Croatian Journal of Philosophy
Vol. XXIII, No. 67, 2023
https://doi.org/10.52685/cjp.23.67.4
Received: July 22, 2022
Accepted: April 10, 2023

Imagination, Thought Experiments, and Personal Identity

MICHAEL OMOGE
University of Alberta – Augustana, Camrose, Canada

Should we descry the nature of the self from thought experiments? Shaun Nichols says 'maybe,' but only if we use thought experiments that do not recruit the indexical "I" (non-I-recruiting). His reason is that the psychology of "I" perforce mandates that imagination responds to thought experiments that recruit it (I-recruiting) peculiarly. Here, I consider whether he is correct about non-I-recruiting personal identity thought experiments. I argue positively using the same framework, i.e., considering the underlying psychology.

Keywords: Propositional imagination; cognitive architecture; personal identity; thought experiments.

1. *Introduction*

In no area of philosophy are thought experiments more used than in personal identity, and yet, in no area are they disparaged than in personal identity. One general reason personal identity thought experiments (PITEs) are said to fail is that propositional imagination[1] (hereafter, simply as 'imagination') breaks down in them. But if so, then this breakdown of imagination in PITEs must be traceable to some

[1] Propositional imagination is a propositional attitude that has linguistically expressed content. It is often contrasted with experiential imagination, which involves consciously entertaining mental imagery. By only talking about propositional imagination here, I do not mean that experiential imagination is not involved in thought experiments, but rather that talk of cognitive architecture—which turns on drawing a similarity between imagination and belief—is often taken to mean that experiential imagination is excluded. But see Omoge (Forthcoming) for how to include it.

faults in the 'cognitive architecture' of imagination. Where cognitive architecture "is a theory about the mind at the functional—as opposed to, say, neurological or biological—level that aims to explain relevant psychological phenomena [by] (literally) drawing out the functional connections between various components of the mind" (Miyazono and Liao 2016: 234).

Shaun Nichols (2008) notices this link between the failure of imagination in PITEs and the cognitive architecture of imagination, deploying the link to expose a shortcoming in how imagination responds to PITEs that recruit the indexical "I" (I-recruiting PITEs). Nichols focuses on Bernard Williams' (1970, 1973) modification of the Lockean body-swap PITE where the psychological properties of person A are transferred to person B. Nichols' diagnosis of why imagination breaks down in this (and other) I-recruiting PITE is that at the psychological level, "I" is semantically impoverished in that it does not come with all the historical details that characterize the speaker of the I-token. He adds that while this poverty renders "I" flexible such that there are no obstacles to imagining scenarios that recruit it, the flexibility makes it possible for an agent to imagine that *I am someone else* even when their defining psychological characteristics are destroyed, which is problematic. Thus, he concludes that we should not use I-recruiting PITEs to draw metaphysical conclusions about the self. He, however, suggests that non-I-recruiting ones may be so used.

My goal in this paper is to consider whether Nichols is right about non-I-recruiting PITEs: do they succeed in leading us to what is essential about the self? It is important to consider this question because if imagination also fails in non-I-recruiting PITEs, such that they, like their I-recruiting counterparts, fail to lead us to what is essential about the self, then that would be the final nail in the coffin for PITEs in general. Philosophers would have been doing something terribly wrong by relying on them. Although things are not so straightforward, I will argue here that Nichols' optimism is warranted. Non-I-recruiting PITEs succeed in leading us to appropriate metaphysical conclusions about the nature of the self.

I begin by discussing the cognitive account of imagination Nichols relies on (Section 2). I then rehearse how he uses the account to show why we should not infer the nature of the self from I-recruiting PITEs, but that we may from non-I-recruiting ones (Section 3). In Section 4, I explain why the cognitive account of imagination Nichols relies on is not straightforwardly compatible with non-I-recruiting PITEs; so, I give an updated version. In Section 5, I show that the updated account is compatible with non-I-recruiting PITEs. In Section 6, I use this compatibility to show why Nichols' optimism about non-I-recruiting PITEs is not misplaced.

