three vertices a, b and c (where a and b are respectively identical to the vertices a and b in G_0) and an edge connecting a and c, in which a and b stand in different relations.

Two assumptions are necessary and sufficient for De Clercq's strategy: (i) that there are no *unlabelled graphs* such that their vertices are objects, and (ii) that if G_0 exists, then G_2 exists. Regarding (i): De Clercq and Leitgeb and Ladyman disagree about what unlabelled graphs are. According to Leitgeb and Ladyman, unlabelled graphs are graphs such that the vertices of an unlabelled graph are distinguished only by their relations within that unlabelled graph, and any isomorphic unlabelled graphs are identical. On this view, there are objects that are the vertices of unlabelled graphs, and vertices of distinct unlabelled graphs do not stand in relations of identity. De Clercq (2012, 666), in contrast, claims that "unlabelled graphs are not graphs but isomorphism classes of graphs" (that is, the equivalence classes into which the set of all graphs is partitioned by the isomorphism relation).[3] On this view, talk of the vertices of unlabelled graphs is not ontologically committing, and there are no unlabelled graphs such that their vertices are objects. Regarding (ii): Since, as specified above, G_2 is a graph some of whose vertices are respectively identical to some vertices of G_0, (ii) presupposes (iii) that some vertices in

[2] Note: this definition is not committed to identifying edges with sets of two vertices.

[3] Note: De Clercq's identification of unlabelled graphs with isomorphism classes is not necessary for his argument. One could instead claim, for example, that unlabelled graphs are mereological atoms that correspond one-to-one with isomorphism classes. But perhaps it is the best motivated of claims that imply (i).

distinct graphs are identical.[4] De Clercq (2012, 665–669) defends (i) and (iii) by appealing to the practice of graph theorists.[5]

Assuming uncontroversially that there exist graphs of three or more vertices and thus that there exists at least one vertex other than a and b, (ii) follows from the following claim:

> Plenitude. For every subset V of the set W of all vertices, for every (possibly empty) set E of sets consisting of two members of V, there exists a graph consisting of the vertices in V and edges connecting every pair of members u, v of V such that $\{u, v\}$ is in E.[6]

In turn, *Plenitude* follows from *Identity*, since any set exists if its members exist, and any ordered pair exists if its members exist. So, De Clercq can accept *Identity* and infer (ii) from *Identity*, but only if he also accepts *Plenitude*. Might one accept (ii) without accepting *Plenitude*? As noted above, (ii) implies (iii) that some vertices in distinct graphs are identical, or, in other words, that some vertices are contained by multiple graphs. De Clercq (2012, 666–667) argues that, while (iii) follows from *Identity*, it is also plausible in light of mathematical practice, independently of the truth of *Identity*. Nonetheless, as long as some vertices are contained by multiple graphs, it would be arbitrary to suppose that some finite graphs that can be formed out of vertices from W and edges connecting them exist but others do not. De Clercq's assumption of (ii), therefore, commits him to *Plenitude*.

It is *Plenitude* that leaves De Clercq's defence of PII vulnerable, even granting (i), to Duguid's reply: for any graph where a and b bear different relations, another graph exists in which a and b are permuted, so that (for instance) corresponding to G_2 there exists an isomorphic graph G_1 which consists of three vertices a, b and c, and an edge connecting b and c. Now, b has the property *being isolated in a graph consisting of three vertices and an edge connecting two of them*, in virtue of G_2, but a has the same property, in virtue of G_1. To

[4] Note: *Identity* is not necessary for De Clercq's strategy because (i) and (ii) are sufficient for it and do not imply *Identity*. *Identity* is also not sufficient for De Clercq's strategy because, although *Identity* implies that there are no graphs that are Leitgeb and Ladyman's unlabelled graphs, *Identity* does not imply that Leitgeb and Ladyman's unlabelled graphs do not exist, and as long as such entities exist, there are counterexamples to PII.

[5] In defending the rejection of (i), Leitgeb and Ladyman (2008, 390) also appeal to the practice of graph theorists.

[6] Note: *Plenitude*, thus formulated, does not presuppose *Identity*, as it would if "a graph consisting [...] is in E" were replaced with "a graph (V, E)."

discern *a* and *b*, De Clercq would have to appeal to properties that distinguish between isomorphic graphs (for example, the property *being isolated in* G_2).[7] But since distinct isomorphic graphs differ only in the identity of their vertices, in order to distinguish between isomorphic graphs, a property "must utilize object names" (Duguid 2016, 472). Let us say that a property or relation is *forbidden* for PII if and only if allowing being discerned by it to count as discernibility would make PII metaphysically uninteresting. For example, a version of PII that allows that *a* and *b* are discernible on the ground that they are discerned by the property *is identical to a* is metaphysically uninteresting, as is a version of PII that allows that *a* and *b* are discernible on the ground that they are weakly discerned by the relation *is distinct from*. ((Rodriguez-Pereyra 2006) and (Muller 2015) discuss what would make PII trivial and as such metaphysically uninteresting.) Following Muller (2015), Duguid considers properties in which object names occur to be forbidden for PII. De Clercq, Duguid concludes, fails to save PII.

