**BELGRADE PHILOSOPHICAL ANNUAL/
FILOZOFSKI GODIŠNJAK 33/2020
Institute of Philosophy, Faculty of Philosophy,
University of Belgrade
Belgrade, Čika Ljubina 18–20**

Belgrade

Year XXXII

YU ISSN 0353-3891

UDK-1

Belgrade Philosophical Annual online at http: //www.f.bg.ac.rs/bpa/index.html

Printed by
Službeni glasnik, Belgrade

This issue is financially supported by
Ministry of Education, Science and Technological Development of the Republic of Serbia

The statement on publication ethics can be found at the journal website
http://www.f.bg.ac.rs/bpa

BELGRADE PHILOSOPHICAL ANNUAL 33/2020

SCEPTICISM

SCEPTICISM

Annalisa Coliva
University of California, Irvine
a.coliva@uci.edu

Original Scientific Paper
UDC 165.72

SKEPTICISM UNHINGED*

Abstract: *The paper explores the anti-skeptical bearing of the kind of hinge epistemology I have developed in* Extended Rationality. A Hinge Epistemology. *It focuses, in particular, on the moderate account of perceptual justification, the constitutive response put forward against Humean skepticism, and the denial of the unconditional validity of the Closure Principle, which is key in rebutting Cartesian skepticism. Along the way, a comparison with Wittgenstein's own views in* On Certainty *and with the positions held by other prominent hinge epistemologists, particularly Moyal-Sharrock, Pritchard and Wright, is provided.*

Keywords: *Hinges, perceptual justification, constitutivism, extended rationality, Humean skepticism, Cartesian skepticism, Closure principle, Transmission failure.*

1. Introduction

Wittgenstein's remarks in *On Certainty* are at the roots of the ever-accelerating trend in contemporary epistemology, which goes under the label of "hinge epistemology". Key to this trend is the acknowledgement of the philosophical significance of the idea that justification and knowledge of empirical propositions always take place within a system of assumptions, or "hinges". Such hinges, Wittgenstein maintains, are the scaffolding of our thoughts (OC 211), the foundations of our research and action, (OC 87–8), and of our doubt and enquiry (OC 151). Here are the passages where Wittgenstein introduces them:

> All testing, all confirmation and disconfirmation of a hypothesis takes place already within a system [of assumptions]. And this system is not a more or less arbitrary and doubtful point of departure for all our arguments; no, it belongs to the essence of what we call an argument.
>
> That is to say, the *questions* that we raise and our *doubts* depend on the fact that some propositions are exempt from doubt, are as it were like hinges on which those turn.
>
> That is to say, it belongs to the logic of our scientific investigations that certain things are *in deed* not doubted.

* I would like to thank an anonymous referee for helpful comments on the penultimate version of this paper.

DOI: 10.5937/BPA2033007C

> But it isn't that the situation is like this: We just *can't* investigate everything, and for that reason we are forced to rest content with assumption. If I want the door to turn, the hinges must stay put.
> (OC 105, 341–343)

In this paper I revisit the main anti-skeptical thrust of the kind of hinge epistemology I have been developing since my *Extended Rationality. A Hinge Epistemology* (Coliva 2015). In doing so, I move away from the letter, if not the spirit of Wittgenstein's *On Certainty*, to claim that propositions like "There is an external world". "I am not a BIV", etc. play a rule-like role, while remaining truth-apt. Furthermore, I maintain that they are constitutive of epistemic rationality and therefore rational, even though unjustifiable. On this extended sense of rationality, understood as comprising both justified beliefs and those assumptions which make the acquisition of justification possible, hinges turn out to be rational and, thanks to them, knowledge of large swaths of reality possible. Thus, the extended rationality view allows us to unhinge skepticism, both in its Cartesian and Humean form.

2. Moderatism and Humean Skepticism

What does it mean to say that all investigations take place within a system of assumptions? Think of "A goal has just been scored." We take the experience of seeing a ball roll between two poles to justify that proposition only thanks to already taking for granted that a football match is being played. For that experience could be just the same if a different game were being played, such that a ball rolling between those poles would not constitute scoring a goal. If so, however, a different proposition (or set thereof) would be justified; for instance, that an own goal has just been scored. This idea can be extended to many different cases. One key move consists in noticing that this insight can actually be brought to bear on the main assumption challenged by (Humean) scepticism;[2] namely, "there are physical objects", understood as mind-independent, continuously existing entities. Consider, for instance, a hand-like experience: just by itself it could equally justify "Here is a hand", "I am hallucinating having a hand"; "I am a BIV (a brain in a vat) who is having

[2] Some Wittgenstein scholars may dispute the legitimacy of this move from an exegetical point of view, by appealing to OC 35 where Wittgenstein declares "there are physical objects" nonsense. However, in Coliva (2010, Ch. 3) I have maintained that, from an exegetical point of view, Wittgenstein is contesting the philosophical use of that sentence, as if there could be a legitimate ontological dispute between realists and idealists and the former, like G. E. Moore, could object to the latter by insisting on that truth. He is not objecting at all, however, to its use as a "piece of terminological instruction", to remind everyone that the category of physical objects belongs to our conceptual scheme (see OC 36). This, in my view, as we will see in sect. 3, can actually be coupled with the idea that "there are physical objects" is true, at least in a minimal sense and is, after all, a proposition, which has a rule-like role, rather than an empirical one.

a hand-like experience", and so on. Hence, taking that experience to partly justify "Here is a hand", rather than any of the other propositions compatible with that very experience, depends on already taking for granted that we are interacting with a world populated by physical objects, that our sense organs mostly work correctly (and, possibly, some other propositions, for example "I am cognitively lucid and not a victim of massive perceptual and cognitive deception"). Hence, we can take our perceptual experience as bearing on the question of what reality is like, i.e. of whether there is in fact a hand in front of us, only by taking for granted that there are physical objects with which we are causally interacting. If we doubted that there were, we could no longer consider that experience as being evidentially significant for that specific enquiry, since we could no longer take for granted that that experience is formed in response to the presence of a mind-independent physical object. Rather, it would then be compatible with alternative hypotheses, such that there are only collections of sense-data for instance. Thus, if we did not accept a hinge like "There are physical objects", it would not be rational for us to rule the alternative sense-data hypothesis out. Hence, to be rational, we should also reinterpret all specific beliefs as being about collections of sense-data, and not as being about specific physical objects *qua* mind-independent entities.

Notice, moreover, that the general propositions I claim must be assumed in order for our experiences to bear legitimately onto other propositions about mid-size objects in our environment, so that the latter are justified, are not needed to give us an indefeasible justification for these more specific empirical propositions. *Ceteris paribus* – that is, given those very assumptions and experiences – we could still be facing papier-mâché hands, for instance. What we need those assumptions for is to be able to overcome what one might call our "cognitive locality" – that is, the representations given to us through perception. Thus, we need those assumptions in order justifiably to go beyond our experiences and bring them to bear on a universe populated by physical objects, whose precise identity and properties can, of course, still escape us in certain circumstances. To be more precise: if a certain kind of evidence *e*, like a perceptual experience, is compatible with mutually incompatible kinds of propositions, namely propositions about mid-size physical objects (P) or about BIVs being stimulated so as to have those experiences, say, absent any causal interaction with the relevant physical objects (Q), in order for *e* to accrue to a justification for propositions of kind P rather than Q, some extra condition has to be met. It is only in this way that we will have a justification for propositions of kind P and will be within our rights in taking a given experience, which is a mind-*dependent* kind of evidence, to bear on propositions about mind-*independent* objects.

Hence, a key claim in *Extended Rationality* is that perceptual justification can take place only thanks to a system of very general assumptions, such as

"There is an external world" (or "There are physical objects"), "My sense organs work mostly reliably", "I am not a victim of massive perceptual and cognitive deception", and so on. A problem as old as the very history of epistemology – epitomized by "Agrippa's trilemma" – concerns the epistemic status of these assumptions. In the quest for justification, each horn of this trilemma is thought to be problematical: either we end up providing circular justifications; or we embark on an infinite regress; or else, we stop with unjustifiable and therefore a-rational and arbitrary assumptions.

Suppose we hold that each assumption, in its turn, needs to be warranted, in order for it to generate perceptual justification, together with the appropriate kind of experience. For, one may think, it is only if these assumptions are justified that our ordinary empirical beliefs will rest on secure grounds and will therefore be justified. Consider the football case: it is only if I am independently justified in believing that a football match is being played that my experience of seeing a ball roll between two poles provides a justification for "A goal has just been scored." I think that in this case there is no dispute. Why not? Because it is indeed very easy to see how that assumption can be *independently* justified, for instance: I know that I paid for a ticket to the football match between teams A and B in the stadium where I am now sitting, watching the game; or, I know that every Sunday a football match is played in the stadium where I am, roughly at this time, and that today is Sunday; or else, if I am watching the match on television, I know that it has been advertised as the football match between the two teams; or that commentators keep repeating that this is a crucial football match, or saying that the team that prevails will win the World Cup, and I know that the World Cup is a football tournament; and so on.

Yet, as soon as we move away from the football example, things become much more complicated, for an independent justification for the relevant background assumptions is impossible to attain. Consider a historical case, like Napoleon's victory at Austerlitz, or the very general proposition that the Earth has existed for a very long time before our birth (see OC 183). One might think that the latter proposition is justified by a lot of our specific historical beliefs based, in their turn, on testimonies, both personal and documentary, often recorded in academic texts. However, those testimonies and documents could be just the same and yet have appeared and been recorded in academic books only a few minutes back. Therefore, clearly, it is not to be expected that a justification for such a general proposition could be obtained by inferring to it starting with premises that are justified just as long as that very proposition is taken for granted. That kind of justification would ultimately be circular and it would be no justification at all.

Nor is it to be expected that justification for it could ensue from coherence between it and our further beliefs. Justifications are epistemic goods – to put it in general terms – that should speak to the truth of what they are supposed to justify. Yet, starting with the same evidence – apparent testimonies, documents and academic records – we could just as well produce a different and yet entirely coherent system of propositions. In that system the general assumption is that the Earth has just been created replete with everything we find in it and the corresponding specific empirical propositions are like "It looks as if Napoleon won at Austerlitz about three centuries ago." Nothing makes the first system of beliefs more likely to be true than the second one. If we deem otherwise it is either because we are more used to it and therefore think that it is epistemically kosher; or else it is because we consider its specific beliefs justified and think that this, in turn, gives us a justification for its basic presuppositions. However, in the former case, we would conflate our willingness to endorse a given system of beliefs with proof of its truth. In the latter case, in contrast, we would try to provide a circular justification for its basic assumptions, starting from beliefs that are justified only insofar as those very assumptions are taken for granted.

Another possibility is to think that we have a priori justification for "The Earth has existed for a very long time." Where would that justification come from, though? Intuition is an appealing answer, but only shortly, because one then faces the problem of explaining its nature and workings. This remains one of the philosophically most arduous tasks.[3] Perhaps we have some kind of a priori yet inferential justification, coming from reflection on the very meaning of the terms involved. Notice, however, that this would immediately be hostage to the particular theory of meaning we are prepared to subscribe to. For it is only by relying on inferential-role semantics, which may take either a holistic or a molecularist form, that we can sensibly claim that, for instance, it is constitutive of the meaning of "Earth" that it has existed for a very long time.[4] Yet, a direct referentialist could simply say that "Earth" refers to the planet we are all living on now, whether it has existed for a very long time or only for five minutes, and that this is the meaning of "Earth."

Faced with this kind of difficulty – to repeat, distrust in justifications for general assumptions, stemming from specific beliefs that would be justified only by already taking them for granted; as well as in coherence theories of justification, and mistrust in intuition and in inferential a priori justifications stemming from meaning-constitutive considerations – recent years have seen the emergence of yet another proposal, which belongs to the a priori camp broadly construed. This proposal provides for non-evidential warrants, called "entitlements", for very general background presuppositions, such as, "The Earth has existed for a very long time." Entitlements however, at least in the way they are currently thought of,[5] are not meant to speak to the truth of these propositions. Yet, if this is the case, it is very hard to see how entitlements could be genuine epistemic warrants for them, since

[3] I discuss some contemporary attempts in Coliva (2015, Ch. 2).

[4] Molecularist semantics identify some core inferences as constitutive of concepts, whereas holistic ones take all inferences licensed by a given concept to be constitutive of it.

[5] Cf. Wright (2004), examined in Coliva (2015, Ch. 2 and 4).

they are neither evidential warrants nor guides to the truth of the relevant propositions, capable of providing a viable solution to the original problem they were meant to address; namely, the problem of how these general assumptions could actually be epistemically justified.

Similar considerations to the ones just rehearsed for "The Earth has existed for a very long time" could be made for "There is an external world", "My sense organs work mostly reliably" and "I am not a victim of massive perceptual and cognitive deception", which, arguably, are the presuppositions thanks to which our sensory experiences can be taken (defeasibly) to justify our beliefs about specific mid-size objects in our environment. If this were the situation, since we can provide neither immediate nor mediate justifications for these propositions, it would seem that the skeptical outcome would ensue. That is to say, it would seem that the only plausible alternative would be to hold that these are just a-rational assumptions and that, even if we think we are justified in believing ordinary empirical propositions, we are not.

I think that in broad outline this is the path that (save for considerations regarding coherence and entitlements) led Hume to his skepticism. However, it is again Hume who, to my mind, offered the first seeds to try to escape it, as paradoxical as that might seem. These seeds were developed much later on, in a different direction, by Wittgenstein in *On Certainty*, as I think Peter Strawson was the first to recognize in his *Scepticism and Naturalism. Some Varieties* (1985).

According to Hume, we cannot help believing that there is an external world, so that our sensory experiences are constantly brought to bear on a world populated by mid-size objects that are taken to exist independently of our minds, even when they are not directly perceived by us. For Hume, it is part of our psychological constitution that we cannot but form beliefs and devise actions accordingly. That is the way we live. That is the human condition; but notice that, for him, the human condition is the Humean condition of being forced by nature to follow certain forms of psychological and practical conduct that fall outside rational sanction. Rationally, however, we have to recognize that our most basic beliefs are not justified and neither are our more specific empirical beliefs based on perceptual evidence.[6]

[6] This is not universally accepted by Hume scholars. Constantine Sandis ("Hume as a hinge epistemologist", paper presented at the Second Hinge Epistemology Conference, Paris July 1–2 2019), for instance, contests this and claims that Hume held that ordinary empirical beliefs are justified. He also thinks that for Hume there might be a sense in which even the general assumption that there are physical objects may be justified. This would turn Hume in an anti-skeptic philosopher. I am not a Hume scholar and I am not in a position to challenge this interpretation on a textual basis. In the following, I will be engaging with a kind of skepticism, inspired by at least some remarks in Hume and by some of their more traditional interpretations whereby we are not epistemically justified in holding that there is an external world and, for that reason, that assumption is not epistemically rational. For an opposite interpretation which, however, aims to block the unwanted consequence that we are blameworhty for having that belief, see Avnur 2015.

Wittgenstein, in contrast, put forward the view that even though we cannot justify these very general assumptions (or indeed, in his view, even more specific ones which are equally necessary for certain sorts of empirical practices and inquiries), we cannot help but make them thanks to our upbringing within a community that shares language and certain epistemic practices or, more generally, a *form of life*. However, his idea was that the human condition is not the Humean one at bottom. Hence, there is no unbridgeable gulf between what reflection imposes on us and what we cannot help doing, given our psychological and more culturally determined nature. That is, between the recognition that all justification for ordinary empirical propositions rests on unwarrantable assumptions, and going on living *as if*, thanks to those assumptions, our ordinary beliefs were justified. Thus, the human condition, in Wittgenstein's view, is one in which we simply have to recognize that whatever degree of justification we possess for our ordinary empirical beliefs, and that we *do* in fact possess, it takes place within a system of assumptions, which are neither justified nor justifiable.[7] Therefore, according to Wittgenstein, the human condition is importantly different from the Humean one, primarily because justifications are indeed possible, at least for ordinary empirical propositions, but only thanks to a system of unwarrantable assumptions.

This is the kind of picture about the structure of perceptual justification that I present and defend in some detail in *Extended Rationality*. It can been seen, among other things, as the attempt to make good one of the horns of Agrippa's alleged trilemma. According to that trilemma, no justification is ever possible because there are no immediately justified propositions, which can serve as the basis for all others,[8] and so the quest for justification ultimately leads to an infinite regress; nor can justification be produced in a circular way[9] or by resting on unjustified assumptions. The view I present

Once the moderate architecture of perceptual justification is endorsed (see below), the possible consequence that also ordinary empirical beliefs may not be epistemically justified, if that general assumption is not, would be blocked. For such a justification is not needed in order to have perceptual justifications for ordinary empirical beliefs. Also Cartesian skepticism would be blocked since Closure would not hold and hence, from the fact that we have no epistemic justification for "I am not a BIV" it would not follow that we would have none for holding "Here is my hand" based on one's current visual experience (see sect. 4).

[7] Recall the citation from OC 105. See also OC 359 and 559.

[8] The attempt to build on that horn of the trilemma would lead to foundationalism. Both Pryor's (2004) and Wright's (2004) views can be seen as different ways of defending it. In Pryor we have immediate justification for ordinary empirical beliefs, thanks to perception and in the absence of defeaters from them, we then derive a justification for very general propositions such as "There is an external world." In Wright, in contrast, we have an entitlement – that is, a non-evidential justification – directly for those very general assumptions and, thanks to it and to an appropriate course of experience, a justification for ordinary empirical beliefs.

[9] The attempt to build on this horn of the trilemma would lead to various forms of coherentism, whose fault is that they could give rise to maximally coherent, yet

and defend in *Extended Rationality* agrees that, when it comes to very general propositions, such as “There is an external world”, we cannot immediately justify them (whatever that might mean as we have briefly explored above). Nor can we justify them in a circular way by dint of beliefs that are justified only as long as these assumptions are already taken for granted. However, it aims to vindicate the idea that even if these assumptions are neither warranted nor warrantable, they can serve to produce a justification for ordinary empirical propositions, once we enjoy the appropriate kinds of experience.

I call this view the “moderate” conception of perceptual warrant, as it can be seen as lying in between the so-called “liberal” view, proposed in recent years by Jim Pryor (2004), and the “conservative” view defended mostly by Crispin Wright (2004). In outline, the first one corresponds to the intuition that perceptual justification is immediate. As long as there are no defeaters, our perceptual experiences give us an immediate justification for ordinary empirical propositions such as “Here is a hand.” In contrast, the conservative view has it that a warrant for ordinary empirical propositions can be had only if certain general assumptions are independently justified.

The idea I defend is that, contrary to the liberal position, we need assumptions to overcome our cognitive locality – that is, if we want to form defeasibly justified beliefs about specific physical objects in our environment based on our experiences. Yet, contrary to the conservative view, these assumptions need not be warranted, for, in fact, they cannot.[10] For present purposes, let me stress that the moderate architecture of the structure of perceptual warrant just says that a specific empirical proposition P, for instance “Here is a hand,” is perceptually justified iff one has the relevant kind of experience, such as a hand-like one, and the background assumption that there is an external world is in place (possibly together with other ones such as, “My sense organs are mostly working reliably,” “I am not the victim of massive perceptual and cognitive deception,” and so on), while there are no defeaters. Since this definition is compatible with various ways of thinking of the status of such an assumption, which range from an externalist positing that the world is just like that, to making it the content of a doxastic attitude of a specific subject, moderatism is introduced as a *family* of possible views and not as just one single position. Yet, they would all be different species of the same genus – the genus I call, following the Wittgensteinian metaphor,

incompatible systems, among which we could make no epistemically sound choice. That is to say, we would have no means to determine which one is the correct one. Or else, we would have to produce locally circular justifications, that is justifications for general propositions like “There is an external world” based on specific propositions, such as “Here is a hand,” which, in their turn, are justified only insofar as we take for granted those very assumptions. In Coliva (2015, Ch. 3) I argue at length why such circular justifications would be no justifications at all.

[10] For a detailed discussion of the reasons why these assumptions cannot be warranted, see Coliva 2015, Ch. 2.

"hinge epistemology" – because they all hold that perceptual justifications take place "within a system" (OC 105) of assumptions, that is of propositions that lie outside the route of inquiry and that make justifications within inquiry possible in the first place.

Furthermore, these species of the same genus are compatible with different accounts of how we should think of the content of perceptual experience for the latter partially to constitute a justification for ordinary empirical beliefs. Indeed, it is my conviction that the moderate architecture of the structure of perceptual warrants has been endorsed, in one version or another, by many different philosophers, like naturalists of a Humean persuasion (provided they were prepared to forsake Hume's skeptical attitude at the reflective level), Wittgenstein in *On Certainty*, and naturalists inspired by him, like Strawson. In addition, pragmatists would turn out to be moderates, in my view, for they would give a pragmatic and therefore non-epistemic justification for hinges. Furthermore, those externalists about the nature of perceptual justification who are prepared to recognize a role for general assumptions, like Ernest Sosa in recent writings, would count as moderates too.[11]

3. Humean Skepticism Unhinged

One could then be tempted to think that moderatism inspired by some of Wittgenstein's considerations in *On Certainty* would offer only momentary relief from skeptical worries for – the train of thought would go – it would remain that if those assumptions are not justifiable, then they may well turn out to be false. Hence, nothing guarantees that our epistemic practices rest on a secure basis. Yet this, according to Wittgenstein, would be right only if it made sense to call those assumptions into question. That is to say, it would be right only if those assumptions were in the business of epistemic appraisal at all. That is, if it made sense to apply to them the very categories of truth and falsity and, more importantly and less contentiously, the very categories of being justified/unjustified, or even known or unknown. But the main thrust of *On Certainty*, at least according to the kind of, so-called, "framework reading" I myself (and others) have put forward,[12] is that those very general assumptions are not like empirical propositions of a more general kind, *contra* what G. E. Moore held. Rather, they are similar to *rules*; that is to say, they play a normative role and, like rules, are not subject to truth or falsity, nor to assessment in terms of justification or lack thereof.[13] Compare with "Stop at traffic lights when red." It is intuitive to think that it does not correspond to a pre-ordinate fact, and so that it does not make sense to think of it as either

[11] For a more detailed discussion of why moderates are legion, see Coliva (2015, Ch. 1).

[12] See Coliva (2010). See also McGinn (1989), Moyal-Sharrock (2004), Wright (1989).

[13] The details of such a reading are developed differently by Moyal-Sharrock (2004) and Coliva (2010) and (2013a, b), but the main thrust is the same.

true or false in any robust sense of that word. Nor, for the same reason, would it make sense to think of it as either justified – that is as supported by further facts or experiences – or as unjustified – as disconfirmed by further facts and experiences. If "There is an external world" or "There are physical objects" are relevantly similar to "Stop at traffic lights when red" then the skeptical worry that, being unjustified, they might turn out to be false would be off target and due to a mistaken conception of the very nature of those "hinges."

I myself embrace the Wittgensteinian view that justifications for ordinary empirical propositions are possible thanks to a system of assumptions – that is, owing to a system of more general propositions, which, as such, cannot be justified. However, I do not wish to endorse the view that these assumptions are rules, devoid of any descriptive content, if that is indeed Wittgenstein's considered view on the topic.[14] Yet, if this is a sensible avenue to explore as far as the status of "There is an external world" is concerned, it actually seems to be in danger of re-opening the door to the skeptical challenge. For now, how would one block the conclusion that this is merely an assumption we make which, however, is actually unjustified and therefore not rational, exactly as a skeptic would hold? This is the challenge the extended rationality view I present and defend in Coliva (2015) is meant to face. Accordingly, if either empirical, or coherentist, or a priori kinds of warrant for "There is an external world" are unattainable and entitlements are only putative epistemic warrants, we may defend the epistemic legitimacy of that hinge by claiming that, even though unwarranted, it is in fact *constitutive* of epistemic rationality itself. Just as both rules and moves are part of any game so, I argue, both constitutive assumptions and perceptual justifications, which are possible thanks to them, are part of epistemic rationality. To ban constitutive assumptions from epistemic rationality simply because they are not warranted (as they cannot be), like skeptics do, is due to too narrow and unmotivated a conception of the extent of epistemic rationality. Namely, one that confines it to perceptually justified beliefs only. In contrast, epistemic rationality extends beyond the latter to those very assumptions that make it possible to produce ordinary perceptual justifications and to have the kind of practice (or

[14] As always, with Wittgenstein, things are not entirely clear. My own reading, presented in Coliva (2010) and further developed in Coliva (2013a, b), is that it is possible to distinguish between the content and the role of a sentence. Hence, Wittgenstein's hinge propositions would indeed be propositions, which, however, have been removed from doubt and inquiry. Therefore, they would play a normative role, while retaining a descriptive content. Think of the draws that serve as instructions to assemble pieces of furniture: they are, at once, pictures, and therefore have a descriptive content, as well as sets of instructions, or rules, regarding how to put pieces together. Indeed, in OC 318–320 Wittgenstein himself points out that the distinction between empirical propositions and norms is not a clear-cut one and that the very concept of proposition is a family resemblance one. I take this to mean that hinges, even though possibly neither true nor false and more akin to rules, would still be regarded by him as propositions. Moyal-Sharrock (2004), in contrast, thinks that they would not.

method) of forming, assessing, and withdrawing from empirical beliefs on the basis of perceptual evidence, which is itself constitutive of our very notion of epistemic rationality. If so, it turns out that we are actually *mandated by epistemic rationality itself* to assume "There is an external world". However, a rational mandate is not an epistemic warrant – namely, an epistemic good that speaks to the truth of what it is meant to warrant. Humean skeptics are right to think that we have no such warrant for "There is an external world" or "There are physical objects". However, they are wrong to think that, for that very reason, these propositions fall outside the scope of epistemic rationality and that, for that very reason, we cannot have perceptual warrants for our ordinary empirical beliefs.