But before I begin, let me give some examples of non-I-recruiting PITEs to clarify the scope of the discussion in this paper. Non-I-recruiting PITEs include but are not limited to the original Lockean

body-swap, Parfit's (1984) fi ssion (where the brain of one of an identical triplet is split into two and put in the bodies of the two members of the triplet), Parfi t's Russian (where a Russian lost his memory but he had already told his wife to share his belongings if that ever happens), Parfi t's teleporter (where someone is broken down into molecules and reassembled somewhere else), and their many variants by other theorists. Though I will only focus on fi ssion in this paper, what I will say about it will generalize to all non-I-recruiting PITEs.

2. *The cognitive architecture of imagination*

Nichols (2008) relies on Nichols and Stich's (2003) cognitive account of imagination to show why imagination behaves peculiarly in I-recruiting PITEs. According to Nichols and Stich, the cognitive architecture of imagination comprises an 'imagination box,' which is a workspace and storage unit where imaginings are temporarily stored and manipulated, a 'script elaborator' that generates and embellishes imaginings, and an 'UpDater' that enables reasoning with imaginings. For Nichols and Stich, these cognitive structures help to explain what happens when we practically imagine, for instance, in pretense and mindreading.

In pretending to have a tea party, the representation *We are going to have a tea party* is generated as the imagination premise by the script elaborator and placed in the imagination box. The content of the belief box is then (copied and) put inside the imagination box as further premises. The UpDater then fi lters out the beliefs that are incompatible with the imagination premise. Since what is left after this fi ltering would be insuffi cient to yield the target imagining, Nichols and Stich say that some of the unfiltered-out beliefs contain 'scripts' (e.g., a script for how tea parties typically unfold), where scripts are psychological paradigms that describe appropriate sequences of events in a particular context (Schank and Abelson 1977). Since scripts are unrestrictive—actors often go off-script, improvising their acts—the script elaborator teases out elaborations on the sequences of events detailed by scripts. For Nichols and Stich, this is how imagination operates psychologically.

One component of this account is that imagination interacts with the same inference mechanisms with which belief interacts. This, according to Nichols and Stich, is why the UpDater, which is part of our inference mechanisms, is also at work in belief episodes. We update our beliefs all the time without needing to upend everything we know. In short, imagination and belief are in *the same code*, i.e., they have the same contents and logical form, and they interact with the same inference mechanisms. Put differently, inference mechanisms will treat imagination and belief in much the same ways. Nichols (2004) calls this component of the cognitive account of imagination the 'single code hypothesis.' Nichols (2008) thinks this hypothesis holds the secret to why I-recruiting PITEs should not be used to infer the nature of the self.

3. *Nichols on I-recruiting and non-I-recruiting PITEs*

Nichols' goal is to explain why imagination responds in the way Williams (1970, 1973) describes. According to Williams, imagining the Lockean body-swap PITE from a 1st person rather than a 3rd person perspective (i.e., turning it into an I-recruiting PITE) problematizes the psychological accounts of personal identity (e.g., Parfit 1984). When imagined from the 1st person's perspective and adding the constraint that one of the swapped bodies would be tortured after the swap, Williams argues that the imaginer would lack one vital respect. The respect of knowing "what was going to happen—torture, which one can indeed expect to happen to oneself, and to be preceded by certain mental derangements as well" (1970: 168). Lacking this respect, William concludes, suggests that the imaginer survives the destruction of their psychological properties, contradicting the psychological accounts of personal identity.

Nichols thinks the reason imagination responds to I-recruiting PITEs in this way "turn on peculiar features of imagining with indexicals" (2008: 521). What peculiar features? According to him, to accommodate indexicals in psychology, an internal mental symbol that corresponds to their semantics must be postulated. Now, the semantics of indexicals is not determined by contents. People with Alzheimer's disease, for example, use "I" frequently and appropriately, even in the late stages of the disease. Likewise, you can wake up in the dark with (a temporary) total amnesia and still be able to think *I have a headache*. Rather, the semantics of indexicals "is determined […] by the sparse character ('the speaker of this token of "I"') plus the context" (Nichols 2008: 523). Nichols calls the internal mental symbol that corresponds to this impoverished semantics of indexicals the 'I-concept.' He then argues that the I-concept is why imagination responds peculiarly to I-recruiting PITEs.