2 Weakly Discerning Vertices

Ladyman, Linnebo and Pettigrew (2012) show in their Theorem 6.4 that two objects are weakly discernible in a language *L* if and only if they are in any way discernible in the language that includes a constant for every element of the domain of *L* (i.e., the language that includes names for all of its objects). It follows, as Duguid accepts, that, if there are object-name-containing properties that discern two vertices *a* and *b*, there is a non-object-name-containing relation that weakly discerns *a* and *b*. Nonetheless, Duguid writes (2016, 473): "such a relation has not yet been provided. And neither can I see what it might be."

Here is one:

$$\Phi(x, y) := \exists g\,((g \text{ is a graph}) \,\&\, (g \text{ contains } x \text{ and } y) \,\&\, (x \text{ has greater degree in g than } y))$$

Given De Clercq's assumptions, this relation, *x has greater degree in some graph than y*, holds between any two vertices *a* and *b* in both directions, but not

[7] Duguid (2016, 472) says that De Clercq must appeal to a property that is "specific enough to single out a single graph." This is not correct, since the property *is isolated in a graph consisting of a, b and some third vertex and an edge connecting a and the third vertex*, which does not single out a single graph, would also do.

doi: 10.48106/dial.v76.i4.06

between either vertex and itself. Hence, *contra* Duguid, any two vertices are weakly discernible, and Leitgeb and Ladyman's case is not a counterexample to the version of PII that requires only weak discernibility.

Whether De Clercq's strategy for saving PII from purported graph-theoretic counterexamples is ultimately successful depends on the plausibility of its assumptions about graph ontology: *Plenitude*, and that there are no unlabelled graphs such that their vertices are objects. Granted these assumptions, however, any two vertices are indeed weakly discernible.[*]

Eric E. Sheng
University of Oxford
eric.sheng@philosophy.ox.ac.uk

References

Black, Max. 1952. "The Identity of Indiscernibles." *Mind* 61(242): 152–164. Reprinted in Black (1954, 80–92), doi:10.1093/mind/LXI.242.153.

—. 1954. *Problems of Analysis: Philosophical Essays*. Ithaca, New York: Cornell University Press.

Caulton, Adam and Butterfield, Jeremy. 2012. "On Kinds of Indiscernibility in Logic and Metaphysics." *The British Journal for the Philosophy of Science* 63(1): 27–84, doi:10.1093/bjps/axr007.

de Clercq, Rafael. 2012. "On some Putative Graph-Theoretic Counterexamples to the Principle of the Identity of Indiscernibles." *Synthese* 187(2): 661–672, doi:10.1007/s11229-010-9867-3.

Duguid, Callum. 2016. "Graph Theory and the Identity of Indiscernibles." *Dialectica* 70(3): 463–474, doi:10.1111/1746-8361.12151.

Ladyman, James. 2005. "Mathematical Structuralism and the Identity of Indiscernibles." *Analysis* 65(3): 218–221, doi:10.1111/j.1467-8284.2005.00552.x.

Ladyman, James, Linnebo, Øystein and Pettigrew, Richard. 2012. "Identity and Discernibility in Philosophy and Logic." *The Review of Symbolic Logic* 5(1): 162–186, doi:10.1017/S1755020311000281.

Leitgeb, Hannes and Ladyman, James. 2008. "Criteria of Identity and Structuralist Ontology." *Philosophia Mathematica* 16(3): 388–396, doi:10.1093/philmat/nkm039.

Muller, F. A. 2015. "The Rise of Relationals." *Mind* 124(493): 201–237, doi:10.1093/mind/fzu175.

Rodriguez-Pereyra, Gonzalo. 2006. "How Not to Trivialiize the Identity of Indiscernibles." in *Universals, Concepts and Qualities: New Essays on the Meaning*

[*] THANKS

of Predicates, edited by Peter Frederick STRAWSON and Arindam CHAKRABARTI, pp. 205–224. Aldershot, Hampshire: Ashgate Publishing Limited.

SAUNDERS, Simon W. 2003. “Physics and Leibniz’s Principles.” in *Symmetries in Physics: Philosophical Reflections*, edited by Katherine BRADING and Elena CASTELLANI, pp. 289–307. Cambridge: Cambridge University Press, doi:10.1017/CBO9780511535369.017.

Published by *Philosophie.ch*

Verein philosophie.ch
Fabrikgässli 1
2502 Biel/Bienne
Switzerland
dialectica@philosophie.ch

https://dialectica.philosophie.ch/

ISSN 0012-2017

This work is licensed under a Creative Commons Attribution 4.0 International License.

Dialectica is supported by the Swiss Academy of Humanities and Social Sciences.

Abstracting and Indexing Services

The journal is indexed by the 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