One may then worry that even if "There is an external world" and "There are physical objects" are epistemically rationally mandated, they might still be false and hence that the extended rationality view has done little to counter the skeptical challenge. It is here, however, that I think we should ponder more on the semantic assessment of that proposition and, in particular, on what it means to say that it is true. As is familiar, there are at least two broad notions of truth: a realist, mind-independent one, and an anti-realist, evidence-dependent one. According to the former, no matter what we think or judge, a proposition is true (or false) in its own right, because it corresponds (or fails to correspond) to some pre-ordinate, mind-independent fact. What is seldom noticed is that it is only on such a conception of truth that broadly Cartesian skeptical concerns with respect to "There is an external world" make sense. For it is only on such a realist conception of truth that, despite the fact that nothing we take ourselves to know speaks against that proposition, it might still be false. Yet, in order to counter the skeptical challenge we cannot revert to a familiar anti-realist, evidence-dependent view of truth either. For, it is a tenet of hinge epistemology that all specific empirical truths are known (or justifiably believed) only by taking that very general proposition for granted. Yet, as remarked, I do not wish to endorse the (allegedly) Wittgensteinian view, according to which hinges are not truth-evaluable at all.

It is at this junction that I propose to endorse a minimalist view of truth with respect to them. Accordingly, they satisfy certain platitudes: they may enter the disquotational schema, and allow for meaningful negation and embedding in suppositional contexts. So much suffices for predicating their truth. However, the kind of truth-property they enjoy is neither of a robustly realist, correspondentist kind, nor of a familiar anti-realist, evidentialist kind. For, to repeat, on the one hand, the realist conception of truth is the most powerful ally of the kind of skepticism that finds its impetus in the intuition that despite all the evidence we have in favour of any given empirical proposition, and even about hinge assumptions, they could nevertheless all be false. On the other, no evidentialist account of truth could confirm hinges for those hinges are needed in order to have justification in the first place. Hence, all there is to hinges' truth is what is made explicit through the platitudes we have just rehearsed. In particular, they are not true because they correspond to a mind-independent reality. Rather, they themselves are conditions of representation of entire swaths of "reality". For instance, those concerning specific mind-independent physical objects (other minds, the past, the uniformity of nature, etc.). In addition, in a Wittgensteinian (indeed Kantian) spirit, when we are dealing with conditions of possibility of representation, they ultimately depend on us. That is, they depend on the fact that we have a conceptual scheme that countenances mind-independent objects. Hence, hinges like "There is an external world" are true, in a minimal sense, because they belong to our conceptual scheme and make it possible for us to represent specific mind-independent object and to acquire justification and knowledge of ordinary empirical propositions. To suppose that despite all we take ourselves to know hinges such as "There is an external world" and "There are physical objects" might after all be false would depend on still being in the grip of a realist conception of truth, which one would be entitled to endorse in this connection only if there were no other options.[15] In short, it would be the result of a kind of "nostalgia" for a realist conception of truth, which results in our inability to let it go, as it were. Such a realist conception of truth is at the root of many of our philosophical puzzles and anxieties, according to Wittgenstein and several other "anti-representationalists" (a deceptive label, which suggests the impossibility of representing anything, while in fact the idea would be that representations are a function of conceptual schemes that are not themselves reflections of a predeterminate reality). It is in connection with this kind of feeling and attitude toward the realist conception of truth that therapy, in the form of acting on our will, is needed, according to Wittgenstein. For initially a picture of truth holds us captive. Through philosophical reflection, we recognize that much and see how it could be thought of differently and yet cannot help going back to it. It is here that our will has to become stronger and make us finally turn our backs to that picture. Temptations may still occur along the road of our thinking about reality. Yet, each time we will have to fight them. In this sense, philosophy is a constant battle against the bewitchment of our intelligence, as Wittgenstein points out in *Philosophical Investigations* (1953, 109).

Hence, the final and specific version of hinge epistemology I endorse has it that thanks to (minimally) true and epistemically rationally mandated assumptions such as "There is an external world," or "There are physical objects" (and possibly other ones), together with appropriate courses of experience, we can and do have perceptual justifications for ordinary empirical beliefs such as "Here is a hand". However, to repeat, this is the species of the hinge epistemology genus I endorse. It is not the only possible one; even though I

[15] Or else, if we were not aware of those options or had decisive arguments against them. This does not seem sustainable with respect to minimalist (or deflationary) accounts of truth. For further discussion of hinges' minimalist truth, see Coliva 2018a and 2019.

am convinced it is the one that has the best prospects of success, because it speaks to the skeptical challenge, albeit by developing an indirect response to it – that is, not contradicting the skeptic by providing ordinary epistemic warrants for "There is an external world". Rather, the extended rationality view is a response that shows that the skeptical quest is somehow illegitimate when it comes to very general propositions like "There is an external world," as it asks for justifications that cannot be obtained and it is based on too narrow and unmotivated a conception of epistemic rationality and on a realist conception of truth that are by no means the only possible option.

4. Cartesian Skepticism Unhinged

A number of important consequences follow from such a general picture. For example, it follows that the Principle of Closure for justification under known entailment is not unconditionally valid.[16] For "Here is my hand" entails "There is an external world". Yet, while we can justifiably believe the former (and the entailment), we cannot justifiably believe the latter. Still, in my view, this does not lead to any "abominable conjunction"[17] of the kind "I justifiably believe there is my hand here, but I don't justifiably believe there is an external world" *sic et simpliciter*. Rather, the kind of conjunction we get, once the extended rationality view is endorsed, is "I justifiably believe that here is my hand, although I don't justifiably believe there is an external world, I am epistemically rationally mandated to assume there is." As Harman and Sherman (2011) have pointed out, the threat of abominable conjunctions depends on not paying enough attention to the possibility of there being, in the vicinity of the repudiated notions (i.e. "epistemic justification for beliefs"), subtler ones, such as, in our case, the notion of "rationally mandated assumptions".[18]

Furthermore, we have to recognize that beside the kind of warrant transmission-failure principle originally presented by Wright,[19] according to which an argument cannot generate (or enhance one's previous) warrant for a conclusion if, and only if, the *warrantedness* of its premises depends on already possessing a warrant for its conclusion, there is another kind

[16] The precise rendition of the Principle of Closure is a matter of contention. I take it to consist in the following: if P is justified or known, and it is justifiably believed or known that P entails Q, then Q is justified or known too. My reading of the Principle of Closure is therefore such to impose merely a consistency requirement between the epistemic status of the propositions figuring in the entailment. It does not see Closure as a principle capable of generating or enhancing the epistemic status of those propositions. The latter, by contrast, is a property of the Principle of Transmission of epistemic goods such as justification (or warrant) and knowledge.

[17] Famously, this is Keith DeRose's (1995) phrase.

[18] There will presently be more on the key notion of assumption.

[19] Cf. Wright (1985, 2004).

of warrant transmission-failure principle, which is indeed at issue in the kinds of cases that are of most interest to philosophers.[20] Namely, the one according to which an argument cannot generate (or enhance one's previous) warrant for a conclusion if, and only if, the warrantedness of its premises depends simply on the very *assumption* of its conclusion. It is for this reason that also on the moderate architecture of perceptual warrant, and not only on its conservative counterpart, Moore's argument ("Here is a hand. If there is a hand here, there is an external world. Therefore, there is an external world") is not cogent. Furthermore, it is because of this kind of transmission-failure that bootstrapping arguments designed to produce warrants for very general beliefs, such as "My sense organs are mostly working correctly," out of specific perceptual beliefs justified by means of occurrent perceptions, would not be cogent either.

Denying the unconditional validity of Closure for principled reasons – that is, because of the moderate account of perceptual justification and the latter kind of transmission failure – is a key move to block Cartesian skepticism. For, as is customary nowadays, that form of skepticism can be seen as depending on two crucial ideas. First, that we are not in a position to exclude radically skeptical scenarios, since all our presently available evidence would be compatible with their occurrence. Second, that if we cannot exclude their obtaining, we cannot know (or justifiably believe) ordinary empirical propositions, such as (P) "Here is my hand", based on one's current visual experience. This second conclusion is indeed based on Closure. For, if that principle holds, if one cannot know (or justifiably believe) that one is not a BIV (Q), by contraposition, one cannot know (or justifiably believe) that there is a hand (P) where one seems to see it. Thus, if the Closure Principle does not hold unconditionally, it is indeed possible to know (or justifiably believe) P, even if one cannot know (or justifiably believe) (Q) "I am not a BIV", and Cartesian skepticism is therefore blocked.

Compared with other kinds of hinge epistemology, mine does not claim that not-Q is ultimately unintelligible;[21] nor does it claim that Q is not a proposition or the object of a propositional attitude, such that it could not figure in the entailment or as a possible instance of Closure (or of Transmission).[22]

[20] I am adopting Wright's terminology here and accordingly speaking of warrants rather than justifications. I take the terms to be safely interchangeable in this context.

[21] For such a position in contemporary epistemology, see Schönbaumsfeld 2016. This is also very much in keeping with Wittgenstein's own pronouncements in *On Certainty* against the very intelligibility of the dreaming hypothesis. I discuss them at length in Coliva (2010, Ch. 3). Arguably, Wittgenstein's remarks are also at the origin of Putnam's (1981) brains in a vat argument.

[22] See Moyal-Sharrock (2004) and Pritchard (2016) respectively. Pritchard in my view conflates Closure with Transmission because he thinks that Closure would be a principle, which would allow us to rationally come to believe the consequences of certain premises we already rationally believe. Crucialy, for Pritchard rational belief is belief held for a

To repeat, in my view, “I am not a BIV” (or “I am not the victim of a lucid and sustained dream (or of any other massively cognitive deception”) is a hinge of all our (empirical) inquiries and cannot be independently justified. Rather, it is constitutive of epistemic rationality and, for that reason, it cannot rationally be doubted either. For it is mandated by any rational activity and inquiry into (empirical) reality. Yet, it is truth-apt, albeit in a minimalist sense, and is a proposition, which, as such, can be the object of a propositional attitude and figure in truth-preserving (though non-epistemic generative[23]) entailments. In my view, the kind of attitude we bear to it is not belief, though, if belief is understood as an attitude of holding a proposition true based on reasons and evidence in its favor. That is why I prefer to talk about assuming, rather than believing, in connection with hinges. For assuming is still an attitude of holding a proposition true, which, however, does not have to be mediated by supporting reasons in favour of its contents.[24]

Yet, it should be realized that rejecting the unconditional validity of Closure is not a terrible price to pay. For, after all, Closure, remains valid in ordinary cases. That is to say, in those cases in which the propositions on both sides of the entailment are not hinges. Thus, insisting on failure of Closure as a fatal blow to hinge epistemology, at least of the kind I have been defending, is once again the symptom of a kind of nostalgia for certain pictures or “truths”, which, however, there is no reason to consider sacrosanct, especially when all is being suggested is simply redefining their boundaries.

5. Conclusions

In this paper we have seen how the specific version of hinge epistemology I have been developing since *Extended Rationality* can counter skepticism of both Humean and Cartesian descent. The key move is to realize, in a Wittgensteinian spirit, if not by following the letter of *On Certainty*, that propositions like “There is an external world”, “there are physical objects” and “I am not a BIV” play a rule-like role, as they are constitutive of epistemic rationality and are therefore mandated by epistemic rationality itself. That is, they allow us to represent reality as populated by mind-independent objects and to confidently exercise our cognitive powers to form justified or even

reason. Since, for him, hinges are not the object of any rational belief, let alone one we form through reasoning, and are the object of visceral commitments instead, Closure does not apply to them and is therefore protected by counterexamples. I have discussed Pritchard’s views at length in Coliva 2016, 2018b. For a different characterization of Closure and a discussion of the difference between it and Transmission, see Coliva (2015, Ch. 3), cf. fn. 15.

[23] See fn. 15.

[24] For an extended discussion of assumptions, of how they are manifested in action and can be attributed also to a– or pre-linguistic creatures based on certain forms of behavior, see Coliva (2015, Ch. 1).

knowledgeable beliefs about them. This is compatible with retaining the idea that they are true, albeit in a minimalist sense, and can thus figure in entailments. Still, even if "Here is my hand" entails "There are physical objects" and "I am not a BIV", it does not follow that if we can, and do in fact know the former, we also can and do know the latter. For the Principle of Closure for epistemic operators holds only for ordinary empirical propositions and does so because these very general assumptions cannot in any way be warranted or known. Yet, thanks to the moderate account of perceptual justification, this is in turn compatible with the commonsensical idea that we do in fact have plenty of justified beliefs in and knowledge of ordinary empirical propositions like "Here is my hand". By retaining this large swath of knowledge and by seeing its assumptions as not lying outside epistemic rationality, thanks to constitutivism and an extended view of epistemic rationality, skepticism can actually be unhinged.

References

Avbur, Y. (2015). "Excuses for Hume's Skepticism". *Philosophy and Phenomenological Research* 92/2: 264-306.

Coliva, A. (2010). *Moore and Wittgenstein. Scepticism, Certainty and Common Sense*. London: Palgrave.

Coliva, A. (2013a). "Hinges and Certainty. A Précis of Moore and Wittgenstein. Scepticism, Certainty and Common Sense," *Philosophia. The Philosophical Quarterly of Israel* 41: 1–12.

Coliva, A. (2013b). "Replies," *Philosophia. The Philosophical Quarterly of Israel* 41: 81–96.

Coliva, A. (2015). *Extended Rationality. A Hinge Epistemology*. London: Palgrave.

Coliva, A. (2016). "Review of Duncan Pritchard Epistemic Angst and the Groundlessness of Our Believing," *Notre Dame Philosophical Reviews*. http://ndpr.nd.edu/news/68056-epistemic-angst-radical-skepticism-and-the-groundlessness-of-our-believing/

Coliva, A. (2018a). "What anti-realism about hinges could possibly be". In *Epistemic Realism and Anti-Realism: Approaches to Metaepistemology*. Edited by R. McKenna and C. Kyriacou. London: Palgrave: 267–288.

Coliva, A. (2018b). "Strange bedfellows. On Pritchard disjunctivist hinge epistemology," *Synthese*: 1–11. https://doi.org/10.1007/s11229–018–02046-z

Coliva, A. (2019). "Hinges, radical skepticism, relativism and alethic pluralism". In *Non-Evidentialist Epistemology*. Edited by N. Pedersen and L. Moretti. Leiden: Brill, forthcoming.

DeRose, K. (1995). “How to Solve the Sceptical Paradox,” *Philosophical Review* 104: 1–52.

Harman G. and Sherman B. (2011). “Knowledge and assumptions,” *Philosophical Studies* 156/1: 131–140.

McGinn, M. (1989). *Sense and Certainty: A Dissolution of Scepticism*. Oxford: Blackwell.

Moyal-Sharrock, D. (2004). *Understanding Wittgenstein’s* On Certainty. London: Palgrave.

Pritchard, D. (2016). *Epistemic Angst and the Groundlessness of Our Believing*. Princeton: Princeton University Press.

Pryor, J. (2004). “What’s wrong with Moore’s paradox?,” *Philosophical Issues* 14: 349–378.

Putnam H. (1981). “Brains in a vat”. In *Reason, Truth and History*. Cambridge: Cambridge University Press: 1–21.

Schönbaumsfeld, G. (2016). *The Illusion of Doubt*. Oxford: Oxford University Press.

Strawson, P. (1985). *Scepticism and Naturalism. Some Varieties*. New York: Columbia University Press.

Wittgenstein, L. (1953). *Philosophical Investigations*. Oxford: Blackwell.

Wittgenstein, L. (1969). *On Certainty*. Oxford: Blackwell.

Wright, C. (1985). “Facts and certainty,” *Proceedings of the British Academy* 41: 429–472.

Wright, C. (2004). “Warrant for nothing (and foundations for free?),” *The Aristotelian Society Supplementary Volume* 78: 167–212.

Michael Blome-Tillmann
McGill University

Original Scientific Paper
UDC 165.4

NON-REDUCTIVE SAFETY*

Abstract: *Safety principles in epistemology are often hailed as providing us with an explanation of why we fail to have knowledge in Gettier cases and lottery examples, while at the same time allowing for the fact that we know the negations of sceptical hypotheses. In a recent paper, Sinhababu and Williams have produced an example—the Backward Clock—that is meant to spell trouble for safety accounts of knowledge. I argue that the Backward Clock case is, in fact, unproblematic for the more sophisticated formulations of safety in the literature. However, I then proceed to construct two novel examples that turn out problematic for those formulations—one that provides us with a lottery-style case of safe ignorance and one that is a straightforward case of unsafe knowledge. If these examples succeed, then safety as it is usually conceived in the current debate cannot account for ignorance in all Gettier and lottery-style cases, and neither is it a necessary condition for knowledge. I conclude from these troublesome examples that modal epistemologists ought to embrace a much more simple and non-reductive version of safety, according to which the notion of similarity between possible worlds that determines in which worlds the subject must believe truly is an epistemic notion that cannot be defined or reduced to notions independent of knowledge. The resulting view is shown to also lead to desirable results with respect to lottery cases, certain quantum phenomena, and a puzzling case involving a cautious brain-in-a-vat.*

1. Classical Safety

Since the turn of the century, a number of epistemologists have defended a necessary condition on knowledge that is familiar as the *safety condition*. Safety is meant to provide us with a plausible response to scepticism, by offering us an explanation of how we know both ordinary propositions and the negations of sceptical hypotheses, and thus by delivering a response to sceptical arguments that succeeds without giving up closure. Roughly, according to authors such as Ernest Sosa, Duncan Pritchard, and Timothy Williamson, a subject *S* knows *p* only if *S* could not have easily been wrong with respect to *p*. Even though Sosa has given up on safety in more recent writing, let us begin the discussion with his formulation of the principle, which represents the most straightforward and familiar way to articulate the general idea underlying safety. Here is Sosa's (1999: 146) definition of what I shall call *classical safety*:

[*] I am indebted to Sven Bernecker and Duncan Pritchard for discussion of earlier versions of this paper, and to an anonymous referee for very helpful comments and suggestions.

DOI 10.5937/BPA2033025B

(ES) *S*'s belief that *p* is classically safe $=_{\mathrm{df}}$
[if *S* were to believe *p*, then *p*].

Given (ES), a belief is classically safe iff it could not have been false easily. In terms of possible worlds, (ES) says that one's belief that *p* is classically safe just in case one believes *p* in a nearby world *w*, only if *p* is true in *w*. Sosa (1999) further defends the view that classical safety is a property of knowledge:

(SAFE_C) Necessarily, *S* knows *p* only if:
[if *S* were to believe *p*, then *p*].

The main motivation of (SAFE_C) consists, according to Sosa (1999), in the fact that it accounts neatly for the fact that we lack knowledge in Gettier cases and lottery examples. In such examples, the explanation goes, we fail to know that *p* because our belief that *p* is not classically safe—our belief could have been false easily as there are many nearby $\neg p$-worlds in which we (falsely) believe that *p*.

Consider, for illustration, the following version of a lottery case, inspired by LJ Cohen (1977):

> *The Gatecrasher*:
> The organizers of the local rodeo decide to sue John for gatecrashing their Saturday afternoon event. Their evidence is as follows: John attended the Saturday afternoon event—he was seen and photographed on the main ranks during the rodeo. No tickets were issued at the entrance, so John cannot be expected to prove having bought a ticket with a ticket stub. However, while more than 1,000 people were counted in the seats, only 157 paid for admission. No further evidence is presented in court.

In the Gatecrasher example, the judge is epistemically rather well justified in believing that John gatecrashed, but she crucially does not know that proposition: for all the judge knows, John was one of the 157 honest fee-paying people in attendance. Thus, while the statistical evidence available to the judge can justify her belief that John gatecrashed,[1] it intuitively cannot ground her *knowledge* that he gatecrashed.

Next, note that Sosa's notion of classical safety provides us with an elegant explanation of this *prima facie* surprising datum. According to (SAFE_C), the judge does not know that John gatecrashed because there are numerous nearby possible worlds in which the judge believes falsely that John gatecrashed—namely, precisely those worlds in which John paid the entrance fee instead of climbing the fence. Thus, by requiring that knowledge be free from what many theorists have called *epistemic luck*,[2] (SAFE_C) seems to provide us with an elegant explanation of our intuitions—not only in the Gatecrasher example, but also in other lottery-style examples and Gettier cases.[3]

[1] The probability that John gatecrashed given the judge's evidence is .843.

[2] See, for instance, (Pritchard 2005).

[3] Note also that the judge cannot justly impose liability on the basis of the statistical evidence available to her. See (Blome-Tillmann 2017a) for discussion.

2. Safe Ignorance: The Backward Clock

Attempts in the literature to discredit safety have usually aimed at producing instances of *unsafe knowledge*—that is, counterexamples to (SAFE$_C$) in which a subject intuitively knows that *p* even though her belief that *p* is classically unsafe.[4] In a recent paper, however, Neil Sinhababu and John Williams (2015) have taken a different route—namely, by producing a Gettier-style example of non-knowledge in which safety does not fail. Thus, according to Sinhababu and Williams, safety does not adequately capture the notion of epistemic luck at issue in Gettier examples and cannot explain why we fail to know in Gettier cases. I shall, in this section, briefly describe the example at issue and then show that it does not turn out problematic for some of the more sophisticated formulations of safety in the literature. In Section 3, I shall then offer a different example that in fact achieves Sinhababu and Williams' goal.[5]

Here is Sinhababu and Williams' example:

> *Backward Clock*:
>
> You habitually nap between 4 pm and 5 pm. Your method of ascertaining the time you wake is to look at your clock, one you know has always worked perfectly reliably. Unbeknownst to you, your clock is a special model designed by a cult that regards the hour starting from 4 pm today as cursed, and wants clocks not to run forward during that hour. So your clock is designed to run perfectly reliably backwards during that hour. At 4 pm the hands of the clock jumped to 5 pm, and it has been running reliably backwards since then. This clock is analogue so its hands sweep its face continuously, but it has no second hand so you cannot tell that it is running backwards from a quick glance. Awaking, you look at the clock at exactly 4.30 pm and observe that its hands point to 4.30 pm. Accordingly you form the belief that it is 4.30 pm. (Williams and Sinhababu 2015)

As Sinhababu and Williams point out, Backward Clock is problematic for classical safety, since your belief that it is 4.30 pm is, intuitively, not knowledge despite being classically safe. It is classically safe because, in nearby worlds in which it is not 4.30 pm when you look at the clock, you do not believe that it is 4.30 pm (in those worlds you (falsely) believe that it is 4.31 pm, 4.32 pm, 4.29 pm, etc.). However, intuitively, you could have easily believed falsely in Backward Clock, and that is why your belief, despite being classically safe, is not knowledge: it is, intuitively, true as a matter of mere luck. Consequently,

[4] See, for instance, (Neta and Rohrbaugh 2004).

[5] Adams and Clarke (2016) also criticize Williams and Sinhababu's example, but they do so by pointing out that it is not a counterexample to *sensitivity*. See also fn. 12 for the topic of sensitivity.

Backward Clock is a Gettier-style example in which classical safety cannot account for the absence of knowledge.

Sinhababu and Williams claim that their example is problematic for more sophisticated formulations of safety, too. To back up this claim they consider the following version of the safety principle, which they ascribe to Duncan Pritchard (2012):

(SAFE$_B$) Necessarily, *S*'s knows *p* on basis *B* only if:
[*S* could not have easily formed a false belief on basis *B*].

Sinhababu and Williams argue that (SAFE$_B$)—let us call the principle *Basis Safety*—falls prey to their example, too, and to establish this conclusion they point out that, in Backwards Clock, the basis on which you believe that it is 4.30 pm

> must be that the hands point to 4.30 pm. That you look at the clock is not a sufficient basis for believing that it is 4.30 pm, as this leaves open where the hands are pointing. You need to see that the hands point to 4.30 pm to have grounds for believing that it is 4:30 pm. (Williams and Sinhababu 2015: 53)

While Sinhababu and Williams might be right that their example spells trouble for both (SAFE$_C$) and (SAFE$_B$), there are other versions of safety that clearly avoid the problem. Instead of formulating safety in terms of belief bases, for instance, we might—following some of Pritchard's earlier work—formulate it by appeal to *belief-forming methods*. Call the following principle *Method Safety*:

(SAFE$_M$) Necessarily, *S* knows *p* via method *M* only if:
[*S* could not have easily formed a false belief via *M*].

According to (SAFE$_M$), a belief is safe just in case it was produced by a method that leads to true beliefs not only in the actual, but also in nearby worlds. Interestingly, this condition is not satisfied in Backward Clock. This is so because, in Backward Clock, you formed your belief that it is 4.30 pm by the method of *reading the clock in front of you*. By this method, however, you form, in a nearby world, the false belief that it is 4.31 pm when it in fact is 4.29 pm. In Backward Clock, there are, as a consequence, numerous nearby worlds in which the method that you actually apply leads to false beliefs. Thus, in Backward Clock, your belief that it is 4.30 pm is not method-safe and, given (SAFE$_M$), does not qualify as knowledge. Consequently, (SAFE$_M$) provides us with an effective response to the problem posed by Backward Clock.