Since inference mechanisms respond to the format, not (simply the) denotation of representations (Fodor, 1987), Nichols says that inference mechanisms will respond to the format of indexical representations, i.e., the I-concept. If so, then the poverty of the I-concept explains why there is no limitation to imagining with the *I*, not even when your psychological properties are destroyed: "In particular, the fact that all of my distinctive psychological properties are gone is no obstacle whatsoever. Given the poverty of [the I-concept], there is no constraint against the representation *I exist in this location with completely different psychological properties*" (Nichols 2008: 527). But once it is clear why we can imagine with the *I*, even with different psychological properties, it becomes clearer that we must be careful about what we make of the imagined I-scenarios.

Given the single code hypothesis (Section 2), inference mechanisms interact with the I-concept in the belief context in much the same way they interact with it in the imagination context. However, in the belief

context, there is no problem arising from the poverty of the I-concept. When I wake up in the dark with total amnesia, there is still a plausible sense in which I am the referent of the I-concept, perhaps because my psychological properties still subsist, although I have no conscious access to them at the time. But there is no such sense in the imagination context under discussion (i.e., I-recruiting PITEs) precisely because my psychological properties are now destroyed, and so imagining that I persist in their absence is problematic. Consequently, Nichols warns:

> Thus, it is dangerous to draw any metaphysical conclusions from these imaginative exercises with the *I*. More generally, we should be exceedingly wary of trying to descry the nature of the self through thought experiments that invoke the *I*. Imagining with the *I* sends us on wild thought experiment rides, but the resulting intuitions are likely not a reliable guide to what the self *really* is. (2008: 529, original italics)[2]

While I-recruiting PITEs may be unreliable guides to metaphysical conclusions about the self, Nichols signals that non-I-recruiting ones may fare better: "If we are to use thought experiments to assess what is and isn't essential to the self, we would do well to exclude the cases that trade on the I-concept" (2008: 529). This optimism, however, will not get off the ground unless some of Nichols' other commitments are addressed.

4. *Metaphysical modality and the cognitive architecture of imagination*

Elsewhere (Nichols 2006a), Nichols argues that Nichols and Stich's cognitive account shows that imagination is an unreliable guide to metaphysical modality. Given the single code hypothesis, which suggests that inference mechanisms will balk at contradictions in the belief context, it follows that they will also balk at contradictions in the imagination context. This, Nichols says, is why we face imaginative blocks when we attempt to imagine metaphysical impossibilities,[3] leading him to the conclusion that imagination is an unreliable guide to metaphysical modality. Imagination's natural domain is practical modalizing (e.g., pretense), not metaphysical modalizing (e.g., personal identity).

But if so, then non-I-recruiting PITEs, like their I-recruiting counterparts, will become unreliable guides to metaphysical conclusions about the self as well, although for different reasons. Where I-recruiting PITEs are unreliable because the psychology of the *I* does not mix

[2] Outside PITE, Williams also raises a puzzle for imagining in the 1st person perspective—namely, why is it much easier to imagine that *I am Napoleon* than imagine that *Someone else is Napoleon*? Nichols also responds to this puzzle. I will say something about his response later in Section 6.

[3] Beyond metaphysical modalizing, Nichols also uses the same argument to explain why we face imaginative resistance in fiction (Nichols 2004, 2006b), and why we face difficulty in imagining our own nonexistence (Nichols 2007).

well with imagination, non-I-recruiting ones will be unreliable because they are not the natural domain of imagination. In short, as things stand, non-I-recruiting PITEs are not compatible with the cognitive architecture of imagination. Thankfully, Omoge (2021) has shown that Nichols' skepticism about using imagination to metaphysically modalize is unwarranted. Though his argument is layered, I will recap the relevant aspects here, and together with a caveat I will add later, I will argue that non-I-recruiting PITEs are not incapacitated by the cognitive architecture of imagination.