As already mentioned, method-safety is inspired by some of Duncan Pritchard's earlier work on safety. In particular, in light of examples including necessary truths and other problem cases, Pritchard (2007a: 292, 2007b: 40, 2009: 34) proposes the following definition of what I shall call *weak safety*:

(SAFE_{W}) Necessarily, *S* knows *p* only if:
[in most near-by possible worlds in which *S* continues to form her belief about the target proposition in the same way as in the actual world, and in all very close near-by possible worlds in which *S* continues to form her belief about the target proposition in the same way as in the actual world, the belief continues to be true.]"[6]

Pritchard's appeal to 'ways of forming a belief' in this passage clearly bears a strong similarity to the notion of belief-forming methods. Thus, both Methods Safety and Weak Safety seem to provide us with an attractive response to the Backward Clock example presented by Sinhababu and Williams.

There are further ways for the safety theorist to respond to Backward Clock that are worth mentioning here. Consider the following version of safety, which is a variant of (SAFE_{B}) and is inspired by Timothy Williamson's (2009b: 325) discussion of safety principles:

($\text{SAFE}_{\text{B*}}$) Necessarily, *S*'s knows *p* on basis *B* only if:
[*S* could not have easily formed a false belief on basis *B* or a similar basis *B**].[7]

It is fairly straightforward to see why ($\text{SAFE}_{\text{B*}}$) is not troubled by Backward Clock. For, in Backward Clock, there are numerous nearby worlds in which you believe a falsehood on a basis that is very similar to your actual belief's basis. For instance, in a nearby world in which you look at the clock at 4.29 pm, you believe, on the basis of looking at the clock and seeing the hands point to 4.31, the falsehood that it is 4.31 pm.

Let me sum up. While Sinhababu and Williams' objection to safety is effective with respect to classical safety as formulated by Ernest Sosa in the early days of modal epistemology, alternative and more sophisticated notions of safety are well-positioned to capture the sense in which our beliefs in Backward Clock are true as a matter of epistemic luck.

3. Safe Ignorance: The Opportunistic Gatecrasher

The general strategy pursued by Sinhababu and Williams—namely, to produce a Gettier-type example of epistemic luck that involves safe ignorance—is interesting, and I shall here attempt to produce an example that is better suited to achieve this goal. The case I have in mind is a variant of the lottery-style example mentioned in Section 1 of this paper. Consider what I shall call the *Opportunistic Gatecrasher*. The details in this example

[6] For a predecessor of this definition, see (Pritchard 2005: 163).

[7] Cp. also (Williamson 2009b: 325): "If in a case α one knows *p* on a basis *b*, then in any case close to α in which one believes a proposition *p** close to *p* on a basis [*b**] close to *b*, *p** is true."

are exactly as in the Gatecrasher example from Section 1, but we fill in the background story as follows:

> *The Opportunistic Gatecrasher*:
>
> John is on his way to the bowling alley to meet his friends, as he does on every Saturday afternoon. John would love to watch the rodeo, but he has not been able to afford the ever-rising entrance fee for many years now. This weekend, however, when he passes by the rodeo on his way to the bowling alley, John sees that a lot of people are climbing the fences. Seizing the opportunity to watch the rodeo for free, John decides to join in and gatecrashes.
>
> Realizing that something is at odds, the organizers of the rodeo decide to sue John for gatecrashing their Saturday afternoon event. Their evidence is as follows: John attended the Saturday afternoon event—he was seen and photographed on the main ranks during the rodeo. No tickets were issued at the entrance, so John cannot be expected to prove having bought a ticket with a ticket stub. However, while more than 1,000 people were counted in the seats, only 157 paid for admission. No further evidence is presented in court.

As in our initial example, the judge is, in the Opportunistic Gatecrasher, rather well justified in believing that John gatecrashed, but she crucially does not *know* that John gatecrashed. Again, the statistical evidence available to her cannot ground knowledge: for all the judge knows, John was one of the honest fee-paying attendees at the rodeo.

What is important about the Opportunistic Gatecrasher, however, is that this time (SAFE$_C$) cannot account for the datum that the judge does not have knowledge. To see this note that the judge's belief that John gatecrashed is classically safe: in all nearby worlds in which the judge believes that John gatecrashed, he in fact gatecrashed. And that is so because, if John had not gatecrashed, he would have gone bowling with his friends and, therefore, could not have been spotted or photographed at the rodeo. Thus, in those nearby worlds in which John does not gatecrash, the judge does not form the (false-in-those-worlds) belief that John gatecrashed. Consequently, the judge's belief that John gatecrashed is classically safe, true, and well-justified. But, crucially, it is not knowledge. The *Opportunistic Gatecrasher* is, therefore, a lottery-style example of problematic epistemic luck that cannot be accounted for by means of classical safety.

One might wonder at this stage whether the alternative and more sophisticated notions of safety discussed in the previous section are better suited to capture the notion of epistemic luck at play in the above example. Consider first Method Safety, reproduced here for convenience:

(SAFE$_M$) Necessarily, S knows p via method M only if:
[S could not have easily formed a false belief via M].

Is the judge's belief in the Opportunistic Gatecrasher method-safe? It is iff there is no nearby world in which the judge formed a false belief via the relevant method. But what is the relevant method? If the relevant method is *believing on the basis of photographic evidence documenting John's presence at the rodeo and the pertinent statistical evidence*, then the method is safe, since the judge does not have photographic evidence of John's presence at the rodeo in nearby worlds in which he did not gatecrash. Remember that, in those worlds where John did not gatecrash, he went bowling instead of attending the rodeo, and so was not photographed at the rodeo in the first place. If, however, the relevant method is *believing that* x *gatecrashed on the basis of photographic evidence of* x*'s presence and the pertinent statistical evidence*, then the judge's belief that John gatecrashed is not method-safe. And that is so because there are many nearby worlds in which a subject other than John is sued for compensation—and, importantly, in some of those worlds the organizers have picked a defendant for their lawsuit who paid the entrance fee and thus did not gatecrash. Since the judge believes, in those nearby worlds and on the basis of the relevant photographic and statistical evidence, that those fee-paying defendants gatecrashed, the belief at issue is not method-safe. Thus, depending on how we specify the belief-forming method at hand, the judge's belief either is or is not method-safe.

What about Williamson's version of safety (SAFE$_{B*}$), also reproduced here?

(SAFE$_{B*}$) Necessarily, *S*'s knows *p* on basis *B* only if:
[*S* could not have easily formed a false belief on basis *B* or a similar basis *B**].

In the Opportunistic Gatecrasher the judge believes that John gatecrashed on the basis of the conjunction of photographic evidence of John's presence and the pertinent statistical evidence. Since John was picked at random, there are nearby worlds in which the judge believes of a fee-paying customer on a very *similar* (or even identical) basis that they gatecrashed. The judge's belief that John gatecrashed is, therefore, not basis*-safe, and (SAFE$_{B*}$) offers us a plausible explanation of why the judge fails to know that John gatecrashed in the Opportunistic Gatecrasher.

While the mentioned principles (SAFE$_{M}$) and (SAFE$_{B*}$) may very well both be able to handle the example as it was presented above, I take it that the case nevertheless illustrates an important point about safety. For we can fairly easily amend the example presented above to the effect that only gatecrashers are being sued by the organizers in nearby worlds. One way to insure this is by stipulating that the real reason for which the organizers sue John is because they do not like him very much, for reasons entirely independent of his propensity to gatecrash the rodeo. Once we have added such stipulations to the effect that there are no nearby worlds in which the organizers sue somebody other than John, the example illustrates the inadequacy of both Method Safety and Basis* Safety.

In summary, we can, with some imagination, construe lottery-style examples in which certain belief-forming methods are only applied in nearby worlds, if they lead to true beliefs, or, in Williamson's terminology, in which certain beliefs are only formed on a particular kind of basis, if the resulting beliefs are true. The Opportunistic Gatecrasher is, therefore, a fairly simple and straightforward lottery-style example of safe ignorance, giving rise to rather strong and clear intuitions.

4. Testimony and Unsafe Knowledge

While the previous section provided a lottery-style example in which safety cannot explain the absence of knowledge, I shall, in this section, produce a case of *unsafe knowledge* and thus aim to show that safety is not necessary for knowledge. While there are several attempts to produce examples of unsafe knowledge in the literature already, the example I propose here is attractive because of its comparative simplicity.[8] Consider the following case of testimonial knowledge, which I borrow from Jennifer Lackey:

> *Chicago Visitor*:
>
> Having just arrived at the train station in Chicago, Morris wishes to obtain directions to the Sears Tower. He looks around, approaches the first adult passerby that he sees, and asks how to get to his desired destination. The passerby, who happens to be a lifelong resident of Chicago and knows the city extraordinarily well, provides Morris with impeccable directions to the Sears Tower by telling him that it is located two blocks east of the train station. Morris unhesitatingly forms the corresponding true belief. (Lackey 2009: 29)

I assume that Morris acquires testimonial knowledge that Sears Tower is two blocks east of the train station in this example. What is important about the example in the present context, however, is that we can amend the case slightly to the effect that the passerby would have told Morris a lie, if he had asked for directions to a different location. Imagine, for instance, that the passerby is an overenthusiastic Democrat, who would have sent Morris in the wrong direction had he asked for directions to the Republican National Convention (RNC).[9] In this scenario, Morris' belief that Sears Tower is two blocks east of the train station is classically safe, and even basis safe, but it is neither method-safe nor basis*-safe. It is classically safe (basis safe) because there is no nearby world in which Morris believes falsely (on the basis of the passerby's testimony) that Sears Tower is two blocks east of the train station. However, it fails to be method-safe because the method of asking a passerby

[8] See, for instance, (Comesaña 2005).

[9] Of course, I do not mean to suggest that Democrats are prone to lying or deceiving as a political tactic.

for directions leads to false beliefs in nearby worlds in which Morris asks for directions to the RNC rather than for directions to Sears Tower. Similarly, Morris' belief fails to be basis*-safe, because, in those nearby worlds in which Morris asks for directions to the RNC, he believes a falsehood on a basis very similar to the basis of his actual belief that Sears Tower is two blocks east of the train station—namely, on the basis of testimony from the mentioned passerby.[10]

5. Non-Reductive Safety

One might wonder whether any of the above examples spells the end of safety accounts of knowledge. The outlook is, to my mind, not quite as bleak. Consider another principle, also defended by Timothy Williamson (2000: 147, 2009a), which defines what I shall call *Simple Safety*:

> (SAFE_S) Necessarily, if one knows *p*, one could not easily have been wrong in a similar case.

Simple Safety offers us, I believe, a straightforward response to both the Opportunistic Gatecrasher and our amended version of the Chicago Visitor. In the Opportunistic Gatecrasher, the judge does not know that John gatecrashed because she could have easily been wrong in similar (even though far away) cases—namely, in precisely those cases in which John paid the entrance fee. Thus, even though the closest worlds in which John pays the entrance fee are overall rather dissimilar to John's actuality, they are nevertheless very similar to John's actuality *in those respects that are relevant for knowledge*. Call this type of similarity *epistemic similarity*. Then, a world *w* can be epistemically similar to a world *w′*, even though *w* is overall rather dissimilar (and thus 'far away') from *w′*. An analogous explanation can be given of our amended version of the Chicago Visitor. In the example, we do not count the case in which Morris asks for directions to the RNC as epistemically similar to the case in which he asks for directions to Sears Tower—one possible explanation being that Sears Tower is not a politically loaded venue, whereas the RNC is, thus potentially rendering a random passerby's testimony unreliable.

Can we give a more informative characterization of epistemic similarity? While it would be desirable to have a reductive account of the notion that allows us to explain in detail how epistemic similarity differs from the intuitive notion of overall resemblance, the demand of an explicit definition or analysis is misplaced. Firstly, it is, as the Gettier literature suggests, rather unlikely that any reductive definition or analysis of knowledge will be

[10] Thanks to Sven Bernecker here, who has drawn my attention to Lackey's example (pc) in the context of safety. See also (Bernecker forthcoming) for critical discussion of safety principles similar to what I have called Basis*-Safety.

resistant to counterexample. Secondly, from a methodological point of view, it is perfectly sufficient to explicate a theoretical term of which one has an intuitive grasp—such as the notion of epistemic similarity—by relating it to other intuitive concepts in our theory—such as the notion of knowledge. (SAFE$_S$) does exactly that: it relates the concepts of knowledge and epistemic similarity to each other in a way that allows us to account for the fact that we do not have knowledge in Gettier examples, lottery cases, and the abovementioned examples—a feat that no other conception of safety has so far achieved.[11]

Do we have an intuitive grasp of the notion of epistemic resemblance? We can determine, for a vast array of examples, whether or not a given case qualifies as epistemically similar to the subject's actuality. As mentioned above, there is an intuitive sense in which worlds in which John pays the entrance fee to the rodeo are relevantly similar to his actuality—despite the fact that they are overall not very close to it. Similarly, it is intuitively plausible that worlds in which John is a brain in a vat do not qualify as epistemically similar to John's actuality. Our grasp of the notion of epistemic similarity closely tracks, in the relevant cases, our intuitions as to whether the subject knows. Thus, in the light of a more holistic or non-reductive approach to epistemological theory building, the demand for an explicit definition or analysis of the notion of safety or epistemic similarity appears unwarranted.[12]

[11] The non-reductive account has further explanatory virtues. It can, for instance, also explain why a reliable eyewitness who saw John climb the fence does not fail to know that John gatecrashed: a reliable eyewitness could *not* have easily been wrong in a similar case.

[12] It is worthwhile noting at this point that the *Opportunistic Gatecrasher* is as problematic for sensitivity accounts of knowledge as it is for classical or reductivist accounts of safety. Here is Nozick's (1981: 179ff.) formulation of sensitivity in terms of the ordinary language counterfactual conditional:

(SEN) Necessarily, *S* knows *p* via method *M* only if:
[if *p* were false, then *S* would not believe *p* via *M*].

Next, note that the following counterfactual conditional is true with respect to the *Opportunistic Gatecrasher*:

(A) If John had not gatecrashed, then the judge would not believe, by inferring from the evidence presented in court, that he gatecrashed.

(A) is true with respect to the *Opportunistic Gatecrasher* because the closest worlds in which John does not gatecrash are worlds in which he goes bowling and does not attend the rodeo. In those worlds John was not singled out by the organizers of the rodeo and, consequently, has never been taken to court. Thus, in the closest worlds in which John does not gatecrash, the judge does not believe falsely that John gatecrashed. The judge's belief that John gatecrashed is, as a consequence, sensitive but it is not knowledge. Sensitivity, accordingly, cannot account for the problematic type of epistemic luck we find in the example and does not provide us with an appropriate response to the challenge of lottery-style examples. Of course, many will consider sensitivity accounts of knowledge problematic for independent reasons. As Nozick (1981: 227–229) himself and many others have pointed out, sensitivity accounts of knowledge entail closure failure. For further criticism of sensitivity see (Blome-Tillmann 2017b).

6. Further Advantages: Lotteries and Cautious Brains-in-Vats

Before concluding, it is worthwhile noting two further potential advantages of the view proposed here.[13] First, consider the case of the *Cautious Brain in a Vat*—a variant of the *New Evil Demon* problem.[14] The problem arises from the observation that, intuitively, a brain in a vat (henceforth 'biv') doesn't know that it has less than three hands, despite the fact that that belief is perfectly safe in the classical sense. In all nearby possible worlds in which the cautious biv forms the belief that it has less than three hands, it is true that it has less than three hands. Method safety ($SAFE_M$) or similar basis safety ($SAFE_{B*}$) cannot solve the problem either, if we think of the cautious biv as a thinker who would never believe that it has two hands, but only that it has less than three hands and similarly for all kinds of other beliefs ('I own less than four bicycles', 'My epistemology class has less than 21 students', 'I see at most one sunrise', and so on). The cautious biv's belief is thus both method safe and basis* safe, since beliefs on similar bases and formed by similar methods in similar worlds are also true. According to Simple Safety, however, the cautious biv's belief that it has less than three hands isn't safe, because worlds in which the cautious biv has three or more hands are, intuitively, epistemically similar to the cautious biv's actuality—despite the fact that they are overall rather dissimilar to the biv's actuality, and thus 'remote'. In contrast, my current belief that I'm not a biv is safe in the way proposed by Simple Safety, because possible worlds in which I am a biv are not only remote but also epistemically dissimilar to actuality.[15]

Second, Simple Safety is plausibly also helpful for dealing with lottery cases. As has been pointed out in the literature,[16] there is a tension between the safety theorists' claim that we don't know that my lottery ticket has lost (because winning the lottery is a very similar case) and her claim that we know all kinds of ordinary propositions about the external world. Consider, for instance, the proposition that the book is on the table. Given classical safety, a strong case can be made that we do not know that the book is on the table because there is a very nearby world in which the book has, due to an extremely unlikely quantum phenomenon, tunnelled through the table the very moment we turned away. In that nearby world we thus believe falsely, by means of the same method and on the very same basis as we actually do, that

[13] I am greatly indebted to an anonymous referee for this journal, who pointed out the following two advantages of Simple Safety.

[14] See (Cohen 1984) for the New Evil Demon Problem.

[15] Another response to the example might be to deny that the cautious biv fails to know that it has less than three hands. I shall, however, not pursue this strategy further here.

[16] See, for instance, (Blome-Tillmann 2014: ch 5.2; Dodd 2012).

the book is still on the table. However, if epistemic similarity is a basic and irreducible notion, then we can easily uphold the idea that the mentioned world is epistemically rather dissimilar to our actuality (which, again, it intuitively is), whereas other safety theorists need an explanation of why winning the lottery is a close world, but one where the book tunnels through the table isn't.

7. Conclusion

Alleged counterexamples to safety principles in epistemology are often complex and convoluted, and usually give rise to diverging intuitions. I have here developed two novel, rather simple and intuitive examples that are problematic for the traditional, reductionist safety accounts familiar from the literature. I have further argued that an explanation of the data emerging from those cases comes at the price of abandoning the idea that safety can be reductively defined in favour of an account of safety in terms of *epistemic similarity*: only non-reductive accounts of safety seem immune to the problems outlined in this paper. Finally, I have argued that non-reductionism is far from problematic or explanatorily idle. To the contrary, once we abandon the reductionist dogma underlying much of 20th century epistemology, we have cleared the way for a fruitful, albeit non-reductive account of safety. Within the framework of a modal epistemology, simple safety can play an important explanatory role with respect to both responses to sceptical arguments and solutions to the Gettier and lottery problems for knowledge.

Appendix – Table of Safety Principles

Table of Safety Principles

S's belief *p* (which is based on basis *B* and formed via method *M*) is safe iff	
Name	Condition
Classical Safety	if *S* were to believe *p*, then *p*
Basis Safety	if *S* were to believe *p* on basis *B*, then *p*
Method Safety	if *S* were to believe some proposition *p** via method *M*, then *p**
Similar Basis Safety	if *S* were to believe a similar *p** on a similar basis *B**, then *p**
Simple Safety	if *S* were to believe *p** in an epistemically similar case, then *p**

References

Adams, Fred and Murray Clarke (2016). „Beat the (Backward) Clock.“ Logos and Episteme **VII**(3): 353–361.

Bernecker, Sven (forthcoming). „Against global method safety.“ Synthese.

Blome-Tillmann, Michael (2014). Knowledge and Presuppositions. Oxford, Oxford University Press.

Blome-Tillmann, Michael (2017a). 'More likely than not' – Knowledge First and the Role of Bare Statistical Evidence in Courts of Law. Knowledge First – Approaches to Epistemology and Mind. A. Carter, E. C. Gordon and B. Jarvis. Oxford, Oxford University Press**:** 278–292.

Blome-Tillmann, Michael (2017b). „Sensitivity Actually.“ Philosophy and Phenomenological Research **94**(3): 606–625.

Cohen, L. Jonathan (1977). The probable and the provable. Oxford, Clarendon Press.

Cohen, Stewart (1984). „Justification and truth.“ Philosophical Studies **46**(3): 279–295.

Comesaña, Juan (2005). „Unsafe Knowledge.“ Synthese **146**(2): 395–404.

Dodd, Dylan (2012). „Safety, Skepticism, and Lotteries.“ Erkenntnis **77**(1): 95–120.

Lackey, Jennifer (2009). „Knowledge and Credit.“ Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition **142**(1): 27–42.

Neta, Ram and G. Rohrbaugh (2004). „Luminosity and the Safety of Knowledge.“ Pacific Philosophical Quarterly **85**(4): 396–406.

Nozick, Robert (1981). Philosophical Explanations. Oxford, OUP.

Pritchard, Duncan (2005). Epistemic Luck. Oxford, Clarendon.

Pritchard, Duncan (2007a). „Anti-luck epistemology.“ Synthese **158**(3): 277–297.

Pritchard, Duncan (2007b). Knowledge, Luck and Lotteries. New Waves in Epistemology. V. Hendricks and D. Pritchard. Basingstoke, Palgrave Macmillan**:** 28–51.

Pritchard, Duncan (2009). „Safety-Based Epistemology: Whither Now?“ Journal of Philosophical Research **34**: 33–45.

Pritchard, Duncan (2012). „Anti-Luck Virtue Epistemology.“ Journal of Philosophy **109**(3): 247–279.

Sosa, Ernest (1999). „How to Defeat Opposition to Moore.“ Philosophical Perspectives – Epistemology **13**: 141–153.

Williams, John N. and Neil Sinhababu (2015). „The Backward Clock, Truth-Tracking, and Safety.“ Journal of Philosophy **112**(1): 46–55.

Williamson, Timothy (2000). Knowledge and Its Limits. Oxford, OUP.

Williamson, Timothy (2009a). „Probability and Danger.“ The Amherst Lecture in Philosophy **4**: 1–35.

Williamson, Timothy (2009b). Replies to Critics. Williamson on Knowledge. P. Greenough and D. Pritchard. Oxford, OUP**:** 279–384.

Tim Kraft
Institut für Philosophie,
Universität Regensburg
tim.kraft@ur.de

Original Scientific Paper
UDC 165.82 Патнам Х.
165.72

BRAINS IN A VAT AND MEMORY: HOW (NOT) TO RESPOND TO PUTNAM'S ARGUMENT

Abstract: *Putnam's argument that we are not brains in a vat has recently seen a resurgence in interest. Although objections to it are legion, an emerging consensus seems to be that even if it successfully refutes one version of the brain in a vat scenario, lifelong envatment, it is powerless against a different one, recent envatment. Although initially appealing, I argue in this paper that this response – merely replacing lifelong envatment by recent envatment – is a bad response to Putnam's argument. Yet there is a different version of the brain in a vat scenario, recent memory-altering envatment, that Putnam's argument does not refute and is also sufficiently radical. The crucial issue turns out to be which epistemic sources sceptical scenarios may attack. I argue that there's no convincing reason for exempting memory from the sceptical attack: Sceptical scenarios must target memory to be sufficiently radical and they can do so without violating any constraint on sceptical scenarios. In the end Putnam's argument doesn't fail because of some 'deep' philosophical mistake, but because it overlooks how flexible and adjustable sceptical scenarios are.*

Keywords: *Putnam, brains in a vat, Cartesian scepticism, constraints on sceptical scenarios, scepticism and content externalism, memory*

1 Putnam's argument: The state of the debate

Putnam's argument that we are not brains in a vat (=*BIV*) has recently seen a resurgence in interest (Putnam 1981, cf. Button 2013, Madden 2013, Goldberg 2016, Thorpe 2018, 2019). Putnam's argument is ambitious: According to Putnam, we can know that we are not *BIV*s and we can even know this based on apriori reasoning and reflective self-knowledge alone. Some thought experiments about intentionality and reference together with reflection about what what we are currently thinking about and referring to suffice to rule out being a *BIV*.

Few philosophers have been persuaded by Putnam's argument. Objections to it are legion: It has been accused of being question-begging, of confusing claims about language with claims about reality, of taking a kind of self-knowledge for granted that is inconsistent with semantic externalism and of being pointless because a *BIV* can repeat it verbatim

DOI 10.5937/BPA2033039K

(for surveys cf. Brueckner 2012, Goldberg 2016). Yet the most prominent response is to concede for the sake of the argument that Putnam's argument successfully refutes *some* version of the *BIV* scenario, lifelong envatment, but to object that it does *not* refute *all* versions of it; in particular, it is said to be powerless against recent envatment. If this is so, Putnam's argument is at most a partial response to scepticism. The sceptical challenge remains alive as long as there is at least one sceptical scenario left that we cannot rule out.

Although *prima facie* convincing, this concessive response leads to problems of its own: It is doubtful that recent envatment is a truly *sceptical* scenario. In fact, as I shall argue below, the concessive strategy as defended in the literature fails for exactly this reason. Recent envatment is not a sceptical scenario. But I shall also argue that with some modifications the concessive strategy can be revived. For there is a different version of recent envatment that is both a truly sceptical scenario and is not refuted by Putnam's argument. The central idea here is this: The classical *BIV* scenario only questions perception as a source of knowledge, but there is no reason why memory should not be included among the epistemic sources under attack in the *BIV* scenario. Relying on this idea I argue that Putnam's argument indeed fails to refute scepticism because it fails to rule out *all* sceptical scenarios.[1] Interestingly, it does not fail because of some 'deep' philosophical mistake, but because it overlooks how flexible and adjustable sceptical scenarios are. If that is so, we can put the debates on whether Putnam's argument is question-begging, whether it relies on an implausible kind of self-knowledge, and so on to rest. No matter how these debates turn out, Putnam's argument cannot succeed since it fails to rule out all sceptical scenarios.