Central to Omoge's view is ascribing a larger role to scripts than Nichols and Stich do. He argues that scripts are (1) activated conceptually given the imaginer's theoretical assumptions such that a script type is rarely similarly tokened by two imaginers and (2) often compositional given the debate's etiology such that the manner of their composition explains how the imaginers get different imaginative outcomes. For instance, when Chalmers (1996) says zombies are possible, and Shoemaker (1999) says they are impossible, not only do they each token different zombie scripts, their differently tokened script explains their different individual stances. Since Chalmers says human actions are decomposable into phenomenal and functional descriptions, his zombie script decomposes into scripts for those descriptions such that his phenomenal action script leads him to the possibility of zombies. Since Shoemakers says human actions are both phenomenal and functional, his zombie script does not so decompose, and so it can only lead to the impossibility of zombies.

Omoge also foregrounds Schank and Abelson's notion of 'interference' to account for the correct usage of imagination in metaphysical modalizing. Where interferences are mental states that prevent the normal unfolding of a script and which often sneak into the imaginative process during the composition of scripts. For instance, in reasoning his way to how functional properties fail to neatly supervene on phenomenal ones, Chalmers may have made some invalid reasoning steps, such that there are some interferences lurking in his zombie script. If so, then he would have wrongly used imagination to reach his view that zombies are possible.

Omoge thinks that due to theoretical assumptions, interferences often go unnoticed, and so are left uncorrected, and even when pointed out, the involved theories may make the imaginer resolute. This, he says, shows that the psychology of imagination and metaphysical modality come apart. For we now have an account of how an agent's imaginative processes can be faulty, which says nothing about the metaphysical conclusions they arrive at via imagination—after all, Chalmers could also use another cognitive faculty, e.g., intuition, to reach the same conclusions, and, certainly, the cognitive architecture of imagination is not identical with that of intuition.

Lastly, Omoge gives an evolutionary psychological argument

against Nichols' skepticism about the usage of imagination in metaphysical modalizing. In his view, talk of a natural domain matters little, if at all, because evolution does not ready-make all our cognitive faculties; some are appropriations of others. For example, spatial reasoning, which we have gone to appropriate for geometry. Omoge says the same appropriation holds for practical and metaphysical modalizing. Metaphysical modalizing may not be the natural domain of imagination, but that does not mean we are thereby barred from so using imagination. After all, geometrical reasoning is not the natural domain of spatial reasoning, yet it is indispensable. Talk of a natural domain matters little when considering the usefulness of a cognitive faculty.

While this view is commendable, Omoge does not address why Nichols is skeptical about the usage of imagination for metaphysical modalizing, which, recall, is that the single code hypothesis predicts that inference mechanisms would balk at contradictory imaginings because they balk at contradictory beliefs. I will conclude this section by supplying a rebuttal to this claim.

Here is the fact: imagination can be used to reason about contradictions, so it is factually incorrect that inference mechanisms balk when we so use imagination. In fact, it is factually incorrect that they balk at mathematical impossibilities like 1+1=7, which are the examples Nichols uses—Graham Priest (2016), for example, says he can perfectly imagine them. But he should not be able to do so if Nichols is correct. How, then, should we explain the imaginative processes of outliers like Priest? And Nichols should want to explain their imaginative processes since he says his view maps onto the cognitive architecture of imagination, which is identical for everyone. My own view is that Nichols gives up too quickly. The way out, as I see it, lies with the UpDater.

Nichols and Stich (2003: 32) set up the UpDater as though it works only in the involuntary mode, i.e., automatically. But I think it can also work in the (semi)voluntary mode. In the contexts of belief and practical modalizing, nomological laws are fundamental to how the UpDater filters out incompatible beliefs. Thus, the UpDater can work independently of what the agent wants to achieve—it just needs to follow the dictates of nomological laws, which, supposedly, are mentally filed in some determinate ways. Believing and practical modalizing are typically automated processes (Connors and Halligan 2015). You may withhold believing that your child, who was asleep in the bedroom, is the person giggling in the living room, at least until you peep to confirm, but believing so was triggered by the giggles you heard (assuming that both of you are alone in the house). Not so for metaphysical modalizing since everyone agrees that nomological laws are suspended therein. Without the guidance of the mental file for nomological laws, the UpDater falls back to what the agent wants to achieve. Simply, in metaphysical modalizing, the agent seizes control of the UpDater, telling it which beliefs to filter out, thereby making the UpDater sensitive to the agent's goal. Thus, outliers like Priest are voluntarily filtering out

beliefs that would block them from imagining metaphysical impossibilities. Not everyone can do this, however; relevant beliefs are needed. I will return to this in Section 6.