This completes my outline of the dialectical situation surrounding Putnam's argument. I will now go through all the steps in detail in order to defend how we should and how we should not respond to Putnam's argument. After briefly summarising Putnam's argument (§2), I discuss why replacing Putnam's original scenario with recent envatment is a bad objection against Putnam's argument (§3), what a better response looks like (§4) and why the latter is indeed a good response, i. e. why it is permissible to target memory without violating any constraints on sceptical scenarios (§5).

[1] It is sometimes argued that Putnam's argument is not meant to refute (Cartesian) scepticism, but to refute only metaphysical realism, i. e. to refute a picture of mind and world that underlies and motivates a specific kind of sceptical worry, but is not equivalent to scepticism. In this paper I argue only that Putnam's argument fails to refute scepticism. I think this is instructive even if Putnam's official target is not scepticism since, on the one hand, Putnam's argument is often taken to be relevant to this debate and, on the other hand, it is by no means obvious that his argument bears only on metaphysical realism.

2 A sketch of Putnam's argument

Cartesian scepticism argues that we cannot know anything we ordinarily think we know about the external world because we cannot rule out radical sceptical scenarios. One of those scenarios is the brain in a vat scenario. In its bare outline the scenario invites us to imagine not having a body, but being a brain kept alive in a vat while being connected to a supercomputer. This scenario is *radical* because it targets not just a small number of our ordinary beliefs and it is *sceptical* because it is difficult to see how we could ever be in a position to rule out being the victim of this scenario. Even a quick look at the literature, however, shows that there is no such thing as *the* brain in a vat scenario. Instead there is a shared template that can be embellished in myriad ways: Where are the *BIV* and the supercomputer located? What else exists in the universe? How long has the *BIV* been envatted? How and why was the *BIV* created? And these are just the basic questions. Additional questions can be raised about what happened to other sentient beings, the laws of physics and so on.

In discussions of Putnam's argument the version of the *BIV* scenario under consideration is usually lifelong envatment in its most radical form:[2]

> **Lifelong envatment.** By sheer chance the whole universe consists of nothing but the supercomputer and a brain in a vat attached to it. All sensory experiences of the envatted brain are the result of the supercomputer stimulating it in such a way that its experiences are indistinguishable from the ones I actually have.

Lifelong envatment is a good choice when discussing Putnam's argument for two reasons: On the one hand, it is hard to imagine a more radical scenario so that attempting to refute it is ambitious indeed. On the other hand, Putnam's core idea is easier to motivate when considering this version of the scenario: Putnam introduces and defends a causal constraint on reference and points out that by hypothesis lifelong *BIV*s do not meet this necessary condition for ordinary external world objects (like brains, hands, and so on): Since there are no hands in the scenario, there is *a fortiori* no causal connection to them. And although there is a brain, a vat and a computer, the causal connection to them is deviant and not of the kind required for reference. If, however, a *BIV* cannot refer to brains, vats, and so on, it can neither think nor state that it is a *BIV*. I, however, can think about whether I am a *BIV* – this is exactly what I am doing right now.[3] Hence, whenever I entertain thoughts about whether I am a *BIV*, I cannot be one.

[2] Putnam himself mentions both lifelong envatment (1981: 6, 12, 50) and recent envatment by an evil scientist (1981: 5f.).

[3] Moreover, if I were unable to even entertain the thought that I am a *BIV*, there would *a fortiori* be no sceptical threat, no possibility the sceptical argument could be based on.

In order to understand why this argument, if convincing, shows that we can rule out being a *BIV* based solely on apriori and reflective reasoning it is useful to spell out Putnam's argument explicitly:[4]

(1) In the language I am using right now "hand" refers to hands. (disquotation)

(2a) A *BIV* is not in causal contact with any hands. (from the description of the scenario)

(2b) Causal contact is necessary for reference. (from thought experiments about reference)

(2) In the language used by a *BIV* "hand" does not refer to hands. (from 2a and 2b)

(3) *Therefore:* I am not a *BIV*. (from 1 and 2, indiscernibility of identicals)

The first premise is trivial disquotational truth that I can know reflectively and the second premise is based on apriori thought experiments and on the description of the scenario. Since the conclusion follows deductively from premises which are based on apriori and reflective reasoning, it is is known based on apriori and reflective reasoning as well.

As already mentioned in the introduction, I will not discuss the various objections raised against this argument. I will not discuss whether the first premise already presupposes that I am not a *BIV* so that the argument is question-begging. Although it may seem that it presupposes that there is something I can refer to, it is also difficult to see how a disquotational triviality like this could be false: How could a word of my own language not refer to what I refer to by using that very same word? Another objection I will not discuss is whether the argument only shows "*I am not a BIV*" *is true* which is distinct from *I am not a BIV*. The idea behind this objection is that we want to find out whether we are *BIV*s, not whether everyone states something true when saying "I am not a *BIV*". A third objection I will not discuss is that a *BIV* could repeat the argument verbatim and show it is not a *BIV* either. The force of this objection depends on whether a *BIV* can in fact repeat the argument or merely think or utter something that looks similar.

3 A bad response to Putnam's argument

The reason why I do not discuss these objections is that a popular and straightforward reply to Putnam's argument (cf. the list of references in Thorpe 2018: 677[5]) is to concede all of the last section, but to point out that

[4] For somewhat similar, somewhat different reconstructions of Putnam's argument cf. Brueckner 1986, Wright 1992, Müller 2003. In the main text I present Putnam's argument as being about words, not sentences or thoughts. For this paper the differences do not matter.

[5] A further indicator of its popularity is that it is often mentioned in textbooks whose focus is *not* on scepticism, cf. e.g. Kallestrup 2014: 173 (a textbook on semantic externalism) or Newen & Schrenk 2013: 38 (a textbook on philosophy of language).

there are *BIV* scenarios whose victims can entertain the thought that they are *BIV*s and who can refer to external world objects, e. g. because of *past* causal connections. Let us call this strategy 'Putnam-proofing': A sceptical scenario is Putnam-proof iff its victim *can* entertain the thought that it is in that scenario. Putnam-proof scenarios cannot be ruled out with the help of Putnam's argument as sketched in the last section.

A natural way of Putnam-proofing the *BIV* scenario is to switch from lifelong to recent envatment. If the envatment happened yesterday, last week or last year, its victim can exploit past causal connections to entertain whatever thoughts she was able to entertain before envatment.

> **Recent envatment.** Last year someone was kidnapped and envatted. The brain's sense experiences are the result of a supercomputer stimulating it so that its experiences are indistinguishable from the ones I actually have.

However, recent envatment by itself cannot be used to challenge all or even most of my empirical beliefs. Beliefs about the past and inductive beliefs based on past observations are outside the scope of the resulting sceptical argument. This restriction has been noted quite often in the literature, but disagreement kicks in as to whether and why this is a problem for the sceptical argument. A minor problem is the *distinction without a difference problem*. The restriction to present empirical beliefs appears to be *ad hoc*. It is the result of Putnam-proofing the scenario, but does not reveal interesting epistemological differences within our empirical beliefs. Perplexingly, empirical beliefs about the past seem to be *better* off than empirical beliefs about the present. This problem need not be a knock-down objection. But even if the sceptical argument could be augmented by an additional step that somehow extends the result about present perceptual beliefs to all empirical beliefs (cf. Brueckner & Altschul 2010 and Smith 2016, see also Kraft 2014: 273–280 for some doubts), a sceptical argument without such epicycles seems to be preferable.

The more pressing problem, however, is the *evidence problem*: It is all to easy to underestimate *how much* evidence we have against recent envatment (for some glimpses cf. Tymoczko 1989: 295, Dennett 1991: 3–7, Kraft 2014: 274–275, Thorpe 2018: 679–682): *First*, there is neurophysiological and technological evidence against recent envatment: Last year human brains could not even be kept alive *in vitro* long enough, electrodes could not yet be connected to brains on a large scale, computers were not powerful enough to run the simulation and so on. *Second*, there is economic evidence: Even if practically possible, envatting humans is bound to consume a lot of resources and is not a routine procedure. *Third*, there is folk psychological evidence: Even if practically possible, there is no plausible motivation for envatting me instead of some other person. Evil guys with funds are on different missions. *Fourth*, there is evidence stemming from the smooth continuity in my life last year. If I was envatted last year, I must have been kidnapped. To cover up the kidnapping, evil scientists must pick the lock silently, shoot my dog before she barks, sedate me without waking me and transport me to their lab without family members or neighbours calling the police – not an impossible feat, but highly improbable. What is worse, the scenario is supposed to work for everybody. Scepticism is not restricted to those who like me are not paranoid and rich enough to sleep in a bunker or fortress, but claims that nobody – independent of their sleeping habits – can rule out the sceptical scenario. *Fifth*, the improbability of the scenario is raised even further if it includes that the earth or even the whole universe – except the *BIV*, of course – has been annihilated after envatment. It is difficult to come up with a more outlandish possibility.

To sum up, lifelong envatment is appealing as a sceptical scenario because it robs me of all evidence so that I cannot even tell what the probability of being in such a scenario is. In contrast, recent envatment leaves me with so much evidence that I can reasonably dismiss it based on circumstantial evidence. Of course, circumstantial evidence does not *guarantee* the scenario's falsity. But that does not rescue the sceptical argument: That our empirical evidence rarely hands out guarantees reminds us of our fallibility, but is a far cry from scepticism (cf. Kraft 2012).

4 A better response to Putnam's argument

The result seems to pose a dilemma: A sceptical scenario is *either* suited for a sceptical argument, but not Putnam-proof *or* it is Putnam-proof, but too easy to dismiss (cf. Thorpe 2018: 668). But that conclusion is premature: So far we have looked only at two versions of the *BIV* scenario. There are other versions in which there are enough causal connections left for the victim to be able to refer to external world objects and to entertain the thought that it is in that scenario, but not sufficient evidence for dismissing the scenario. In fact, going back to Putnam's original description of the *BIV* scenario gives us a hint for how to fix recent envatment:

> "He [= the evil scientist] can also obliterate the memory of the brain operation, so that the victim will seem to himself to have always been in this environment." (1981: 6)[6]

[6] In Nozick's version the evil scientist is even more powerful: "for any reasoning [...] we can imagine the psychologists [...] feeding *it* to their tank-subject, along with the (inaccurate) feeling that the reasoning is cogent" (1981: 167f.). Nozick's evil scientist is similar to Schaffer's debasing demon (Schaffer 2010). The victim of Nozick's scenario has a belief based on an incogent reason, but mistakenly thinks it is cogent. The victim of Schaffer's scenario has a belief based on an incogent reason, but mistakenly thinks it is based on a different reason. The scenarios discussed in the main text do not depend on such powerful scientists or demons.

This remark addresses a worry mentioned already: By obliterating memories the evil scientist can cover up the kidnapping so that the victim does not suspect that something is amiss. But once memory alteration is allowed, the sceptical toolbox suddenly contains many more scenarios. If the scenario tells a convincing story why my memory is untrustworthy, both the evidence problem – if memory is untrustworthy, I no longer have any evidence to dismiss recent envatment – and the distinction without a difference problem – beliefs acquired in the past are no longer treated differently – can be solved.

> **Recent memory-altering envatment.** Last year a member of an alien species living on a planet far away from earth was kidnapped and envatted. It underwent a training session devoted to radically altering its memories. This training session affected all its empirical memories, but not its apriori and conceptual knowledge.[7] Otherwise its memory works properly: It can reliably retrieve memories and its working memory is not affected at all. After the training session is completed, the envatted brain is sent to space. A supercomputer stimulates the brain in such a way that its experiences are indistinguishable from the ones I actually have. This all happens as a means of population control: The alien species prevent overpopulation on their planet by running an envatment lottery. Since they consider it unethical to let the losers know that they have lost, they devised the memory alteration scheme.[8]

This is a radical sceptical scenario: All the beliefs covered by lifelong envatment are also covered by this scenario.[9] The scenario even covers the *BIV*'s beliefs that brains are bihemisperical, grey and weigh approx. three pounds. In the scenario brains may well be octospherical, blue, weigh approx. twenty pounds with the *BIV* only seeming to remember having seen brain scans showing two hemispheres and so on. Thus, since all neurophysiological, technological, folk-psychological etc. beliefs are false, there is no evidence left that could be used to dismiss the scenario. Causal connections, however, are not affected in any way so that the *BIV* can entertain all thoughts it was able to entertain before envatment. Memory alteration is not memory replacement: Causal connections are left intact because memories are not overwritten by new ones, but only altered in a way that results in false beliefs.

[7] In the rare case that the victim lacks some relevant concepts the training session must involve some prior conceptual learning. For example, if the victim lacks the concept *brain*, it may be unable to think about brains for the trivial reason that it never acquired the concept before envatment.

[8] Memory alteration is rarely mentioned in the literature. Brueckner & Altschul 2010: 176, Briesen 2011: 574–576 and Gerken 2012: 72 are exceptions, but none of them discusses the permissibility of memory alteration in sceptical scenarios any further.

[9] Since the scenario is designed to be consistent with semantic externalism, the beliefs that water, Churchill and so on exist/-ed are exceptions. Surprisingly, McKinsey's paradox (1991) works for, not against the sceptical argument here: If these beliefs are non-empirical beliefs, as McKinsey's paradox suggests, they are exempt from sceptical doubts not because they are true in the scenario, but because they are non-empirical.

The fine print of recent memory-altering envatment is worth commenting on: *First*, all empirical memories are altered. One may wonder why a sceptical scenario with partial memory alteration does not suffice, e.g. restricting memory alteration to those memories that are evidence against recent envatment. A convincing sceptical scenario is one whose victim has no evidence – not even weak evidence – against being in that scenario. Again, it should not be underestimated how many memories have to be altered to achieve this goal. Altering all the neurophysiological, technological, folk-psychological etc. memories that may potentially be adduced as evidence requires altering large swaths of memories. *Second*, one may wonder whether there is really no evidence left to dismiss this scenario. What about arguing that running this lottery would consume too many resources on a planet already saddled with overpopulation? But even this, rather weak, evidence is ruled out. The aliens are presented as very ethical. They would never kill or neglect a fellow alien being. The elaborate memory-alteration is also needed for soothing the lottery's winners: Those who continue experiencing alien life will believe that they have not lost because only non-envatted aliens experience alien life. *Third*, in the scenario the victim's memory is altered in a training phase and the supercomputer no longer interferes with the victim's memory once training is completed.[10] This is important since it makes the scenario consistent with memory being distributed over the brain and avoids the need to postulate a 'memory box' in the brain to which a supercomputer could regularly feed new memories. The scenario does not depend on treating perception and memory as being similar. In particular, it does not preuppose that both involve some kind of experience, perceptual experience or memory traces. To the contrary, the scenario is neutral with respect to the various philosophical accounts of memory.

5 A good response to Putnam's argument?

The scenario from last section is likely to be met with resistance: Lifelong envatment is already a far-fetched thought experiment, but aliens running an envatment lottery overstrains the imagination – too much is too much, or so it seems. But recall that the aim of this paper is to argue for the usefulness and permissibility of memory alteration in sceptical scenarios, not to tell a thrilling and fascinating story. The interesting philosophical question is whether the restriction of sceptical scenarios to perception is well-motivated. My aim is to argue that if we take scenarios like lifelong envatment seriously, we cannot stop right there, but should allow recent memory-altering envatment as well.

[10] Recent work on optogenetics and memory in which memories of transgenic mice with light-sensitive neurons are manipulated provides some hints at how such a training phase might look like, cf. Ramirez et al. 2013, Liu et al. 2014, Robins 2016a.

The *first* objection I want to discuss is the *possibility objection*: A common constraint on sceptical scenarios is that they must present (what at least appear to be) genuine metaphysical possibilities. This is often taken to require that the sceptical scenario must be easily conceivable, that it must be consistent with our best philosophical and scientific theories about how the mind works and that it does not merely stipulate *that*, but explains *how* and *why* the beliefs of its victim fall short of knowledge (cf. Cross 2010, Kung 2011). An example for a scenario that does not meet this constraint is the jinn in a lamb scenario, a scenario which suggests you might be ghost living in a lamb waiting to be freed by Aladdin. Since we do not understand how minds can be realised as jinns in lambs and what beliefs and experiences jinns have while being in a lamb, we do not even know what it is we are asked to rule out.

Despite what one might think at first, memory alteration clears that bar. We should not reject the possibility of memory alteration just because we do not yet understand *all* the details of it. For the same is true of super computers feeding sense experiences. Although the rough outline is clear – plug a cable into the optic nerve –, the details are all just science fiction. If feeding sense experiences is thought to be sufficiently supported by science, memory alteration is so, too. After all, there already *is* scientific evidence for the possibility of memory alteration (in animal research, cf. memscience). Regarding easy conceivability the best criterion is to look at science fiction movies and popular science books. Those are open to memory alteration: There are at least two classic science fiction movies, *Blade Runner* (Scott 1982) and *Total Recall* (Verhoeven 1990), that deal with memory implants and at least one bestselling popular science book, *The Memory Illusion* (Shaw 2016), questioning our steadfast belief in the trustworthiness of memory. Hence, memory alteration is not an outlandish possibility discussed only in obscure epistemology circles.

The *second* objection I want to discuss is the *personal identity objection*: Memory is deeply connected with personal identity and, therefore, memory alteration endangers personal identity. If envatment involves near-total memory alteration, envatment creates a new person.

Both the main claim – envatment creates a new person – and the underlying assumption – if a new person is created, the sceptical scenario fails – are dubious. The claim that a new person is created clashes with some intuitions about the case: When suspecting that I may be the victim of such a scenario, I suspect that something bad happened to *me*, I want to go back to *my* old life, I want *my* memories back and so on. Moreover, memory continuity is not broken completely: If the alien had memories of some event, say its fifth birthday, it still has memories of its fifth birthday after envatment. Although the details of the memories are false – it now remembers its fifth birthday as its sixth, and it was not its birthday, but new year's eve –, it still remembers a particular event of its past, albeit falsely. Memory alteration should not be confused with memory replacement (for similar distinctions in different contexts cf. Byrne 2010, Robins 2016b).

But even if a new person is created, the sceptical argument does not fail. The causal constraint on reference does not rule out that the causal connection involves several persons. After all, I can refer to mammoths and other objects from the distant past because of inherited causal connections. As long as the causal connection between the person before envatment and the person after envatment is sufficiently tight, as in recent memory-altering envatment, the latter can inherit reference from the former even if it is a different person.

The *third* objection I want to discuss is the *reference shift objection*: Can a *BIV* whose memory has been radically altered really refer to the things it had causal contact with before envatment? As Evans' "Madagascar" example (Evans 1973) illustrates, errors can result in reference being re-routed: Although there is a causal chain from an area of mainland Africa to current utterances of the proper name "Madagascar", the name does not refer to the mainland area, but to the island.

Even if "the idea that there is a *moment* at which the languages switch just seems faintly ludicrous" (Button 2013: 159), the general consensus is that reference does not switch *instantaneously*. There is nothing magical about referring to something that is completely misremembered. For example, someone can refer to Churchill even if everything she believes about him is based on false memories and even if she recently moved to a place where "Churchill" is commonly used as a name for, say, some living jazz singer. In the end the causal constraint is a double-edged sword when used against scepticism (cf. Burge 2003): It rules out some error-possibilities, e. g. lifelong envatment, but is at the same time consistent with reference despite widespread error, e. g. the example just given or Kripke's Gödel-Schmidt example (reference to Gödel is independent of whether all or most of one's beliefs about him are true, cf. Kripke 1980: 83–84).

Yet, although in the case of "Madagascar" a reference shift occurred only after Marco Polo's error caught on, it may still seem that recent memory-altering envatment is an altogether different case. One way of motivating this claim relies on replacing "causal" in Putnam's original causal constraint by "world-involving abilities" (cf. e. g. Putnam 2013: 25): What really matters for reference are abilities to do something with worldly objects, not mere causal connections. In recent memory-altering envatment the changes are so pervasive that the relevant world-involving abilities are lost and reference is shifted instantaneously. It is not obvious, however, why only one's *present* world-involving abilities should matter for reference. If only present world-involving uses matter, *all* (irreversible) switches from one environment to another would result in instantaneous reference shifts. If both present and past world-involving uses matter, the reformulated constraint does not show that in recent memory-altering envatment instantaneous reference shifts occur. Combining a transfer to a new environment with memory alteration may accelerate an otherwise slower switch, but there is no reason to think that it is turned into an instantaneous one.

The *fourth* objection I want to discuss is the '*causal contact is necessary, not sufficient*' *objection*: So far I have at most shown that a victim of recent memory-altering envatment meets the causal constraint on reference. This, of course, does not show that it actually can refer to external world objects. After all causal contact is not sufficient for reference, but only necessary. One route to take here is to accept Williamson's principle of knowledge maximisation (2007: ch. 8) and the associated idea that:

> "Roughly: a causal connection to an object [...] is a channel for reference to it if and only if it is a channel for the acquisition of knowledge about the object [...]." (2007: 264)

Based on this idea one may argue that even if there is a causal connection between a victim of recent memory-altering envatment and external world objects, it cannot refer to external world objects since it cannot acquire knowledge about those objects via the causal connection.

In response I concede the main point: My aim was to show only that Putnam's argument, with the causal constraint it depends on, cannot refute that we are victims of recent memory-altering envatment. Of course, a different constraint on reference may be able to do that, but that would not be *Putnam's* argument anymore. That being said let me add some worries about relying on a Williamsonian knowledge constraint on reference to refute recent memory-altering envatment. As formulated by Williamson, the constraint is timeless, i. e. it does not state that I can refer *now* only to what I can *now* acquire knowledge about. For example, it allows that someone referred to something and acquired knowledge about it in the past, but due to an undercutting defeater lost her knowledge about it later. It also allows that in cases of dementia the patient can refer to, say, a long dead aunt by her proper name although he has lost all knowledge about her. But to rule out recent memory-altering envatment the constraint must be understood synchronically: I can *now* refer only to what I can *now* acquire knowledge about. Only this stronger constraint has the consequence that a victim of recent memory-altering envatment is unable to refer to external world objects. The synchronic knowledge constraint on reference, however, seems to be too strong as cases of undercutting defeat and severe memory loss show.[11]

The *final* objection I want to discuss is the *epistemic autonomy objection*: This objection is based on a constraint on sceptical scenarios according to which the victim of a sceptical scenario may not lose its epistemic autonomy, i. e. the beliefs must be the victim's own beliefs and the victim must be able to

[11] For further criticism of Williamson's principle of knowledge maximisation and the associated knowledge constraint on reference, cf. McGlynn 2012.

reflect rationally on the epistemic standing of her own beliefs. This constraint on sceptical scenarios is meant to rule out a variety of uninteresting sceptical scenarios such as:

> **Robot.** There is a robot all of whose 'beliefs' are regularly externally updated via WiFi, including its 'beliefs' that its 'beliefs' are based on experiences and reasons. It happens that the robot's 'beliefs' and 'experiences' are indistinguishable from the ones I actually have.
>
> **Shortcuts.** There is someone in whose brain random shortcuts are occurring all the time. It happens that the random shortcuts result in beliefs and experiences that are indistinguishable from the ones I actually have.
>
> **Confabulation.** There is someone who suffers from a severe confabulation syndrome whose sufferers never realise that they have it. By chance the confabulation results in the beliefs that are indistinguishable from the ones I actually have.

Of course, I cannot rule out being in such a scenario. Yet this does not mean that the sceptical argument is successful. Victims of such scenarios lack minimal epistemic autonomy so that the alleged beliefs are no longer the victim's *own* beliefs and the victim is unable to reflect rationally on the epistemic standing of her beliefs. If the 'beliefs' of the victim are directly controlled by something external or are the result of deviant causal processes in the brain, she does not have false beliefs, but the external agent or the deviant process (at most) cause the victim to store a false representation. Analogously, if a book contains a false account of the world (no matter whether it was written intentionally or came about by chance), the paper on which the book is printed does not have false 'beliefs'. Moreover, rational reflection on the epistemic standing of one's beliefs is impossible since the results of such a reflection are affected by external updating, random shortcuts or confabulation, as well. If I suspect to be in such a scenario, I must suspect that my reasoning about the scenario is affected as well – taking such scenarios seriously is self-undermining.[12] Thus, sceptical arguments must rely on a scenario in which the victim has beliefs of her own and can reason about them.[13]

[12] To see why sceptical arguments must not rely on self-undermining scenarios consider, for example, the closure argument: I know that having hands entails not being a *BIV*. Since knowledge is closed under known entailment, this means that I if I know that I have hands, I also know that I am not a *BIV*. But I do not know whether I am not a *BIV*. Therefore, I do not know whether I have hands. – The uninteresting scenarios mentioned in the main text cannot be relied on in the closure argument: Either I know the entailment or I do not know the entailment. If I do not know the entailment, there is no sceptical threat. If I do know the entailment, I am not in one of the uninteresting scenarios. For victims of these scenarios cannot trust their own reasoning, not even their reasoning about entailments, and therefore lack knowledge of any entailment. Either way there is no sceptical threat.

[13] It is an interesting question whether Nozick's *BIV* scenario or Schaffer's debasing demon (cf. fn. 6) meet this constraint. I am not going to try to answer this question in this paper.

Although the epistemic autonomy constraint is central for understanding sceptical scenarios, it does not rule out memory alteration. In recent memory-altering envatment the victim's conceptual knowledge, reasoning skills and working memory are not put into question. What is put into question are empirical memories, but that is not self-undermining. As long as my *present* rationality and my *present* minimal epistemic autonomy is taken for granted, it is *my* beliefs that I *reason* about. The training session may alter *dispositional* beliefs (it does so on at least some conceptions of dispositional belief). For example, even before the newly envatted alien thinks explicitly about it for the first time, it dispositionally believes that it is on earth. In this regard the scenario looks similar to the scenarios mentioned in the last paragraph: The dispositional beliefs are not really the victim's own beliefs. However, manipulating *dispositional* beliefs is consistent with minimal epistemic autonomy. As long as the dispositional beliefs are open to review and one is able to reason critically about them and to sustain or change them accordingly, one's epistemic autonomy is not threatened. To sum up, minimal epistemic autonomy is not threatened by the kind of memory alteration envisioned in recent memory-altering envatment.

6 Conclusion

If the argument of this paper is successful, Putnam's argument fails independently of more philosophically loaded objections to it. It fails because sceptical scenarios are flexible and adjustable in ways that allow for Putnam-proofing them. Although any sceptical scenario must meet several constraints in order to pose a serious challenge, there is, as I have argued, no constraint that rules out memory alteration. If that is so, recent memory-altering envatment is a sceptical scenario Putnam's argument must refute or else it fails as a general anti-sceptical strategy.

This paves the way for a final observation that is not limited to Putnam's argument: If memory alteration is permissible in a sceptical scenario, the sceptical toolbox turns into a Pandora's box. Once opened, a wide variety of new scenarios emerge in which this or that cognitive process is manipulated in a way undermining knowledge (cf. Schaffer 2010). Unless sceptical scenarios in which both perceptual and non-perceptual cognitive processes are manipulated can be disallowed in a principled way, the prospects for anti-sceptical strategies that, like Putnam's, are tailored to the specifics of a particular scenario look dim.[14]

[14] Earlier versions of this paper have been presented to audiences in Cologne and Regensburg. Many thanks for discussion and criticism to Alexander Dinges, Andrea Klonschinski, Christoph Michel, Hans Rott, Joshua Thorpe and a reviewer for this journal.

References

Briesen, Jochen (2011): „Antiskeptische Trittbrettfahrer des semantischen Externalismus“, in: *Zeitschrift für philosophische Forschung* 65: 563–585.

Brueckner, Anthony (1986): “Brains in a Vat”, in: *Journal of Philosophy* 83: 148–167. (Reprinted in *Essays on Skepticism.* Oxford: OUP, 2010: 115–132.)

Brueckner, Anthony & Altschul, Jon (2010): “Terms of envatment”, in: Brueckner, Anthony: *Essays on Skepticism.* Oxford: OUP, 174–176.

Brueckner, Anthony (2012): “Skepticism and content externalism”, in: Zalta, Edward (ed.): *The Stanford Encyclopedia of Philosophy* (Spring 2012 Edition), https://plato.stanford.edu/archives/spr2012/entries/skepticism-content-externalism/

Burge, Tyler (2003): “Some reflections on scepticism: Reply to Stroud”, in: Hahn, Martin & Ramberg, Bjørn: *Reflections and Replies. Essays on the Philosophy of Tyler Burge.* Cambridge/Ms.: MIT Press, 335–346.

Button, Tim (2013): *The Limits of Realism.* Oxford: OUP.

Byrne, Alex (2010): “Recollection, perception, imagination”, in: *Philosophical Studies* 148: 15–26.

Cross, Troy (2010): “Skeptical success”, in: *Oxford Studies in Epistemology* 3: 35–62.

Dennett, Daniel (1991): *Consciousness Explained.* Boston: Little, Brown & Co.

Evans, Gareth (1973): “The causal theory of names”, in: *Proceedings of the Aristotelian Society, Supplementary Volumes* 47: 187–208.

Gerken, Mikkel (2012): “Critical notice: *Essays on Skepticism*”, in: *International Journal for the Study of Skepticism* 2: 65–77.

Goldberg, Sanford (ed.) (2016): *The Brain in a Vat.* Cambridge: CUP.

Kraft, Tim (2012): “Scepticism, infallibilism, fallibilism”, in: *Discipline Filosofiche* 22: 49–70.

Kraft, Tim (2014): “Defending the ignorance view of sceptical scenarios”, in: *International Journal for the Study of Skepticism* 5: 269–295.

Kripke, Saul (1980): *Naming and Necessity.* Cambridge/Ms.: HUP.

Kung, Peter (2011): “On the possibility of skeptical scenarios”, in: *European Journal of Philosophy* 19: 387–407.

Liu, Xu; Ramirez, Steve & Tonegawa, Susumu (2014): “Inception of a false memory by optogenetic manipulation of a hippocampal memory engram”, in: *Philosophical Transactions of the Royal Society B* 369: 20130142.

McGlynn, Aidan (2012): “Interpretation and knowledge maximization”, in: *Philosophical Studies* 160: 391–405.

McKinsey, Michael (1991): “Anti-individualism and privileged access”, in: *Analysis* 51: 9–16.

Madden, Rory (2013): “Could a brain in a vat self-refer?”, in: *European Journal of Philosophy* 21: 74–93.

Müller, Olaf L. (2003): *Wirklichkeit ohne Illusionen.* 2 Vols., Paderborn: mentis.

Newen, Albert & Schrenk, Markus (2013): *Einführung in die Sprachphilosophie.* 2nd ed., Darmstadt: WBG.

Nozick, Robert (1981): *Philosophical Explanations.* Cambridge/Ms.: HUP.

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

Putnam, Hilary (2013): “From quantum mechanics to ethics and back again”, in: Baghramian, Maria: *Reading Putnam.* London: Routledge, 19–36.

Ramirez, Steve; Liu, Xu; Lin, Pei-Ann; Suh, Junghyup; Pignatelli, Michele; Redondo, Roger L.; Ryan, Tomás J. & Tonegawa, Susumu (2013): “Creating a false memory in the hippocampus”, in: *Science* 341 (6144): 387–391.

Robins, Sarah (2016a): “Optogenetics and the mechanism of false memory”, in: *Synthese* 193: 1561–1583.

Robins, Sarah (2016b): “Misremembering”, in: *Philosophical Psychology* 29: 432–447.

Schaffer, Jonathan (2010): “The debasing demon”, in: *Analysis* 70: 228–237.

Scott, Ridley (1982): *Blade Runner* [Motion Picture]. USA: Warner Bros.

Shaw, Julia (2016): *The Memory Illusion. Remembering, Forgetting, and the Science of False Memories.* London: Random House.

Smith, Martin (2016): “Scepticism by a thousand cuts”, in: *International Journal for the Study of Skepticism* 6: 44–52.

Thorpe, Joshua (2018): “Closure scepticism and the vat argument”, in: *Mind* 127: 667–690.

Thorpe, Joshua (2019): “Semantic self-knowledge and the vat argument”, in: *Philosophical Studies* 176: 2289–2306.

Tymoczko, Thomas (1989): “In Defense of Putnam’s Brains”, in: *Philosophical Studies* 57: 281–297.

Verhoeven, Paul (1990): *Total Recall* [Motion Picture]. USA: TriStar Pictures.

Williamson, Timothy (2007): *The Philosophy of Philosophy.* Oxford: Blackwell.

Wright, Crispin (1992): “On Putnam’s proof that we are not brains-in-a-vat”, in: *Proceedings of the Aristotelian Society* 92: 67–94.

Peter Murphy,
University of Indianapolis
murphyp@uindy.edu

Original Scientific Paper
UDC 161/162:165.4
164.05:165.72

SUSPENSION-TO-SUSPENSION JUSTIFICATION PRINCIPLES

Abstract: *We will be in a better position to evaluate some important skeptical theses if we first investigate two questions about justified suspended judgment. One question is this: when, if ever, does one justified suspension confer justification on another suspension? And the other is this: what is the structure of justified suspension? The goal of this essay is to make headway at answering these questions. After surveying the four main views about the non-normative nature of suspended judgment and offering a taxonomy of the epistemic principles that might govern which suspended judgments are justified, I will isolate five important principles that might govern which suspended judgments are justified. I will call these suspension-to-suspension principles. I will then evaluate these principles by the lights of each of the four views about what suspensions are. I close by drawing some conclusions about the prospects for skepticism, the structure of justified suspended judgment, and the importance of theorizing about justified suspended judgment.*

Keywords: *coherentism, foundationalism, infinitism, skepticism, suspended judgment*

While sophisticated theories of justified belief have proliferated, especially in the recent history of western epistemology, the same is not true for theories of justified suspended judgments (hereafter “suspensions”). This is unfortunate since it may mean that that we are not yet in a strong position to fully understand and evaluate various skeptical theses. The skeptical theses I have in mind are those that say that suspension is the only stance we are justified in taking to the claims in some domain. Fortunately, times are changing. Epistemologists have recently started to study suspension much more closely than they have before.[1] This paper adds to this new trend by offering some new arguments that bear on the nature of justified suspension, and consequently on skepticism’s prospects.

My focus will be on two questions that any adequate theory of justified suspension must answer. First, when, if ever, does one justified suspension

[1] Due in large part to the groundbreaking work of Jane Friedman. See especially Friedman (2013) and (2017). For some interesting criticisms of some of Friedman’s work, see Archer (2018) and Archer (forthcoming).

DOI 10.5937/BPA2033055M

confer justification on another suspension? And second, how do clusters of justified suspensions hang together? Possible answers to these questions consist in epistemic principles that might govern justified suspensions.[2] After providing a taxonomy of the kinds of principles that might govern the realm of justified suspensions, I will argue that some of the key principles that might govern when one justified suspension confers justification on another suspension and some of the key principles that might govern how justified suspensions hang together should be rejected. I will then look at what these negative findings mean for the skeptic's prospects, the structure of justified suspension, and the importance of theorizing about justified suspension.

The paper has five main sections. Since the non-normative nature of suspension is an important determinant of the justification norms that govern suspensions, and since the non-normative nature of suspension is a matter of dispute, I will begin, in Section 1, by reviewing the main contending views about the nature of suspension. Rather than trying to decide among these views (something that would require a very different, and much longer, essay), I will proceed in a theory-neutral manner and look at how each of the contending views about the nature of suspension fits with various suspension principles. In Section 2, I step back and offer a taxonomy of the kinds of principles that might govern justified suspensions. Within that taxonomy, I locate what I will be calling *suspension-to-suspension principles*. These principles are modeled on the familiar Closure, Transmission, and Counter-Closure principles that are sometimes thought to govern justified belief. In Section 3, I say why these principles are especially important for the theory of justified suspension. In Section 4, I determine how each of these principles fares on the theories identified in Section 1. Then in Section 5, I draw some lessons.

1. Views About The Nature of Suspension

Both of my main questions concern the normative nature of suspension. To answer these questions, though, we may first have to identify suspension's non-normative nature. So I will begin by briefly stating the main competing views about the non-normative nature of suspension.[3] I am interested here in views with ontological ambitions – views, that is, that try to capture what

[2] If there are no true epistemic principles that govern suspensions, then the correct theory of justified suspension might be a particularist one. My working assumption is that the epistemic status of particular suspensions is governed by general principles.

[3] Epistemologists almost invariably theorize about the normative nature of belief without first taking a view about the ontology of belief. Below I provide reason to think that this same way of proceeding is not suitable for theorizing about the normative nature of suspension.

suspensions are identical to, and not views which merely state necessary and sufficient conditions on someone's suspending judgment. With this in mind, here are the views[4]:

> *The Sharp Credence View:* to suspend about *p* is to equally divide one's credence between *p* and *not-p* by having a 0.5 credence that *p* and a 0.5 credence that *not-p*.[5]
>
> *The Maximally Mushy Credence View:* to suspend about *p* is to be in a maximally mushy credence state regarding *p*, where this state is spread out across the full 0–1 interval.[6]
>
> *Higher-Order Belief Views:* to suspend about *p* is to have some distinctively epistemic higher-order belief – for example, the belief that one is neither justified in believing *p* nor justified in disbelieving *p*.[7]
>
> *The Inquiry View:* to suspend about *p* is to inquire into whether or not *p* is true.[8]

Though much can be said in favor of, and against, each of these views about suspension's non-normative nature, my main concern will be with the epistemic norms that govern suspensions. Consequently, I am going to remain as neutral as possible about the (non-normative) nature of suspension.[9] Still there are obvious connections between views about the (non-normative) nature of suspension and theories of justified suspension. After all, it seems that Proponents of The Sharp Credence View must take the epistemology of sharp credences as providing the correct theory of justified

[4] I omit the view, now widely dismissed, which says that suspending regarding *p* is identical to not believing *p* and not believing *not-p*. This view has several problems, among them that we do not suspend about propositions that we never entertain. More generally, in offering these four candidate views of the nature of suspension, I assume that suspending is an attitude of some kind; for a defense of this assumption, see Friedman (2013a).

[5] *The Sharp Credence View* is an example of a middling sharp credence view, where views of this kind say that to suspend about *p* is to equally, *or to approximately*, divide one's credence between *p* and *not-p* by (i) having a 0.5, *or close to 0.5*, credence that *p*; and (ii) having a 0.5, *or close to 0.5*, credence that *not-p*. The points that I go on to make about *The Sharp Credence View* also apply to other middling credence views. For detailed discussion and criticism of middling sharp credence views, see Friedman (2013b).

[6] For discussion and development of this view, see Sturgeon (2010). I borrow the term "mushy credence" from White (2010).

[7] For discussion and development of this view, see Raleigh (forthcoming); Rosenkranz (2007) is also relevant.

[8] See Friedman (2017), though Friedman is primarily interested in defending the view that suspension and inquiry are biconditionally related, and not the stronger view that they are identical.

[9] If all four of the views I have outlined are mistaken, and some fifth view is correct, I hope to have at least uncovered a few things about the epistemology of the attitudes isolated in these four views. Of course, if one of those four views is correct, we will have learned about three other important kinds of attitudes as well.

suspension; that proponents of The Maximally Mushy Credence View must take the epistemology of mushy credences as providing the correct theory of justified suspension; that proponents of Higher-Order Belief Views must take the theory of justified belief as providing the correct theory of justified suspension; and that proponents of The Inquiry View must take the epistemology of inquiry (which tells us when inquiring into some question is justified) as providing the correct theory of justified suspension. In the face of these diverse views about the nature and norms of suspension, I will remain neutral and try to argue from inductions across these four views about the nature of suspension and their accompanying theories of justified suspension. This will allow us to determine where there is, and where there is not, unanimity about whether some suspension principle is true. As we will see, there is some impressive unanimity. Before getting to the supporting inductions though, we need to survey the general principles that might govern the justification of suspensions.

2. A Taxonomy of Candidate Suspension Principles

My next task is to provide a taxonomy of the candidate principles that might govern whether some given suspension is justified. Note that I am after *candidate* principles. These are the principles that we need to consider when constructing a full theory of justified suspension. Though I will ultimately reject some of these principles, I need to begin by identifying and organizing all of the candidate suspension principles in a way that is helpful for understanding and evaluating those principles.

2.1 Strong and Weak Suspension Principles

The candidate principles divide into strong principles and weak principles. These two kinds of principles are closely connected to the two organizing questions that I mentioned at the outset. Recall that the first question was this: when, if ever, does one justified suspension *confer* justification on another suspension? Proposed answers to this question will cite strong principles. A strong principle is a generalization that tells us what *confers* justification on some justified suspensions. For example, a subset of strong principles consists in those principles that tell us when some justified suspensions confer justification on other suspensions. More generally though, any generalization which claims that some specified kind of fact confers justification on some specified suspensions counts as a strong principle. I refer to these as *strong principles* because they go beyond stating necessary or sufficient conditions for a suspension's being justified, and identify what it is that *confers* justification on suspensions. The notion of conferring is crucial to a principle being a strong principle. This notion is meant to be an ecumenical one that can cover a variety of ideologies. So you can think of it in terms of what metaphysically grounds the justification of the relevant suspensions, or you can think of it in terms of what the truth-makers are for ascriptions of justified suspensions, or you can think of it in terms of what it is in-virtue of which those suspensions are justified, etc.

Weak principles, by contrast, carry no implications about what confers justification on suspensions. Weak principles are less ambitious: they simply identify necessary, or sufficient, conditions for a suspension's being justified. Still weak principles are important because they tell us how justified suspensions hang together. This makes them crucial for determining the answer to my second organizing question. Recall that was this question: how do clusters of justified suspensions hang together? One justified suspension, as I will put it, hangs together with another suspension, as long as one necessitates, or suffices, for the other. However, since neither necessitating nor sufficing requires that one suspension confers justification on another suspension, these principles are logically weaker than strong principles.

The distinction between strong and weak principles has some important implications. Since a strong principle might be false, but only because what it cites does not *confer* justification on suspensions, though what it cites does suffice for that suspension to be justified, strong principles have weak counterpart principles that cite a sufficient condition for the relevant suspension to be justified. A corollary of this is that a strong principle entails the corresponding weak principle that cites a sufficient condition on the justification of some suspensions, while that weak principle does not entail that strong principle. It also follows that weak principles that cite a *necessary condition* on the justification of some suspensions do not have strong principles as counterparts.[10]

2.2 A Taxonomy of Kinds of Suspension Principles

To generate the candidate strong and weak principles that might govern justified suspensions, I am going to look to the familiar strong and weak principles that have been proposed for justified belief; I will transpose the principles that I find there to arrive at a set of principles that might govern justified suspensions.

In the arena of justified belief, strong principles can be distinguished by the kinds of items that are claimed to confer justification on some of our beliefs. Those candidate items are:

1. other justified beliefs,
2. other justified doxastic states of the subject (e.g. justified credences or justified suspensions),
3. non-doxastic states of the subject (e.g. perceptual experiences),

[10] This is because the satisfaction of a necessary condition on the truth of *x is f* does not entail that *x is f* and so it does not entail that anything confers *f* on *x*.

4. the reliability of the belief-forming process (and methods) that produced the belief,
5. the indispensable role that the belief plays in some inquiry (as proponents of justified belief in hinge propositions contend),
6. a feature of the belief (or its context) that confers default justification on it,
7. a combination of the previous items.

Transposing this to the realm of justified suspensions yields candidate strong principles, which claim that the following items confer justification on some of our suspensions:

1. other justified suspensions,
2. other justified doxastic states of the subject (e.g. justified beliefs or justified credences),
3. non-doxastic states of the subject (e.g. perceptual experiences),
4. some property of the suspension-forming process (and methods) that produced the suspension,
5. the indispensable role that the suspension plays in some inquiry,
6. a feature of the suspension (or its context) that confers default justification,
7. a combination of the previous items.

This delivers the branch of the taxonomy that consists in the candidate strong principles.

To complete the taxonomy, we need to add the candidate weak principles. Recall that some weak principles cite a necessary condition on the justification of some specified suspensions, and others cite a sufficient condition on the justification of some specified suspensions. Recruiting from the last list of seven items yields seven weak principles, each claiming that one kind of item suffices for some specified suspensions to be justified. The same can be done to yield principles which say of the respective items that they are necessary for some specified suspensions to be justified – this yields seven more principles.[11] In total then we have fourteen kinds of weak principles to put alongside the seven kinds of strong principles that were identified in the previous paragraph.

2.3 Two Kinds of Suspension-to-Suspension Principles

I am now going to narrow my focus to strong and weak principles that recruit the first kind of item on our list, namely other justified suspensions. I will call principles of this kind, *suspension-to-suspension* principles.[12]

[11] For a useful discussion of some weak principles that connect beliefs to suspensions, see Rosa (forthcoming).

[12] Perhaps the two kinds of weak suspension-to-suspension principles (i.e. one kind that offers a justified suspension that is necessary for a target suspension to be justified, and the other that offers a justified suspension that is sufficient for the same target suspension

My focus will be even narrower though. This is because suspension-to-suspension principles divide into two different subtypes. One subtype consists in principles that link two suspensions whose contents are logically related, while the other subtype consists in principles that link two suspensions whose contents are not logically related. My focus will be on the first subtype.

Let me briefly illustrate the second subtype, before leaving principles belonging to that type behind. An example of a principle of the second subtype, where the contents are not logically related, is a principle that recruits a justified suspension about the reliability of one's own epistemic faculties. A principle like this arguably plays a central role in the arguments of Descartes's First Meditation. Here is an example of this kind of principle: if a person has a justified suspension about whether the beliefs that are produced by one of her faculties, *f*, are certain, then this confers justification on all of her suspensions that are produced by *f*.[13] There is much to be said about principles of this type. I mention them, and offer this example, though, only to illustrate that some suspension-to-suspension principles connect suspensions whose contents are not logically related. My focus in what remains will be on the other subtype, namely principles that connect suspensions whose contents are logically related to one another. I will call these *content-connecting suspension principles*.

3. The Importance of Content-Connecting Principles

Before examining some of the leading content-connecting principles, I want to highlight two reasons why content-connecting principles are especially important in the theory of justified suspension.

One concerns implications for skepticism. Recall the construal of skepticism as covering views, which say, for some domain of claims, that the only justified stance that we can take to any of the claims in such a domain is one of suspension. With this in mind, notice this important point: if it turns out that there are suspension-to-suspension principles of the content-connecting kind, then skepticism will be *infectious*. This is because *if* some claim is one for which suspension is the only justified stance, then the correct content-connecting principles will ensure that suspension is the only justified stance to take with respect to some other claim. In this way,

to be justified) can be collapsed into one kind. Though the two kinds of principles have the same form, the justified suspensions that are necessary for some target suspension to be justified could be very different from the justified suspensions that are sufficient for that same suspension to be justified.

[13] This principle allows that the recruited suspension confers justification on the target suspensions in a mediated way, by first operating as a justification defeater for the beliefs produced by the relevant faculty, which in turn confers justification on the suspensions produced by that faculty.

justified suspension will spread. Of course, how justified suspension spreads in this way, from claim to claim, and the extent to which it does so, will depend on exactly which content-connecting principles are correct and on what logical relations hold among the relevant claims. Still the basic point holds: suspension-to-suspension principles of the content-connecting kind can spread justified suspension, both through a domain, and perhaps also to other domains.

On the other hand, if no (or few) content-connecting principles are true, then the skeptic cannot argue from the fact that a subject is justified in suspending about one proposition to the claim that the subject is also justified in suspending about some logically related claim. The skeptic will have to find some other strategy to try to show that the subject is justified in suspending about the second claim, and this will require the skeptic to defend and deploy some other kind of suspension principle.

The second reason it is important to evaluate content-connecting, suspension-to-suspension principles is that doing so will help to reveal the structure of justified suspension. Or, to use some of my earlier language, it will help us to see how justified suspensions hang together. To see this, return to the arena of justified belief. This time consider the debate between foundationalists, coherentists, infinitists, and skeptics over how our justified beliefs hang together. This debate is triggered by the familiar regress argument. That argument, recall, has, as a crucial premise, a belief-to-belief justification principle. This formulation of such a principle will serve our present purposes: if a belief is inferred from some other beliefs, then the inferred belief is justified only if the beliefs that it is inferred from are justified.[14] A principle like this can be repeatedly applied to trigger a regress of justified belief. In turn, that regress forces us to choose one of four views. One, foundationalism, says that the regress terminates with justified beliefs that are not justified by other beliefs. A second, infinitism, says that the regress goes on ad infinitum. A third, coherentism, says that the best way to resolve the regress is to take justification to primarily attach to sets of beliefs, rather than individual beliefs. And a fourth, a form of skepticism, says that the best resolution is to conclude that there are no justified beliefs after all.

If there is a parallel suspension-to-suspension principle that triggers a regress of justified suspensions, then we will have to choose one of four parallel views about how justified suspensions hang together. One, a foundationalist view, says that the regress of justified suspensions terminates with justified suspensions that are not justified by any other suspensions. A second, a form of infinitism, says that the regress of justified suspensions goes on ad infinitum. A third, coherentism about the structure of justified

[14] To be defensible, this principle needs to be refined in several ways; such refinements do not, however, take away from my main point here. For some of the needed refinements, see Luzzi (2014).

suspensions, says that the best way to resolve this regress is to take justification to primarily attach to sets of suspensions, rather than individual suspensions. And, a fourth is a form of skepticism, which says that the best resolution to this regress is to conclude that there are no justified suspensions after all.

Alternatively, if it turns out that all of the suspension-to-suspension principles that could trigger a regress of justified suspensions are false, then we must look elsewhere to determine how justified suspensions hang together. Regardless, any adequate theory of justified suspension needs to tell us how justified suspensions hang together; or it needs to tell us why they don't hang together in any interesting way. I will return to the structure of justified suspensions and the infectiousness of justified suspension as I work through some specific content-connecting, suspension-to-suspension principles.

4. Five Content-Connecting Principles

Let's now examine some specific suspension-to-suspension principles of the content-connecting kind. In this section, I will formulate and evaluate five principles of this kind.

4.1 The Weak and Strong Dual Principles

I begin with this very plausible principle:

> **Weak Dual:** If S's suspension about *p* is justified, then S's suspension about *not-p* is also justified.

Notice that this is a weak principle since it does not say that the justified suspension about *p* is what confers justification on the suspension about *not-p*.

This principle is very plausible; there is surely something normatively incoherent about suspending about *p*, while believing, or even disbelieving, not-p. Further support for Weak Dual comes when we consider how this principle comes out on the theories of justified suspension that naturally accompany the four views of suspension that were cataloged earlier. Consider then the main claims that are delivered by orthodox versions of the epistemologies that naturally go along with each of those views of suspension. First, on orthodox epistemologies of sharp credence, the relevant claim is highly plausible. It is this claim: if S's 0.5 credence that *p* is justified, then S's 0.5 credence that *not-p* is also justified. Second, on one epistemology of mushy credence, the relevant claim is the following highly plausible claim: if S's maximally mushy credence that *p* is justified, then S's maximally mushy credence that *not-p* is also justified. Third, on a plausible epistemology of belief, the key claim that follows from the conjunction of Weak Dual and a sample Higher-Order Belief View is highly plausible, though its logical form is complex. It is this claim: if S has a justified higher-order belief that *n* (where *n* is: S is neither justified in believing *p* nor justified in believing *not-p*), then S's higher-order belief that *not-n* is also justified (where *not-n* is: S is neither justified in believing *not-p* nor justified in believing *p*) – in fact, these are arguably one and the same higher-order belief.[15] And, last, on at least one plausible epistemology of inquiry, the key claim is also a highly plausible one. It is this claim: if S's inquiry into whether *p* is true is a reasonable act of inquiry, then her inquiry into whether not-p is true is also a reasonable act of inquiry – in fact, these are arguably one and the same inquiry. Weak Dual, I conclude, enjoys strong support.

What about Weak Dual's strong counterpart? That is this principle:

> **Strong Dual:** If S's suspension about *p* is justified, then this confers justification on S's suspension about *not-p*.

This principle is far less plausible. One problem it faces is a symmetry problem. It is highly plausible that the relation of conferring justification is asymmetric[16]; but there does not seem to be any epistemic asymmetry between suspensions about *p* and suspensions about *not-p*, which would give one of these the needed priority over the other, so that the one always confers justification on the other, but the second never confers justification on the first. There is more to be said here about Strong Dual, but this is a serious strike against it.

Even if Strong Dual is true though, not much follows about either of the two issues highlighted in the previous section. The skeptic will not get far by using either Strong Dual or Weak Dual since these principles only allow justified suspension to spread from a person's required stance regarding *p* to her stance regarding *not-p*. As for the structure of justified suspension, the most these principles imply is that justified suspensions about contradictory propositions will either both be justified or both be unjustified. This, however, tells us nothing about the structure of any other sets of justified suspensions besides those that are directed at contradictories. So neither dual principle can help us decide among the four structural options that were

[15] Notice that the only possible difference between these two beliefs lies in the order of the disjuncts in the negated disjunctions that are the contents of the respective higher-order beliefs. The logical form of the content of the first belief is ~(JB*p* v JB~*p*), and the logical form of the content of the second is ~(JB~*p* v JB*p*).

[16] Someone might resist the claim that conferring justification is an asymmetric relation by claiming that coherentist theories of justification provide us with models of beliefs that mutually confer justification on one another. Two points in response. First, plausible coherentist theories do not say that two beliefs in logically equivalent propositions can confer justification on one another – such a view would be much too permissive. Plus coherentists typically hold that much larger sets of beliefs need to be in place for there to be any justified beliefs. And, second, on the best formulations of coherentism, justification primarily attaches to groups of beliefs, not individual beliefs – so one individual belief never confers justification on another individual belief.

outlined earlier.[17] However, the fact that justified suspensions are paired in this way is compatible with the view that either the justification enjoyed by the suspension about *p* or the justification enjoyed by the suspension about *not-p* has its justification conferred on it by something outside this pair of suspensions. And the operative principle that governs that conferring might be some other strong suspension-to-suspension principle; moreover, that principle might trigger a regress of justified suspensions. Let's examine some principles that might do this.

4.2 The Closure and Transmission Principles

Next are some principles that are modeled on some familiar belief-to-belief principles. Two are counterparts of one another. The weak principle in the pair is modeled on the familiar idea that justified belief is closed under known entailment. We can work with this rendering of that idea, a single-premise closure principle, which says:

> **Closure for Belief:** If (i) S's belief that *p* is justified, (ii) S competently deduces *q* from *p*, and (iii) S thereby comes to believe *q* while retaining her justified belief that *p* throughout, then S's belief that *q* is justified.[18]

Here is the parallel principle for suspension:

> **Closure for Suspension:** If (i) S's suspension about *p* is justified, (ii) S competently deduces *q* from *p*, and (iii) S thereby comes to suspend about *q* while retaining her justified suspension about *p* throughout, then S's suspension about *q* is justified.

To see the strong counterparts of these principle, return to belief. There Closure for Belief has as its strong counterpart the so-called "transmission principle", which adds to Closure for Belief the claim that it also follows from

[17] There are other candidate suspension-to-suspension principles, besides Weak Dual and Strong Dual, that have no weighty consequences for skepticism's prospects or for revealing the structure of justified suspensions because they can't (even if true) spread justified suspension across claims in some significant way. Other examples include principles modelled on disjunction introduction (e.g. if S's suspension about *p* is justified, then S's suspension about *p or q* is justified), principles modelled on conjunction introduction (e.g. if S's suspension about *p* is justified and S's suspension about *q* is justified, then S's suspension about *p & q* is justified), and principles modeled on reasoning across a known biconditional (e.g. that if S has a justified suspension about *p*, and S knows that *p iff q*, then S's suspension about *q* is also justified). Notice that the claim that these principles do not carry weighty consequences for skepticism or for the structure of justified suspensions is about what these principles imply; it is not a claim about whether they are true. See Rosa (forthcoming) for arguments against simple versions of these principles.

[18] Read 'competently deduces' as a success term that implies knowledge; so *S competently deduces q from p* entails that *S knows that p entails q*. My formulation here is closely modeled on Hawthorne's Single-Premise Closure principle for knowledge in his (2004).

the antecedent of Closure for Belief that it is the satisfaction of (i)-(iii) that *confers* justification on S's belief that *q*. Here is a version of that principle:

> **Transmission for Belief:** If (i) S's belief that *p* is justified, (ii) S competently deduces *q* from *p*, and (iii) S thereby comes to believe *q* while retaining her justified belief that *p* throughout, then (i)-(iii) confer justification on S's belief that *q*.

And here is the parallel principle for suspension:

> **Transmission for Suspension:** If (i) S's suspension about *p* is justified, (ii) S competently deduces *q* from *p*, and (iii) S thereby comes to suspend about *q* while retaining her justified suspension about *p* throughout, then (i)-(iii) confer justification on S's suspension about *q*.

Because Transmission for Suspension is stronger than Closure for Suspension, if Closure for Suspension is false, then Transmission for Suspension is also false. I will now argue that Closure for Suspension is false.

Consider a coin that is about to be flipped, and a subject who has a justified suspension about whether the coin will land heads. Suppose that our subject knows that *the coin will land heads* entails propositions like *coins exist*, *coins will be flipped*, and *coins either land heads or tails*. Suppose also that she competently reasons from the claim that the coin will land heads to one of these last claims, and that while doing so she retains her suspension about whether the coin will land heads. Closure for Suspension implies that she is justified in suspending about *coins exist*, *coins will be flipped*, and *coins land either heads or tails*. But this is clearly false.

The implausibility of Closure for Suspension is obvious when the point is put in terms of sufficient conditions. Here it is helpful to think in terms of a sufficient condition, *s*, on the truth of a proposition, *p*. Thought of in terms of necessary and sufficient conditions, Closure for Suspension says that if someone has a justified suspension about whether some sufficient condition, *s*, for the truth of *p* has been satisfied, then she is justified in suspending about *p*. This is clearly false though. Having this kind of justified suspension does not preclude a subject from having a justified belief (or knowledge) that another sufficient condition for *p is met*. Since the latter will put her in an excellent position to have a justified belief that *p* (or even to know *p*), it precludes her from being justified in suspending about *p*.[19]

This is confirmed when we run this last point through the four views of suspension. First, The Sharp Credence View allows that one can have a justified 0.5 credence that *s*, and yet not be justified in having a 0.5 credence that *p*, all while knowing that *s entails p*. This will happen when one has a sufficiently high justified credence that some other sufficient condition on the

[19] Here, and elsewhere, I assume the following weak uniqueness principle: if S is justified in believing *p*, then S is not justified in suspending about *p*.

truth of *p* is satisfied. Similarly, The Maximally Mushy Credence View allows that one can have a justified maximally mushy credence about *s*, yet not be justified in having a maximally mushy credence about *p*, all while knowing that *s entails p*. This will happen when one has a sufficiently high justified credence that some other sufficient condition on the truth of *p* is satisfied. The same is true on Higher-Order Belief Views. Here one can have a (higher-order) justified belief that one is neither justified in believing, nor justified in disbelieving, *s*, yet not be justified in believing (at the higher-order) that one is neither justified in believing *p* nor justified in disbelieving *p*, again all while knowing that *s entails p*. This will happen when one is justified in believing that some other sufficient condition on *p* is satisfied. And, last, The Inquiry View of suspended judgment allows that it can be reasonable for one to inquire into whether *s* is true, yet it not be reasonable for one to inquire into whether *p* is true, again, while also knowing that *s entails p*. All three of these things are true when one knows that some other sufficient condition on the truth of *p* is satisfied. This last piece of knowledge makes it unreasonable to inquire into whether *p* is true or false.[20] This review of the four leading views of the nature of suspension provides additional support against Closure for Suspension.

4.3 Excursus: Can Suspensions Figure Into Inferential Reasoning?

It is worth pausing at this point to consider an important issue that bears on Closure for Suspension, Transmission for Suspension, and another principle that I will soon examine. The issue concerns a potentially problematic assumption that might underlie these principles. The concern is that these principles assume that suspensions can figure into inferential reasoning, as beliefs do. But it is far from clear what is involved in, say, reasoning and inferring to a conclusion, about which one suspends judgment.

The issue of whether suspensions can figure into inferential reasoning is important and underexplored. But it is an issue that can only be adequately answered with a much different, and longer, paper. More importantly, since it concerns the non-normative nature of suspensions, rather than the normative nature of suspensions, I am going to set it aside. Still, two points are worth keeping in mind. The first is that it is not clear that Closure for Suspension and Transmission for Suspension really do require that suspensions figure into inferential reasoning. There are two reasons for this. One is that *competently deducing q from p* does not require either suspending about *q* or suspending about *p* – keep in mind that it is propositions, not attitudes towards propositions, that are deduced from one another. Second, *thereby coming to suspend about q* can occur as the last in a series of mental states without that series constituting an episode of inferential reasoning. In fact,

[20] For defense of this claim, see Friedman (2017).

suspensions might show up as both the first and last items in such a series and yet that series not be an episode of inferential reasoning. This is so with belief: a series of mental states can begin with one belief and end with another belief, and yet that series not constitute an episode of inferential reasoning. For example, on some models of higher-order belief, a person can begin with the belief that grass is green, then go into some self-monitoring mental states, and then form the higher-order belief that she believes that grass is green – and yet no inferential reasoning needs to have occurred during this time.[21] Similarly, suspensions might figure into a series of mental states, perhaps one that also includes deducing one proposition from another, even if that series does not constitute an episode of inferential reasoning.

Second, it is worth going back over the views about the nature of suspension canvassed in Section 1, and asking whether they allow for suspensions to figure into inferential reasoning. When we do this, we find that on Higher-Order Belief Views, there is no problem with suspensions figuring into inferential reasoning, since suspensions are themselves just beliefs (and surely beliefs can figure into inferential reasoning). As for The Sharp Credence View and The Maximally Mushy Credence View, we can at least say that proponents of these views should be motivated to model suspensions so that they can figure into inferential reasoning since this is just an instance of the more general ambition of modelling credences so that they can figure into inferential reasoning. It would be a serious cost to theorizing with degrees of belief (whether sharp or mushy), if degrees of belief cannot figure into our inferential reasoning.[22] Things are admittedly different with The Inquiry View. Here it seems much less obvious that suspensions can figure into inferential reasoning, since, according to this view, suspensions are directed at questions, not propositions. There might however be some non-obvious way to show that questions can figure into inferential reasoning.

Let's return to the normative nature of suspensions and look at one last content-connecting suspension-to-suspension principle.

4.4 The Counter-Closure Principle

The last point I argued for before the excursus concerned justified suspensions about known sufficient conditions. The point was this: having a justified suspension about some sufficient condition for the truth of a proposition does not entail being justified in suspending about that proposition. This is because one can have a justified belief (or know) that some other sufficient condition on the truth of that proposition is satisfied – when this is so, one is justified in believing that proposition, and one is not justified in suspending about it.

[21] In such cases one belief is non-inferentially based on another belief.

[22] For a partial defense of the view that credences can figure into inferential reasoning, see Staffel (2013).

What about having a justified suspension about whether a *necessary condition* on the truth of a proposition is met? This is equivalent to the question of whether a principle that I will call *Counter-Closure for Suspension* is true. To see what this principle is, consider first this version of the Counter-Closure principle for justified belief:

> **Counter-Closure for Belief:** If (i) S has a justified belief that *q* (ii) that is solely based on a competent inference from *p* to *q*, and (iii) S believes that *p*, then S's belief that *p* is justified.[23]

Here is the parallel principle for suspension:

> **Counter-Closure for Suspension:** If S has a justified suspension about *q*, (ii) S competently infers *q* from *p*, and (iii) S suspends about *p*, then S's suspension about *p* is justified.[24]

This principle is a weak one since it does not say that the justified suspension about *q confers* justification on the suspension about *p*.

Counter-Closure for Suspension is the best candidate for a principle that can trigger a regress of justified suspensions. For, if this principle is true, then each proposition that is known to entail a proposition about which one is justified in suspending will itself be a proposition about which one is justified in suspending. Counter-Closure for Suspension could then iterate, and thereby spread justified suspension further back through entailing propositions, thus triggering a regress of justified suspensions. Of course, if Counter-Closure for Suspension is false, then it won't trigger any such regress and will therefore not shed any light on the structure of justified suspensions.

Initially Counter-Closure for Suspension might seem plausible. For example, I know that a necessary condition on a coin having landed heads is that the coin was flipped. But if I have a justified suspension about whether the coin was flipped, then surely I am also justified in suspending about whether it landed heads. Despite this, however, the principle is false. The reason is simple: in addition to having a justified suspension about whether a necessary condition for *p* is met, one might have a justified belief (or knowledge) that some other necessary condition for *p* is *not met.* When this happens, one is *not* justified in suspending about *p* – instead one is justified in believing that *p* is false.

This too is confirmed by each of the four views of suspension. Here, in somewhat compressed form, are the crucial points. First, The Sharp Credence

[23] This formulation is fine for present purposes. Again, for some of the needed refinements, see Luzzi (2014).

[24] Notice that no reference is made in this principle to any inferential reasoning that involves a transition from one suspension to another suspension, and thus reflects the thought from the last section that perhaps no inferential reasoning needs to figure into any content-connecting suspension-to-suspension principles.

View allows that one can have a justified 0.5 credence that *n*, a justified 0.5 credence that *not-n*, know that *p entails n*, and yet not be justified in having a 0.5 credence that *p*. All of this will be true if one has a sufficiently high justified credence that some other necessary condition on the truth of *p* is not satisfied. Similarly, The Maximally Mushy Credence View allows that one can have a justified maximally mushy credence that *n*, know that *p entails n*, yet not be justified in having a maximally mushy credence that *p*. This will happen when one has a sufficiently high justified credence that some other necessary condition on the truth of *p* is not met. The same pattern holds on Higher-Order Belief Views. One can have a higher-order justified belief that one is neither justified in believing, nor justified in disbelieving, *n*, know that *p entails n*, and yet not be justified in having a higher-order belief that one is neither justified in believing, nor justified in disbelieving, *p*. Once again this will happen when one is justified in believing that some other necessary condition on *p* is not satisfied. Last is The Inquiry View. This view allows for cases in which it is reasonable for one to inquire into whether *n* is true, when one knows that *p entails q*, and yet it is unreasonable for one to inquire into whether *p* is true. All of this will hold when one knows that some other necessary condition on the truth of *p* is not satisfied. This last piece of knowledge makes it unreasonable to inquire into whether *p* is true or false. This induction across the four views is more support for rejecting Counter-Closure for Suspension.

5. Lessons

Where does all of this leave us? In particular, what lessons can we now draw about the skeptic's prospects and about the structure of justified suspensions?

Take the skeptic's prospects first. Since we have, at the very most, only identified two suspension-to-suspension principles that are true (namely Weak Dual and Strong Dual), and since those two principles can do very little to spread justified suspension across a body of claims, the skeptic's prospects look a little dimmer. Principles like Closure for Suspension, Transmission for Suspension, or Counter-Closure are not available to show that justified suspension is infectious, since those three principles are false.

What about the structure of justified suspension? Based on the finding that Counter-Closure for Suspension is false, we can at least say this much: a regress of justified suspensions is not triggered in the same way that a regress of justified beliefs is typically thought to be triggered. Since it is unclear how else a regress of justified suspensions could be triggered, it remains unclear how justified suspensions might hang together. Much more work needs to be done to reveal the structure of justified suspensions.[25]

[25] Perhaps because the approach that borrows from the theory of justified belief has not yielded anything, an entirely different approach is needed to reveal the structure of justified suspension.

I end with a third lesson. It is based on the kinds of considerations and the cogency of the considerations that I offered against Closure for Suspension, Transmission for Suspension, and Counter-Closure for Suspension. Those considerations, both in kind and in cogency, are very different from the familiar considerations that are offered against the parallel principles for justified belief. Epistemologists on both sides of the debates about Closure for Belief, Transmission for Belief, and Counter-Closure for Belief will, I think, agree that there are no simple, highly cogent considerations about the epistemology of necessary and sufficient conditions that can be offered against any of these three principles. Debates about those principles continue because there are no simple highly cogent arguments on either side of the debate. As we have seen though things are quite different for the parallel principles governing suspension: each of those principles has been definitively shown to be false, both on general grounds about the epistemology of necessary and sufficient conditions, and on the basis of the inductions across the four leading views about the nature of suspension. The difference in both the kinds and cogency of the considerations that bear on the parallel principles in the realms of suspension and belief is some reason to think that theorizing about justified suspension can unfold very differently from how theorizing about justified belief unfolds. This, I submit, gives us even more reason to continue to theorize about justified suspended judgment.

Acknowledgments

For helpful remarks on a much earlier version of this paper, thanks to my commentator, Baron Reed, and the audience at the 2018 Central Division Meetings of the American Philosophical Association. Thanks also to Adam Sanders and an anonymous referee for this journal.

References

Archer, A. (2018). Wondering What You Know. *Analysis* 78: 596–604.

Archer, A. (2019). Agnosticism, Inquiry, and Unanswerable Questions. *Disputatio* 11: 63–88.

Friedman, J. (2013a). Suspended Judgment. *Philosophical Studies* 162: 165–181.

Friedman, J. (2013b). Rational Agnosticism and Degrees of Belief. *Oxford Studies in Epistemology* 4: 57–81.

Friedman, J. (2017). Why Suspend Judging? *Nous* 51: 302–326.

Hawthorne, J. (2004). *Knowledge and Lotteries*. New York: Oxford University Press.

Luzzi, F. (2014). What Does Knowledge-Yielding Deduction Require of Its Premises? *Episteme* 11: 261–275.

Raleigh, T. (forthcoming). Suspending is Believing. *Synthese*.

Rosa, L. (2019). Logical Principles of Agnosticism. *Erkenntnis* 84: 1263–83.

Rosenkranz, S. (2007). Agnosticism as a Third Stance. *Mind* 116: 55–104.

Staffel, J. (2013). Can There Be Reasoning With Degrees of Belief? *Synthese* 190: 3535–3551.

Sturgeon, S. (2010). Confidence and Coarse-Grained Attitudes. *Oxford Studies in Epistemology* 3: 126–149.

White, R. (2010). Evidential Symmetry and Mushy Credence. *Oxford Studies in Epistemology* 3: 161–186.

Živan Lazović
Department of Philosophy
University of Belgrade

Original Scientific Paper
UDC 165.82 Патнам Х.
165.3/.4

IS PUTNAM'S 'BRAIN IN A VAT' HYPOTHESIS SELF-REFUTING?

Abstract: *In this paper, I provide a detailed analysis of Putnam's conclusion (derived from the externalist interpretation of meaning and mental content) that the skeptical hypothesis, according to which we have always been brains in vats, is self-refuting. I confine my attention to the following question: If we assume that semantic externalism is plausible on independent grounds, does it follow that the semantic argument against skepticism (as articulated by Putnam) is indeed successful? In the first section, I briefly review the basic contention of Putnam's semantic externalism. In the second section, I outline and reexamine Putnam's, Brueckner's, and Warfield's version of the semantic argument. I hope to show that Putnam's version of this argument remains on a purely meta-linguistic level, which means that it can only prove that the phrase 'We are brains in a vat' must be false when it is considered in the context of the argument, although it most certainly does not prove that we are not brains in a vat after all. In the third section, I argue that Brueckner's and Warfield's attempt to modify Putnam's argument, and consequently provide an a priori proof that we are not brains in a vat, are ultimately unsuccessful, for both attempts beg the question against the skeptic. In the final section, I draw a comparison between the skeptical hypothesis and other cases of self-refuting statements and conclude that Putnam was ultimately right in claiming that the skeptical hypothesis is self-refuting in a weak sense, in which it is unassertible, although it might be true nevertheless.*

Keywords: *semantic externalism, reference, disquotation, self-knowledge, truth, assertibility.*

1. Introduction

In (1973, 1981), Putnam presents and articulates his externalist view of the meaning of referring (singular and general) terms, as well as of the content of our thoughts about physical objects. This view is widely known in the relevant philosophical literature as *semantic externalism* (SE). Although Putnam's primary intention in presenting his version of SE is to elucidate the nature of the relationship between our mind and the world, he indicates that SE can also be used to show that the skeptical hypothesis (SH)—in its most radical form, according to which we have always been brains in a vat (BIV) in an otherwise empty universe—is self-refuting (1981: 7). On the basis of these indications, later proponents of SE have offered the so-called *semantic*

DOI 10.5937/BPA2033073L

argument (SA) against such a Cartesian-inspired form of skepticism. Given that both SE and SA have generated many philosophical discussions, I will not go into their detailed analysis here. Instead, I will focus on the following question: If we assume that SE is plausible on independent grounds, does it follow that SA (as articulated by Putnam) is indeed successful? In other words, the question I will be dealing with is whether SA shows that the skeptic's position is self-refuting in the sense in which, under the assumption that SE represents the correct view, it follows that SH must be false. I will argue in this paper that, given SE, the skeptic's position is in fact self-refuting in the weak sense, according to which SH is unassertible, although it might be true nevertheless.

2. Putnam's Semantic Externalism and BIV Hypothesis

Before presenting and analyzing SA, I will provide a brief explanation of SE. Thus, in Putnam's view, the basic thesis of SE could be formulated as follows:

> (SE): Reference and hence the meaning of referring terms (such as 'water', 'tree', 'table' and the like), as well as the content of our thoughts about the objects to which these terms refer, is at least partially determined by environmental factors, among which the causal link between the use of these terms and the objects to which they refer plays a prominent role.

In order to obtain a fuller understanding of how the causal constraint, expressed in the above formulation, determines the meaning of the referring terms and the content of sentences and thoughts that include them, three points are especially worth noting. First, the meaning of a referring term is at least partially determined by the direct causal link, established *via* its original introduction into the linguistic practice, as well as by the indirect causal chain of its later applications by the members of the language community. Thus, for instance, the term 'water' is first introduced by the speakers who have had direct causal encounters with individual samples of a substance with such-and-such properties. This direct causal link is then extended *via* the appropriate chain of communication, wherein the speakers use the term 'water' to speak and think about water.

Second, if it turns out that one and (linguistically) the same referring term has a different causal history in two language communities, then it will also have a different meaning in these language communities. This is shown by Putnam's famous 'Twin Earth' thought experiment (Putnam 1973), where we are told to imagine a scenario in which on a planet (nearly) identical to ours—the so-called *Twin Earth*—there are people who represent our physical and phenomenological duplicates and who, perhaps unsurprisingly, speak the same language as we do. Now, both on Earth and on Twin Earth, there is a substance that Earthlings and Twin Earthlings identify by their superficial (i.e. observable), stereotypical properties as water and refer to its samples by applying the word 'water'. The only difference is that on Earth, the molecular structure of this substance is H_2O, while on Twin Earth, it is an entirely different substance with the molecular structure XYZ. Under the assumption that the molecular structure constitutes the identity of a substance, from the fact that Earth and Twin Earth differ with respect to the external environment and the causal history of the term 'water', it follows that this term has different meanings in English and vat-English. That is to say, when Earthlings use the word 'water', they refer to a substance composed of H_2O, whereas their duplicates on Twin Earth refer to a substance with the molecular structure XYZ. This fact brings about a difference with respect to the content of the appropriate sentences that we and our duplicates on Twin Earth formulate when we talk about water, as well as to the content of our thoughts about water, for the truth conditions of the sentences (and thoughts) about water on Earth differ from their truth conditions on Twin Earth. Namely, when we (on Earth) point to a sample of the liquid in a glass in front of us and say 'This is water', and when our duplicates on Twin Earth do the same, our sentence will be true if the liquid in the glass has the molecular structure of H_2O, while the sentence of our duplicates on Twin Earth will be true if the liquid in the glass in front of them has the molecular structure XYZ.

Third, if the speaker fails to meet the causal constraint on a referring term—that is, if she has never been in direct or indirect causal contact with the object to which she applies the referring term—then she cannot form the corresponding concept, make any assertions or, ultimately, have any thoughts *about* the object to which this term refers. We can thus see that the semantic significance of the corresponding causal link between the referring expressions and the objects to which they refer has a strong impact on the linguistic competence of the speaker. Namely, in order to be able to properly understand the meaning of a referring expression—that is, to use it correctly in speech and to think about the objects to which it refers—it is necessary to be in the appropriate direct or indirect causal contact with these objects. It is also important to note that, although Earthlings and Twin Earthlings use the term 'water' with different meanings, they can still successfully satisfy the causal constraint within their own language communities. Yet, someone—whether on Earth or Twin Earth—who has never been (directly or indirectly) in the appropriate causal contact with any sample of water, would be utterly unable to understand the meaning of the word 'water' or, consequently, formulate sentences and form thoughts about water (see, Putnam 1981: 12, 16; Kallestrup 2012: 36).

Now, as mentioned at the outset, Putnam was convinced that SE could serve as a powerful argumentative tool against the Cartesian-inspired philosophical skeptic. In order to avoid any possible quandaries about whether we can meet the causal constraint in hypothetical situations in which we are only temporarily victims of the systematic deception, or in which we are constantly deceived by any other subject that otherwise meets the causal constraint, Putnam introduces (for the sake of argument) the most radically updated version of the Cartesian skeptical hypothesis:

> (SH) In an otherwise completely empty world, as a result of some cosmic accident, we are always disembodied brains envatted in a nutrient fluid, connected to a super-computer and having experiences, including thoughts, that are caused only by computer-generated electrical impulses.

It is worth pointing out that the scenario described in SH shows striking similarity to the above-mentioned Twin Earth thought experiment. Namely, observe that BIVs in SH should be understood as our phenomenological twins; that is, they represent our exact psychological duplicates with respect to sensory evidence, thoughts and interior monologue. In a BIVs' world in which there are no physical objects, the super-computer produces experiential experiences in the BIVs' minds that are qualitatively indistinguishable from the experiences that we have in our actual environment. We can thus see that in SH, the semantic point about the reference of the term 'water' from the Twin Earth thought experiment is extended to cover *all* referring terms in the BIVs' world. Suppose a glass containing liquid is in front of me and that, on the basis of my sensory evidence, I identify that liquid as water. Suppose further that I say, 'This is water', expressing with this sentence the content of my thought about the liquid in front of me. According to SH, it follows that my BIV—in its otherwise empty world—has the same sensory evidence produced by the appropriate electrochemical stimulation and that—in its own interior monologue—it utters the same sentence 'This is water'. Now, in this waterless world, the BIV cannot have any causal contact with water as a physical liquid, but rather with entities that in *its* own world play a causal role with respect to *its* uses of 'water' that is analogous to the causal role that the instances of water play with respect to *my* uses of 'water'. If these entities in the BIVs' world are, say, the recurring computer program features, then, according to the causal constraint of SE, the BIVs' word 'water' does not refer to water but rather to the recurring computer program feature <W>, which causes electrical stimuli in BIVs and, in turn, produces experiences that are qualitatively indistinguishable from the experiences of our embodied brains that are stimulated as a result of seeing water in normal circumstances.

The difference between the reference of the word 'water' in the actual world (in which we are normal human beings in our typical physical environment) and the reference of the word 'water' in the BIVs' world brings about an important difference in the semantic content (i.e. the truth-conditions) of my sentence 'This is water' and the BIV's sentence 'This is water' respectively. Namely, in the actual world, my sentence will be true if the liquid in the glass in front of me really is water; in the BIVs' world, however, the sentence 'This is water' will be true if the computer program feature <W> is running. In other words, given SE, with the phrase 'This is water', my BIV and I (each in our own world) assert *different* statements and express *different* thoughts—of course, only if we assume, following Putnam, that BIVs can have any thoughts.

On the basis of this observation, it seems that the semantic difference between our and BIVs' sentences and thoughts can be successfully represented by using a disquotational mechanism, as a device that we use in ordinary (natural) language—given that it is semantically closed (i.e. contains semantic predicates which include both 'referring to' and 'true') and universal (i.e. contains both object– and meta-language)—in order to explicate both the reference of terms and the truth-conditions of declarative sentences.[1] Thus, by applying a disquotation mechanism, the reference of my word 'water' in English is determined by the following sentence:

(R_E) 'Water' refers to water;

and the truth-conditions of my sentence 'This is water' (which I assert while pointing to the liquid in a glass) is determined by the following equivalence:

(T_E) 'This is water' is true iff this is water.

Now, let us suppose that in my case, this truth-condition is obtained; i.e. that the liquid in a glass I am pointing to is, in fact, water. If we attach the same meaning to the words of BIVs—in their waterless and glassless world—the sentence 'This is water' would not be true. But given the causal constraint involved in SE, the word 'water' in vat-English does not refer to water, but rather to the computer program feature <W>. Since this is so, it follows that the truth-condition of BIV's sentence 'This is water' obtains when the computer program feature <W> is running. In other words, at least from the perspective of our language—used as a meta-language for BIVs' object-language—it seems that the usual disquotation mechanism is not applicable in vat-English,[2] for we cannot obtain the reference to the word 'water' or the truth-condition of the sentence 'This is water' in vat-English by simply removing the quotation marks; rather, we would have to use the following formulations in *our own language*:

(R_E) 'Water' refers to a computer's program feature <W>;

that is,

(T_E) 'This is water' is true iff the computer's program feature <W> is running.

[1] When sketching his argument in *Reason, Truth and History* (1981), Putnam does not use disquotation, but he resorts to it in his 'Replies' (1992: 347–408).

[2] Cf. Brueckner 1992: 205.

Although the BIVs in SH are represented as our phenomenological duplicates, their language seems to lose the semantic properties of closedness and universality,[3] while the reference of their individual words and the truth-conditions of their sentences cease to have the disquotational character that is familiar to normal English speakers, due to the fact that the words they use in their empty environment are causally connected to the features of the computer program, and not to familiar physical objects.

Now, what are the consequences of all this for the hypothesis (SH), which the skeptic, as a normal speaker, wants to formulate in a natural language? Suppose this hypothesis simply reads 'We are brains in a vat'. By formulating this hypothesis in *our* language, the skeptic certainly takes the words 'brain' and 'vat' with their *usual* meaning, in which they refer to *brains* and *vats* as physical objects; that is, objects with which they are causally related. Given the formulation of SH, it is clear that the skeptic must not assume that *we have not always been* brains in a vat, or, ultimately, that *she herself is not*—at least at the moment in which she is presenting her skeptical hypothesis—a brain in a vat. But if SE is correct, the skeptic's position appears to be self-refuting for *semantic* reasons, that is, on the basis of the meaning of the terms used in the formulation of her hypothesis, as well as on the basis of its semantic content. This is exactly what Putnam claims. In his view, if SE is correct, the hypothesis that we have always been brains in a vat '*cannot possibly be true*, because it is, in a certain way, self-refuting' (1981: 7); that is to say, it represents a supposition whose truth implies its own falsity. Now, Putnam makes it clear that a statement can be self-refuting in at least two different ways (1981: 7–8). First, there are statements that are self-refuting only because of their semantic content, regardless of whether anyone asserts them or not; such is, for example, the statement 'All statements are false', for if this statement is true, it follows that it must be false. But there are also statements that are self-refuting partly due to their grammatic form; i.e. due to who and in which form asserts or contemplates them. Putnam's example of this particular type of statement is 'I do not exist', which, given the meaning of the pronoun 'I', must be false whenever (and in all circumstances in which) *any* person asserts or contemplates it in the present tense. After recalling this distinction, Putnam states that SH represents an instance of the second group of self-refuting statements:

> What I will show is that the supposition that we are brains in a vat has just this property. If we can consider whether it is true or false, then it is not true (I shall show). Hence it is not true. (1981: 8)

In the remainder of the paper, I will attempt to show that the fact that SH falls into the second group of self-refuting statements significantly diminishes the power of Putnam's version of SA against skepticism; namely, I will argue that

[3] Ibid. 211.

from the fact that, given SE, neither the skeptic nor any of us can claim—in the usual sense in which 'brains' and 'vats' refer to *real* brains and *real* vats—that SH is true without implying that SH is false, it does not follow that SH cannot be true after all. In the next section, I will consider in more detail Putnam's versions of SA, but I will also pay attention to Brueckner's and Warfield's versions of this argument.

3. Putnam's Semantic Argument against the Skeptic

Putnam has tried to show that SH must be false by appealing to SE and its causal constraint on reference. As we have seen, according to SE, the words that BIVs use in their otherwise empty world, although linguistically the same as the words we use in the actual world, cannot refer to ordinary physical objects, given that BIVs have never been in causal contact with these objects. We have already agreed that in the BIVs' world, these words refer to the corresponding computer program features with which the tokens of their uses are causally connected. This also applies to the words 'brain' and 'vat' involved in SH. Thus, just as in the BIVs' world—where the word 'water' does not refer to water as a physical substance, but rather to the computer program features <Ws> with which BIVs' tokens of that word are causally connected—the words 'brain' and 'vat' in SH do not refer to *brains* and *vats* as physical objects, but rather to the computer program features <Bs> and <Vs>.

Following Brueckner (1986), we will present Putnam's argument in the disjunctive form, according to which we are either BIVs or we are not BIVs. The disjunctive formulation of this argument has the following consequences. If we are not BIVs, then by uttering the sentence 'We are BIVs' we mean that *we are BIVs* and, taken with this meaning, the sentence is clearly false. On the other hand, if we are BIVs, then by uttering the sentence 'We are BIVs' we would mean that we *are <Bs> in <Vs>*. However, since SH represents the hypothesis about *real* brains and *real* vats, rather than about the computer program features <Bs> and <Vs>, the BIV's sentence 'We are BIVs' turns out to be false. Given that the sentence 'We are not BIVs' is false whether or not we are BIVs, it follows that the opposite sentence 'We are not BIVs' must be true (cf. Putnam 1981: 14–15).

Yet, as Brueckner (1992) rightly observed, Putnam's argumentation works on a meta-linguistic level, which proves that the *sentence* 'We are BIVs' (when we assert or consider it) must be false. We therefore need at least one additional step in order to reach the conclusion that *we are not BIVs*. Proponents of SE typically maintain that this step could be made either by applying the disquotation mechanism (Putnam 1992; Wright 1992; Brueckner 1986, 1992) or, alternatively, by invoking the assumption—the so-called *self-knowledge thesis* (SK)—that the subject has privileged access to the contents of her mental states (Tymoczko 1989, Warfield 1998, Brueckner 2003). Let us consider these two strategies in turn.

Brueckner (1986) argued that we can reach the conclusion that we are not BIVs by applying the following disquotation principle:

(T) 'We are not BIVs' is true iff we are not BIVs.

Combined with the conclusion of Putnam's original SA, according to which the sentence 'We are not BIVs' must be true, the equivalence (T) leads us to the further conclusion that we are *not BIVs*. However, Brueckner himself (1986: 164–165) expressed concern that the application of the disquotation principle (T) in the context of disjunctive SA begs the question against the skeptic. As Folina (2016) and McKinsey (2018) show, this concern is well grounded. First of all, in the context of argumentation that starts with the disjunctive premise 'Either we are not BIVs or we are BIVs', (T) is most certainly *ambiguous*. Namely, note that whether we speak normal English or vat-English depends on whether or not we are BIVs. In either of these two languages—i.e. as (T_E) or as (T_{VE})—the equivalence is the same: 'We are not BIVs' is true iff *we are not BIVs*. But the truth-conditions for the above-mentioned sentence 'We are not BIVs' are evidently different in these two languages. Thus, if we are not BIVs (i.e. if we speak normal English), the truth-conditions are that *we are not BIVs*. If, on the other hand, we are BIVs (i.e. if we speak vat-English), the truth-conditions are that *we are not <Bs> in <Vs>*. The conclusion in the vat-English sense that *we are not <Bs> in <Vs>* obviously misses the point, since we wanted to get to the conclusion that we are not BIVs in the normal English sense. However, in order to reach this particular conclusion, we would have to employ (T) within normal English (with the subscript $_E$), but given the starting disjunctive premise of SA, it turns out that to assume that we are normal English speakers is to assume in advance the point we wanted to prove; namely, that we are not BIVs (cf. Brueckner 2016: 4; Kallestrup 2018: 170).

In his later reconstruction of SA, Putnam applied a disquotation scheme (R) by arguing that from the fact that our word 'water' refers to *water*, with whose instances we are causally connected, it follows that we are not BIVs in the waterless world (1992: 369). Brueckner's (2003) simple version of this argument runs as follows:

(Br1) If we are BIVs, then our word 'water' does not refer to water.
(Br2) Our word 'water' refers to water.
(Br3) So, we are not BIVs.

However, step (Br2) is controversial for two important reasons. First, in order to know that our word 'water' refers to *water* and not to *<W>*, we would have to *know* that our uses of this word are indeed causally linked to the instances of water as a liquid in our normal physical environment, where this knowledge must be *empirical* and, as such, endangered by SH. Second, our uses of the word 'water' refer to *water* only if we are normal English speakers in a normal physical environment, and since this can only be the case if we are not BIVs, we seem to beg the question against the skeptic once again.

Arguably, within a semantically closed language, we use disquotation as a syntactic means by which we present the reference of the terms and the truth-conditions of the sentences containing these terms. The knowledge that we—as normal competent speakers—possess about the role of quotation marks, as well as of the semantic terms ‘refers’ and ‘true’, is indeed a priori in that it allows us to present the reference of any meaningful referring term ‘*m*’ with the scheme (R): “’*m*’ refers to *m*”, and the truth-conditions of any sentence ‘*s*’ with the equivalence (T): “’*s*’ is true iff *s*”. However, the lesson from the Twin Earth thought experiment is that without additional descriptive information about the objects with which our uses of words are causally connected, disquotation is utterly insufficient to determine the reference of expressions or the truth-conditions of the sentences.

Even before the discovery of the molecular structure of liquids, Earthlings and Twin Earthlings could—each in their own language—successfully apply (R_E or R_{TE}): “’water’ refers to water”, and (T_E or T_{TE}): “’This is water’ iff this is water”. Namely, before the discovery of the difference in the molecular structure of that liquid on Earth and on Twin Earth, we were willing to argue that the word ‘water’ both in Earth English and in Twin Earth English has the same reference, and that sentences about water have the same truth-conditions. However, it is worth pointing out that the meaning of the word ‘water’ and the truth-conditions of the sentences about water in Earth English and Twin Earth English did not become different the moment we came to this discovery (see Putnam 1973: 702). Namely, given SE, it is clear that the uses of the word ‘water’ in Earth English and Twin Earth English had different references even before that discovery and that the utterances of the corresponding sentences had different truth-conditions all along. Of course, we were not in a position to detect this difference by applying (R) and (T), which read the same in both languages, but only through empirical research. Therefore, in order to specify this difference, it is necessary to supplement the right side of (R) and (T) with the appropriate descriptions: in (R_E) ‘water’ refers to water as liquid H_2O, and in (R_{TE}) ‘water’ refers to water as liquid XYZ; also, in (T_E) ‘This is water’ is true if this liquid is H_2O, and in (T_{TE}) ‘This is water’ is true iff this liquid is XYZ.

Now, the same lesson applies to the reference of the word ‘water’ in the context of Brueckner’s simple version of SA. One, perhaps not so important, difference with the Twin Earth experiment is that there is no water in the BIVs’ world and that the most suitable candidates for external causes of BIVs’ uses of the word ‘water’ are the computer program features <Ws>. Hence, according to SE, all BIVs’ uses of the word ‘water’ refer to <Ws>. But, from the perspective of both normal English and vat-English, the application of a disquotation (R) to the word ‘water’ provides the same linguistic outcome: “’water’ refers to water”. As such, in order to express the semantic difference between our and BIVs’ uses of the word ‘water’, we need to supplement the right side with an adequate descriptive characterization that distinguishes the objects with which these uses are causally connected in our and BIVs' environments. That is to say, the word 'water' in our environment refers to the instances of such-and-such physical liquid, and in the BIVs' environment, it refers to the computer program feature <W>. The limitations of this strategy are by now more than obvious. Namely, the main difficulty here arises from our utter inability to know what kind of environment we *de facto* inhabit, that is, from the fact that we can never know whether we are normal English-speaking human beings in an environment with physical objects, or whether we are BIVs in a completely empty environment.

Now, let us see if the observation that vat-English is not semantically closed and that the reference and the truth-conditions in it are not disquotational is of any help. It is precisely on this observation that Brueckner (1992) articulates one version of his SA, but he *relativizes* it with respect to normal English as a meta-language. Since we know in advance from SH that BIVs' uses of the word 'water' do not have the same reference as *our* uses of that word, and that the truth-conditions of the BIVs' sentences 'This is water' are not the same as the truth-conditions of *our* uses of that sentence, by applying the disquotation schemes (R) and (T) in our English to BIVs' use of the word 'water' and to their sentence 'This is water' we will not get accurate results; on the right side of the disquotational schemes (R_E) and (T_E), we need to put *<W>* and *this is <W>* instead of the *water* and *this is water* mentioned on the left side. But does this mean that vat-English is not semantically closed *tout court* and that the reference of words and the truth-conditions of the sentences in this language are not disquotational *independently* of our English? In order to provide a satisfactory answer to this question, I think it is instructive to appeal once again to the Twin Earth thought experiment: if we have no reason to question that Twin Earth English is semantically closed and that disquotation works within this language, then there seems to be no reason whatsoever to question that the same is the case with vat-English.

Similar to the impression that we had, following Brueckner, with respect to the semantic difference between our and BIVs' sentences, when we realize that the word 'water' in Twin Earth English refers to the instances of XYZ, at first glance it might seem to us (from the perspective of Earth English as a meta-language) that in Twin Earth English neither the reference of that word nor the truth-conditions of the corresponding sentences are disquotational, for we should represent them as follows: in Twin Earth language, 'water' refers to *XYZ*, and 'This is water' is true iff *this is XYZ*. However, just like Earth English, Twin Earth English has all the linguistic resources that make it semantically closed and allows for disquotation: it contains the semantic terms 'refer' and 'true' as well as quotation marks, as a syntactic means by which we name linguistic expressions.

The difference in the reference of the word 'water' and the truth-conditions of the sentence 'This is water' in normal English and vat-

English have an empirical rather than a linguistic origin: the application of disquotation in both languages yields the same outcome, but the important difference between our and the Twin Earth environment consists in the fact that the named liquid on Earth is H_2O, whereas on Twin Earth it is XYZ. Due to this empirical discovery, we can specify the reference and the truth-conditions by adding the appropriate descriptive characterization on the right side of the disquotation schemes. There is no reason why inhabitants of Earth and Twin Earth should not continue to use the word 'water', as well as the sentence 'This is water', in the same way as they used them before this discovery, and why they could not (each in their own language) successfully apply the disquotation, while at the same time being aware that in their two languages (owing to the difference with respect to their environment) the word 'water' has different references and the sentence 'This is water' has different truth-conditions.

As for the semantic closeness and the applicability of disquotation, it seems that what is true of Twin Earth English is also true of vat-English. According to SH, vat-English (just like Twin Earth English) has all the linguistic resources that make it semantically closed and allows for disquotation: it contains the semantic terms 'refer' and 'true' as well as quotation marks, as a syntactic means by which we name linguistic expressions. The difference with respect to the reference of the word 'water' and the truth-conditions of the sentence 'This is water' in normal English and vat-English has an empirical rather than a linguistic origin: the use of disquotation gives us the same outcome again, but our and BIVs' environments differ in that the uses of the word 'water' in our language sustain a causal connection to the instances of such-and-such physical liquid, whereas in the BIVs' waterless world, it sustains a causal connection with the computer program features <Ws>. The only important, and yet empirical, difference between the Twin Earth and the BIVs' world is that BIVs (unlike the Twin Earthlings) are utterly unable to discover to which particular objects in the environment the word 'water' is causally connected; that is, they are unable to find an adequate descriptive characterization to determine the reference and truth-conditions of their linguistic expressions and sentences. It is especially worth stressing, however, that the semantic closedness of a language, as well as the possibility of applying disquotation in it, should depend only on the linguistic resources, and not on the epistemic position of the speakers. In other words, it is irrelevant for these linguistic features whether we *sometimes* make errors in descriptively identifying objects of reference (as in the Twin Earth experiment), or whether we *always* and *systematically* make such errors (as in the BIVs' world). Thus, in contrast to the Twin Earth thought experiment, the main point of the skeptical BIV hypothesis is that we might be in the BIVs' position after all. If we are not able to exclude this possibility, we will never know for certain that our uses of the word 'water' refer to the instances of such-and-such liquid. Given SE, the most we can know is that the following conditionals "If we are in a normal environment, ’water’ refers to the instances of such-and-such liquid” and “If we are BIVs, ’water’ refers to the computer program features <Ws>” are true. Unfortunately, we cannot know—either a priori (i.e. by means of disquotation) or a posteriori (i.e. by means of sensory evidence)—which of the two antecedents is in fact true; that is, we cannot know whether we are normal human beings speaking normal English or BIVs speaking vat-English.[4]

We have thus seen that disquotation cannot help us to complete Putnam’s SA and, consequently, prove that we are not BIVs. Some semantic externalists (e.g. Tymoczko 1989, Warfield 1998, Brueckner 2003) have used the fact that SE represents the thesis about the content of our thoughts of external objects, and appealed to the assumption that was traditionally considered to be an internalist ally: with respect to the contents of our thoughts, we have immediate, privileged access that allows us to obtain non-evidential, a priori self-knowledge (SK) about *the content* of our thoughts. One version of SA that relies on SK was offered by Warfield (1998: 78):

(Wr1) I think that this is water.
(Wr2) In its waterless world, no BIV can think that this is water.
(Wr3) So, I am not BIV.

As we can see, the first premise of Warfield’s argument relies on SK, and the second on SE. Of course, this particular combination of premises can only be legitimate if these two theses are compatible. Yet, as is well known, Putnam considered SK to be inconsistent with SE,[5] and his opinion was shared by many authors (e.g. McKensey 1991, Bilgrami 1992, Brown 1995, Boghossian 1997, etc.). Some of them, such as Bilgrami (1992), argued that, if SE (along

[4] The same line of reasoning can be applied in order to refute some recent attempts (e.g. Thorpe 2018) to prove that we are not BIVs on the basis of the assumption that the subject has non-empirical semantic knowledge of the content of his current thoughts and that this knowledge can be expressed in the disquotational form (e.g. “My thought ‘This is water’ has the content that this is water”). Given the limitation of space, I cannot provide a detailed analysis of this proposal. See Falvey & Owens (1994) for the point that we cannot have non-empirical knowledge of the comparative content of our thoughts: namely, in order to find out that the content of my thought ‘This is water’ would be different depending on whether I form this thought in the Earth or Twin Earth environment, I would have to find out the difference in the molecular structure of water, which is something that I can only know empirically.

[5] Putnam found a solution to the conflict between SE and SK by bifurcating the content of thoughts into narrower and wider. The thought of an object in its narrowest content is the subject’s conception of the object as an internal psychological state, while in its wider content, this thought is determined by the external relation to the object. Thus, in the Twin Earth thought experiment, when we have a thought expressed by the phrase ‘This is water’, it turns out that we and our phenomenological duplicates share the same *narrow* thought content (i.e. we are in the same psychological state when the instance of the substance we perceive is called ‘water’), but we have different *wide* thought contents (i.e. our thought is *about* a sample of the liquid H_2O, and our Twin Earth duplicate’s thought is *about* a sample of the liquid XYZ).

with its causal constraint) is correct, then in order to possess knowledge of the content of thoughts (as well as to determine their references and truth-conditions), we would have to include a descriptive characterization of objects from the environment with which these thoughts sustain a causal connection. Yet, Bilgrami continues, we cannot know that such characterizations are accurate without a proper (a posteriori) empirical investigation. Others, such as McKinsey (1991), argued that the combination of SE and SK leads to absurdity: according to SK, we should have privileged access to our thoughts and thus know a priori that we are thinking a water-thought; on the basis of SE, on the other hand, we should know a priori that if we have water-thoughts, then water exists. From these two premises, it follows that we should know a priori that water exists. However, given that we do not have privileged access to the outside world, our knowledge of the existence of water must be a posteriori (cf. Kallestrup 2012: 173).[6]

In the light of many discussions on this topic, the ultimate impression is that the incompatibility of SE and SK cannot be eliminated without rejecting or, at least, significantly modifying one of these theses. Thus, for example, Bilgrami (1992) modifies SE by introducing a fundamentally internalist constraint, according to which, in selecting the object in the environment that is supposed to fix the concept that is being expressed by the given term, one has not only to pick the object which is obviously causally correlated with that term but also to *describe* this external determinant of the concept 'in a way that fits in with the other *contents* one has attributed to the agent' (1992: 257). On the other hand, Nuccetelli (2003) modifies SK by distinguishing between two types of a priori knowledge: in the first, stronger sense, the knowledge of a statement is a priori if it is completely independent of empirical assumptions, while in the second, weaker sense, the knowledge of a statement may be a priori even if it includes certain empirical assumptions in light of which that statement can be challenged a posteriori; in her view, self-knowledge about the content of the thought 'This is water' is a priori in the weaker sense, for, according to SE, it rests on the empirical assumption that the term 'water' does have a reference with which it is causally connected—namely, it refers to the instances of H_2O (2003: 180).

Either way, it turns out that neither disquotation nor SK can help us in completing Putnam's SA. If we accept the original (Putnam's) version of SE, we are forced to reject or at least significantly modify SK: without the additional evidential knowledge of the environment and the objects to which our thoughts are causally connected, we cannot fully know the contents

[6] The thesis that Putnam's versions of SE and SK are incompatible could be defended in another, indirect way. Namely, under the essentialist assumption—according to which the molecular structure essentially determines the identity of substances such as water—someone who does not know the molecular structure of water could entertain the thought 'Water is not H_2O'. If both SE and SK are accepted, this thought should be logically inconsistent. Hence, if we want to preserve SE, without declaring that person irrational, we have no other option but to reject SK (Bilgrami 1992).

of these thoughts. The premise (Wr1) in Warfield's argument is therefore problematic and questionable for similar reasons to the premise (Br2) in Brueckner's argument. Namely, in order to know that our thought of water is, indeed, the thought of water as a physical liquid (i.e. H_2O), and not of the computer program features <Ws>, we would have to know that our thought sustains the appropriate causal connection with the instances of H_2O. We cannot gain such knowledge a priori, but only a posteriori; that is, only through empirical research and on the basis of the appropriate sensory evidence. By formulating SH, however, the skeptic eliminates our possibility of having such evidence: on whatever sensory evidence we base our belief that our thought about water sustains a causal connection to the instances of H_2O, we cannot know for sure that this thought does sustain such a causal connection, for we cannot rule out the possibility that we are BIVs who have the same sensory evidence and the same (though false) beliefs about our environment. That is to say, just as we are convinced that we have a thought of water as a physical liquid, BIVs can be convinced (though wrongly) that they have a thought of water as a physical liquid. If, despite our inability to know this, we endorse (Wr1)—thereby implying that we have the thought of water as an instance of H_2O—we assume in advance that we are in a normal physical environment and, ultimately, beg the question against the skeptic.

4. Self-Refuting Character of the BIV Hypothesis

As we have seen in the previous section, both Brueckner's and Warfield's externalist attempts to complete Putnam's SA and, consequently, reach the conclusion that we are not BIVs have failed. Since there does not seem to be any third way to accomplish these goals, the ultimate reach of Putnam's argument against skepticism is the meta-linguistic conclusion that SH in the form of the sentence 'We are BIVs'—when it is claimed or considered (as is obviously the case in the context of the skeptical argument)—must be false. Now, where does this leave the skeptic?

Assuming that SE is correct and that SH is formulated as Putnam proposes, it turns out that in making his argument, the skeptic is forced into a somewhat precarious position. Namely, by presenting the possibility of a mistake that we seemingly cannot exclude, the skeptic must take the statement 'We are BIVs' in the normal English sense; that is, in the sense in which the words 'brain' and 'vat' refer to actual *brains* and *vats*. As such, the skeptic implies that we are normal English speakers (i.e. that we are not BIVs), thereby acknowledging the point of Brueckner's premise (Br2), as well as of the premise (Wr1) in Warfield's version of SA. If the skeptic starts with the assumption that the 'We are BIVs' hypothesis is true, she undermines the possibility of asserting this hypothesis in the sense she originally intended. In this case, the skeptic herself would be BIV without any causal contact with real brains and real vats; in other words, she would be in a position in which she could only assert the sentence 'We are BIVs' in vat-English, and her words 'brain' and 'vat' would only refer to the computer program features <Bs> and <Vs>. Putnam's disjunctive version of SA points out this skeptic's predicament: whether or not we are BIVs, the 'We are BIVs' hypothesis is shown to be false.

Still, it is important to bear in mind the following restriction that Putnam himself invokes: *whenever we are considering* whether SH is true or false, it follows that it must be false. This restriction creates theoretical space within which the radical skeptic can find at least some sort of escape route. Namely, Putnam's SA proves only the unassertiveness (but not falsehood) of SH: expressed by the sentence 'We are BIVs', SH cannot be truly asserted by us, although the proposition expressed by this sentence might be true nevertheless. Put otherwise, the proposition expressed by SH is in itself perfectly consistent, but when we assert it, we contradict ourselves. Why is this?

Let us once again recall Putnam's observation about the self-refuting character of SH. He reminds us that there are two groups of self-refuting statements (Putnam 1981: 7–8). The first group includes statements that are self-refuting on the basis of their semantic content, regardless of whether anyone asserts them; the statement that falls into this group is the general statement 'All statements are false' which, if true, must be false. As is well known, in a semantically closed and universal language such statements give rise to semantic paradoxes (such as those of the Liar family), for they depend upon the semantic notion of truth and on explicit self-reference (i.e. the sentence refers to itself). In other words, they are self-refuting only due to their semantic content and self-reference, regardless of *linguistic* factors (for instance, the presence of terms that indicate *who* and in *what* circumstances *asserts* them) or *theoretical* factors (such as specific assumptions about the meaning of some terms that occur in a statement). As an example of self-refuting statements that fall into the second group, Putnam cites the statement 'I do not exist'. This sentence is understood to be self-refuting due to its semantic content. But since the indexical term 'I' introduces an element of self-reference, it is obvious that its self-refuting status depends also on *who* and in *what* linguistic form *asserts* the proposition expressed by this sentence: the statement must be false only if it is asserted by the speaker (or speakers) in the first person, present tense. Yet, note that the same proposition *about my non-existence* may be truly asserted by someone else (e.g. 'He (Živan Lazović) does not exist'), or even by me in some other (past or future) tense (e.g. 'I did not exist' or 'I will not exist').

So, despite the fact that the sentence 'I do not exist' is necessarily false when I assert it in the present tense, it does not follow that the *proposition* expressed by this sentence cannot be true after all. The first type of statements, whose self-refuting character depends solely on their semantic content, will be characterized as self-refuting in the *strong* sense, while the second type of statements, whose self-refuting status depends partly on semantic content and partly on linguistic form (i.e. presence of particular expressions that introduce self-references by pointing out *who* and in *what* circumstances asserts them) or on some specific theoretical assumptions (such as the SE thesis), will be characterized as self-refuting in the *weak* sense. I think we should concede Putnam's claim that SH—presented in the form of the statement 'We are BIVs'—is self-refuting in the weak sense.

I will show in the remainder of this paper that the self-refuting character of SH rests partly on its semantic content (which is conditioned by the SE assumption about the meaning of the words 'brain' and 'vat') and partly on the fact that it is asserted in the first person form. I will also show that SH cannot be asserted by *any* human being, but that the proposition expressed by it might be true nevertheless. In this respect, the weak self-refuting status of the statement 'We are BIVs' is no exception. Without going into the exhaustive analysis, I will compare this statement with similar statements whose self-refuting character is partly dependent on their semantic content and partly on additional theoretical (conceptual) assumptions or the use of certain linguistic terms that make them unassertible relative to some particular speaker. Each of the following examples will be accompanied by brief remarks that should account for the fact that those who make such sentences contradict themselves without uttering a contradiction and, consequently, help us clarify the unassertibility of SH.

As an example of semantic paradoxes, the first sentence belongs to the well-known Liar family:

(1) 'All humans are lying.'

It is clear why, in a semantically closed and universal language, this sentence is self-refuting: it depends on the semantic notion of truth, the element of self-reference provides the universal quantifier 'all', and the sentence is unassertible relative to any speaker who belongs to the class—i.e. the class of human beings—about which it talks. Note, however, that this sentence might be consistently and truly asserted by any non-human being, as well as that the proposition expressed by this sentence might be true even if it is not asserted by anyone.

The so-called *Moorean paradox* provides us with yet another interesting example:

(2) '*p*, but I do not believe that *p*.'

Although I can assert *p* at some particular moment and add that I do not believe that *p* at some other moment, it seems that if I simultaneously say '*p*' and 'I do not believe that *p*', I contradict myself. Admittedly, there are numerous interpretations of this paradox in the relevant literature, but according to one of the most popular accounts, assertion and belief are directed to truth in such a way that to assert *p* is to *express* or *imply* the belief that *p* is true. Hence, whoever asserts *p* and conjoins it with the assertion that she does not believe that *p*, obviously contradicts herself in the sense that she believes that *p* and she does not believe that *p*. It is worth stressing, however, that the self-refuting status of this sentence also depends on its formulation in the first person, present tense. Even if it is true that one would not assert that *p* without believing that *p*, one's belief that *p* does not imply *p*. Given that this is so, no problem will occur with the second and third person counterparts of (2)—e.g. '*p*, but *you* do not (*she* does not) believe that *p*'—or with the sentence in the first person past tense, such as '*p*, but I did not believe that *p*'. So, in spite of being unassertible in the first-person present tense, the Moorean sentence expresses a consistent proposition which might be true. Put differently, it may well be that *p*, and—just like in the example with 'I do not exist'—it might be true *about* me (or us) that I (or we) do not believe that *p*.

The third and even subtler example—closely related to some responses to skepticism—is the so-called *abominable conjunction* (see DeRose 1995):

(3) 'I know that I have hands, but I do not know that I am not BIV.'

It seems that even this sentence cannot be asserted in the first person without falling into contradiction. This is so because of the conceptual connection between knowledge and truth, and because of the fact that the statement 'I have hands' implies that I am not a (handless) BIV. This is analogous to the Moorean paradox in all important respects. Namely, if I claim to know that I have hands in the first conjunct, I thereby imply that I am not BIV, whereas in the second conjunct, I explicitly question this implication by allowing the possibility that I am BIV. The inconsistency becomes even more obvious if we assume the principle of deductive closure: from my knowing that I have hands and my knowing that having hands implies that I am not BIV, it should follow that I also know that I am not BIV. Thus, if I am willing to acknowledge that I do not know the implied statement, it follows by modus tollens that I must refrain from claiming to know the antecedent. This point is especially important given that the Cartesian versions of the skeptical argument—including the Putnamian BIV version of this argument—seemingly rest on the principle of deductive closure.

The self-refuting character of this conjunction, however, depends on the particular conception of knowledge that we assume. Thus, for invariantists and infallibilists—e.g. Descartes and Peter Unger respectively—this conjunction is self-refuting in the strongest sense, for both the first and the third person formulation (e.g. 'She knows she has hands, but she does not know that she is not BIV') sound like a contradiction in the same way as 'I know (She knows) that *p*, but it is possible that not-*p*'. Yet, for someone who is a fallibilist, and especially for those who (like Dretske and Nozick) reject the deductive closure principle, the third person formulation of the conjunction will be perfectly consistent. However, it seems that even for these authors, there is a problem with the first-person version of the conjunction 'I know that I have hands, but I do not know that I am not BIV', for it looks self-refuting for a similar reason to the Moorean paradox: the problem is that the speaker negates the point implied by her first conjunct by asserting the second conjunct.

Mark Heller (1999) has shown that the abominable conjunction could be interpreted as self-refuting in the weak sense within a variantist conception of knowledge such as conversational contextualism. As is well known, the central thesis of conversational contextualism is that the concept of knowledge is context-sensitive in the sense that knowledge attributions of the form '*S* knows that *p*' can express different propositions (and thus have different truth-values) depending on the attributor's conversational context. These changes occur because knowledge attributions in different contexts can apply different—i.e. lower or higher—standards for knowledge. Contextualists explain the change in epistemic standards mainly by relativizing Drecke's idea of relevant alternatives with respect to the knowledge attributors: when we evaluate whether someone in a given context knows some statement *p*, we expect that person to exclude (*ceteris paribus*) all those alternatives (i.e. possibilities of error) which we consider relevant in this context. The contextual change of conversational factors—such as intentions, needs or interests of the knowledge attributors—results in the narrowing or widening of the set of relevant alternatives; that is, it results in lowering or raising the standard of knowledge. According to most contextualists, including Heller, in order to make an alternative *relevant* to the assessment of knowledge in a given context, it is sufficient to pay attention to it (Lewis 1996) or to make it salient (Cohen 1988). Thus, for instance, in everyday contexts of knowledge attribution, the epistemic standards are relatively low, which means that remote alternatives—e.g. various skeptical possibilities of error—are not taken into account. These skeptical possibilities of error, however, become relevant in the philosophical context. As such, our knowledge attributions in everyday contexts will (*ceteris paribus*) express truth, whereas in the philosophical context, they will express falsehoods due to the fact that skeptical hypotheses make salient precisely those alternatives which we are unable to exclude.

It is worth noting that, since Heller—like other conversational contextualists—has in mind knowledge attributions from the third-person perspective, the assertion of the abominable conjunction should be self-refuting regardless of whether we (as speakers) are attributing (or denying) knowledge to someone else or to ourselves. What is important is that, if the contextualist explanation is correct, we will be able to truthfully claim in ordinary contexts that we know (*ceteris paribus*) that we have hands in spite of not knowing that we are not BIVs. On the other hand, any emphasis of the possibility that we are BIVs would shift us into the skeptical context, in which—given that we cannot know that we are not BIVs—it will not be true to know that we have hands. The resemblance to the Moorean paradox is now apparent. Each part of conjunction (3) can be asserted independently (of course, in different contexts). Yet, by asserting them simultaneously, we fall into contradiction: if we ascribe to ourselves (or to someone else) the knowledge of having hands, and at the same time maintain that we do not know that we are not (handless) BIVs, we make this skeptical alternative relevant and create a skeptical context, wherein no one can know that she has hands, because no one can exclude the possibility of being a (handless) BIV. As Heller concludes, this point 'explains the abominableness of DeRose's abominable conjunction' and 'makes the conjunction true but unassertible' (Heller 1999: 204). Thus, according to this explanation, the conjunction is true in everyday contexts of knowledge attribution in which *no one asserts it*, but it is unassertible relative to the participants in conversational contexts, for its assertion by any speaker in any context would turn the given context into a skeptical one and, thereby, make it false.

It is easy to see, I think, that all three examples considered above express consistent propositions in the same way as Putnam's example 'I do not exist', and become false only when they are asserted by some particular speaker; that is, when they are asserted in the first person, by the members of the class about which the proposition states something, or by the participants in the conversational context. The propositions expressed by these sentences are, therefore, unassertible relative to particular speakers, although they can be true nevertheless. Given that the same point applies to SH, it is clear that this hypothesis falls into the group of self-refuting statements in the weak sense. I will explain why this is so by drawing a comparison between SH and each of the three examples stated above.

The self-refuting character of Putnam's SH is certainly influenced—as was the case with (1)—by the element of self-reference invoked by the first person, present tense formulation; recall that such self-reference was also observed in the example 'I do not exist'. With respect to the version of SA that invokes the self-knowledge (SK) thesis, it is imperative that the argument is formulated in the first-person by using the indexical expression 'I' (see Wright 1992: 76–7; Kallestrup 2012: 172–3). In the context of Putnam's version of SA, however, the first-person formulation is not mandatory. By formulating SH, the skeptic addresses us as human beings and presents us with the possibility that we are BIVs, which points to a fatal flaw in our epistemic position and, ultimately, compromises our knowledge of the external world. The skeptic can therefore formulate SH in the form of the universal statement 'All humans are BIVs' in which—given that the property of being BIV is ascribed to all members of that class—self-reference occurs if the statement is asserted by any member of the class of human beings (this is yet another similarity to (1)). In this formulation, it is clear that the statement is unassertive relative to *us* as human beings. As such, some other (non-human) being could assert this sentence without any problems, and the proposition expressed by this sentence could be true even if no being asserts it.[7] Putnam's SH is, therefore, self-refuting when it is asserted in the first person, or when it is asserted in the form of a universal statement by any speaker who belongs to the class of human beings, since in both cases it contains self-reference.

We have seen that in (2) the speaker finds herself in a paradoxical situation, for by asserting the second conjunct, she negates what is implied by the first conjunct. By asserting SH in either of the two mentioned forms, the speaker makes the statement false by undermining a necessary condition—i.e. the existence of a proper causal link—under which her words 'brain' and 'vat' can have the desired meaning in the context of considering the skeptical argument; i.e. she makes it impossible for her words to refer to *real* brains and *real* vats. Something similar occurs in the example 'I (we) do not exist'. Namely, when uttering a statement, it is a necessary condition for the proper use of the pronoun in the first-person present tense that the speaker exists at the moment of utterance. So, by denying her own existence in the second part of the sentence, the speaker denies the fulfillment of that necessary condition and thus falls into contradiction.

Finally, SH has in common with (3) that a particular theoretical assumption concerning the meanings of the used terms is crucial for its unassertiveness. In the case of (3), it is a contextualist assumption about the meaning of the concept of knowledge, whereas in the case of SH, it is an externalist assumption about the meaning of referring terms such as 'brain' and 'vat'. We have seen that the (externalist) causal constraint on the reference in the context of Putnam's disjunctive SA leads to the conclusion that the words 'brain' and vat' have different references depending on whether or not we are BIVs: in the second case, these words refer to brains and vats as physical objects, while in the first they refer to the computer program features <*Bs*> and <*Vs*>. We have also seen that in order to derive her conclusion, the skeptic is forced to demonstrate that SH is true in the normal English sense. But due to the causal constraint on the reference and the element of self-reference in both formulations of SH, it follows that the skeptic—when considering whether SH is true either in the first person or as a member of the class SH refers to—finds herself in a paradoxical position. Namely, it turns out that when the skeptic asserts 'We (all humans) are BIVs', she does not assert a true proposition in the normal English sense (and does not mean that we are *real* brains in *real* vats), but rather a quite different, and ultimately false, proposition that we (humans) are <Bs> and <Vs>.

[7] Although SH is construed in such a way to include the assumption that, aside from BIVs, the actual world does not contain any other intelligent beings that could claim that *we* (i.e. human beings) are BIVs, it does not mean that some (logically) possible world in which such intelligent beings exist, and in which they could truly assert the proposition 'All humans are BIVs', is inconceivable. I am strongly inclined to think that this is the status of SH that Crispin Wright had in mind when he made his remark that Putnamian SA does not refute the skeptical nightmare (1992: 73, 93); cf. Kallestrup (2012: 173).

5. Conclusion

On the basis of previous considerations, I think it is safe to say that Putnam was essentially right in claiming that the BIV hypothesis is self-refuting in the weaker sense. However, such self-refuting status has been shown to imply merely the unassertiveness of this hypothesis and is only relative to *us* as normal (human) speakers. In other words, Putnam's SA only proves that the sentence or statement 'We are BIVs' must be false when it is *us* (humans) who question its truthfulness. For this reason, SA has limited reach against the skeptic. The same consistent proposition that is expressed by the sentence 'We are BIVs' can be expressed in some other grammatic form—e.g. 'All humans are BIVs' or 'You are BIVs'—and it may be truthfully asserted or considered by some other (non-human) being, and can even be true when it is not asserted.

In the end, I should clarify that my intention in this article was not to provide a philosophical defense of the skeptic's position. Yet, when faced with the inability to consistently think or assert the truth of SH, the skeptic can do one of the following things. First, she can appeal to some version of the skeptical hypothesis that is completely beyond the reach of Putnamian SA. Second, she can acknowledge Putnam's SH and find at least some sort of satisfaction in the fact that it is still possible for this hypothesis to be true. Given that Putnam's SA was not successfully completed by any of its subsequent versions, the main point of the skeptical argument persists: it is indeed possible that we are BIVs, and there does not seem to be any theoretical or empirical strategy to exclude this possibility. Unfortunately, I have to admit that the same conclusion seemingly applies to the thoughts expressed in this paper. For, despite the fact that I cannot truthfully say or consistently think that *I am* (and have always been) a BIV, the proposition 'Živan Lazović is a BIV' may be true after all. But if this proposition is true, and if SE represents the correct view, then the words and thoughts expressed in this paper ultimately concern computer program features instead of real objects in the external world.[8]

References

Bilgrami, A. 1992. Can Externalism Be Reconciled with Self-Knowledge? *Philosophical Topics*, 20: 233–268.

Boghossian, P. 1997. What the Externalist Can Know A Priori. *Proceedings of the Aristotelian Society*, 97: 161–175.

Brown, J. 1995. The Incompatibility of Anti-individualism and Privileged Access. *Analysis*, 55: 149–156.

[8] I wish to thank the anonymous referees for their very helpful comments on earlier drafts.

Brueckner, A. 1986. Brains in a Vat. *Journal of Philosophy*, 83: 148–167.

–, 1992. Semantic Answers to Skepticism. *Pacific Philosophical Quarterly*, 73: 200–219.

–, 2003. Trees, Computer Program Features, and Skeptical Hypotheses. In Stern, R. (ed.), *Transcendental Arguments: Problems and Prospects*, Oxford: Clarendon Press, pp. 152–162.

–, 2016. Skepticism and Content Externalism, *Stanford Encyclopedia of Philosophy* (Winter 2016 Edition), Edward N. Zalta (ed.), URL = https://plato.stanford.edu/archives/win2016/entries/skepticism-content-externalism/

Cohen, S. 1988. How to Be a Fallibilist. *Philosophical Perspectives*, 2: 91–123.

DeRose, K. 1995. Solving the Skeptical Problem. *The Philosophical Review* 104: 1–52.

Dretske, F. 1970. Epistemic Operators. *The Journal of Philosophy*, 67: 1007–1023

–, 1981. The Pragmatic Dimension of Knowledge. *Philosophical Studies,* 40: 363–378.

Falvey, K., & Owens, J. 1994. Externalism, Self-Knowledge and Scepticism. *The Philosophical Review*, 103: 107–137.

Folina, J. 2016. Realism, Skepticism, and the Brain in a Vat. In Goldberg, S. C. (ed.), *The Brain in a Vat*, Cambridge: Cambridge University Press, pp. 155–173.

Heller, M. 1999. Relevant Alternatives and Closure. *Australasian Journal of Philosophy*, 77: 196–208.

Kallestrup, J. 2012. *Semantic Externalism*. Routledge.

Lewis, D. 1996. Elusive Knowledge. *Australasian Journal of Philosophy*, 74: 549–567.

McKinsey, M. 1991. Anti-Individualism and Privileged Access. *Analysis*, 51: 9–16.

–, 2018. Skepticism and Content Externalism, *The Stanford Encyclopedia of Philosophy* (Summer 2018 Edition), Edward N. Zalta (ed.), URL = https://plato.stanford.edu/archives/sum2018/entries/skepticism-content-externalism/

Nuccetelli, S. 2003. Knowing That One Knows What One Is Talking About. In Nuccetelli, S. (ed.), *New Essays on Semantic Externalism and Self-Knowledge*, Cambridge Mass: The MIT Press, pp. 169–184.

Nozick, R. 1981. *Philosophical Explanations*. Harvard University Press.

Putnam, H. 1973. The Meaning of 'Meaning'. *The Journal of Philosophy*, 70: 699–711.

–, 1981. *Reason, Truth and History*. Cambridge University Press.

–, 1992. Replies. *Philosophical Topics: The Philosophy of Hilary Putnam* 20 (1): 347–408.

Thorpe, J. R. 2019. Semantic Self-Knowledge and the Vat Argument. *Philosophical* Studies, 176: 2289–2306.

Tymoczko, Th. 1989. In Defense of Putnam's Brains. *Philosophical Studies*, 57: 281–297.

Warfield, T. A. 1998. A Priori Knowledge of the World: Knowing the World by Knowing Our Minds. *Philosophical Studies*, 92: 127–147.

Wright, C. 1992. On Putnam's Proof that We Are Not Brains in a Vat. *Proceedings of the Aristotelian Society*, 92: 67–94.

Erratum

In: Miller, A. and Surgener K., (2019). Ecumenical Expressivism and the Frege-Geach Problem, *BPA* 32, 7–25, the text on pp. 9–10 should read as below:

2. Blackburn's Quasi-Realist Expressivism and the Frege-Geach Problem

The fundamental expressivist ideas are that we give an account of the meaning of a sentence in terms of the state of mind that it expresses and that in the case of a moral sentence such as "Murder is wrong" the relevant state of mind is a non-cognitive attitude of disapproval of murder: B!(murder).[1] These ideas, however, leave the expressivist with a problem. While it is plausible to think of the meaning of "Murder is wrong" as it appears in an asserted context such as e.g.

(1) Murder is wrong

in terms of B!(murder), it is difficult to see how this account can be extended to cover the appearance of "murder is wrong" as it appears in an unasserted context such as the antecedent of (2):

(2) If murder is wrong then getting Peter to murder people is wrong,

since someone sincerely asserting (2) needn't have an attitude of disapproval towards murder (or indeed towards getting Peter to murder people) – think of how those who approve of helping the aged can still sincerely utter "If helping the aged is wrong then getting Peter to help the aged is wrong". If this extension turns out not to be possible it looks like the inference from (1) and (2) to

(3) Getting Peter to murder people is wrong

will be vitiated by a fallacy of equivocation, since "Murder is wrong" will have different meanings as it appears in (1) and in the antecedent of (2). And this is highly problematic, as the inference is an instance of Modus Ponens, a valid inference form.[2] This is the Frege-Geach

[1] Ridge characterises expressivism as a form of "ideationalism", where "Ideationalism maintains that facts about the semantic contents of meaningful items in a natural language are constituted by facts about how those items are conventionally used to express states of mind" (2014: 107). For an account of the philosophical motivations for expressivism – in metaphysics, epistemology and moral psychology – see chapters 3–5 in Miller (2013).

[2] Notice that it will not do for the expressivist to simply accept that this aspect of moral discourse is in bad faith: as we noted above the problem in this area extends to most of moral reasoning. Going down this road would leave the expressivist with an account of

Problem, and the challenge to the expressivist is therefore to give an account of the contribution made by the meaning of a moral sentence to the meaning of a more complex sentence in which it appears in terms of the state of mind it expresses when used in an asserted context, in such a way that intuitively valid inferences involving it are not impugned (by, for instance, the commission of fallacies of equivocation).

the meaning of positive, atomic, moral statements but not much else. At this point it is unclear why developing expressivism is preferable to simply adopting an error theory.

CIP – Каталогизација у публикацији
Народна библиотека Србије, Београд

1

FILOZOFSKI godišnjak = Belgrade philosophical annual / editor Slobodan Perović. – God. 1, br. 1 (1988)– . – Belgrade : Institute of Philosophy, Faculty of Philosophy, 1988- (Belgrade : Službeni glasnik). – 24 cm

Godišnje. – Glavni stvarni naslov od br. 28 (2015) Belgrade philosophical annual. – Tekst na engl. jeziku.

ISSN 0353-3891 = Filozofski godišnjak

COBISS.SR-ID 15073792