This view that the UpDater is sensitive to the agent's goal is not an affront to the single code hypothesis, it must be said. Nichols and Stich only say that inference mechanisms will treat beliefs and imaginings in *much* the same way, i.e., the hypothesis admits some differences between beliefs and imaginings. Nichols (2006a) himself discusses some of these differences at length. What I am adding, then, is that the UpDater's sensitivity is another difference in how inference mechanisms treat beliefs and imaginings. In belief and practical modalizing contexts, the UpDater is not sensitive to the agent's goal, but it is in metaphysical modalizing contexts. If so, then Nichols' skepticism is indeed unwarranted because the single code hypothesis does not, in fact, show that imagination cannot lead to metaphysical modality. We only need to build sensitivity to the agent's goal into the UpDater, and the single code hypothesis will accommodate metaphysical modalizing.

Now that we have seen how the cognitive architecture of imagination can be updated to become compatible with metaphysical modalizing, we can proceed to check whether non-I-recruiting PITEs, since they are cases of metaphysical modalizing, do indeed fare better than their I-recruiting counterparts vis-a-vis the nature of the self, as Nichols suspects. First, let us demonstrate the compatibility of non-I-recruiting PITEs so as not to beg the question.

5. *Non-I-recruiting PITEs and the cognitive architecture of imagination*

As I said (Section 1), I will focus on Parfit's fission in the remainder of this paper for simplicity's sake, although what I will say is generalizable to other non-I-recruiting PITEs. In fission, identical triplets were involved in an accident such that the body but not the brain of one is damaged (Brainy), and the brains but not the bodies of the other two are damaged (Lefty and Right). Parfit asks that if Brainy's brain is split into two halves such that Lefty gets the left half and Righty gets the right half, which of Lefty and Righty will be identical to Brainy? His famous answer: neither. From this, he concludes that what matters when identity does not obtain is psychological continuity, not personal identity.

First, let me show how his conclusion is subserved by the (updated) cognitive architecture of imagination, segueing from there to whether he uses imagination correctly to arrive at the conclusion. This second task is important because if fission is to succeed in leading us to what is essential about the self, then a good starting place is whether the conclusions it affords were correctly arrived at in the first place. As we all know, an invalid conclusion cannot be sound.

In fission, the invitation to imagine that "identical triplets were in-

volved in an accident …" will signal to the script elaborator to generate an imagination premise, which will be put inside Parfit's imagination box. The contents of his belief box will then be put inside the imagination box as further premises to yield the target imagining—namely, when identity does not obtain, what matters? His UpDater will then filter out any beliefs he may have that will be incompatible with the imagination premise, for example, some nomological beliefs about the physical impossibility of splitting brains into two. Here, as I said (Section 4), the UpDater is operating in a voluntary mode in that Parfit is manually controlling it, telling it to filter out the incompatible nomological beliefs, even though his UpDater will not filter the beliefs out were he not metaphysically modalizing. He can do this because he is a seasoned personal identity thinker, such that he has the relevant theoretical assumptions to maintain a coherent thought process despite manually hijacking the UpDater. For comparison, a first-year philosophy student may not be able to suspend the influence of nomological laws if they suppose that fission is possible.

Being a seasoned personal identity thinker would also enable some of Parfit's UpDater-unfiltered-out beliefs to contain a script that details how PITEs typically proceed, i.e., he has PITE scripts or, in our case, a fission script. Like any script, this fission script will be unrestrictive in that further details about thought experiments can be teased out from it independently of Parfit's theoretical assumptions. Simply, Parfit's script elaborator will embellish the imaginative scenario in ways not informed by his theoretical assumptions without straying from the scope set by the fission script. Thus, from what he imports into the imaginative process—which, of course, are the UpDater-unfiltered-out beliefs—imagination will continue in an autonomous mode, fleshing out other relevant details.

Now, as we have seen (Section 4), the fission script will be activated conceptually, i.e., when key concepts like 'personal identity' and 'psychological properties' are instantiated in Parfit's imaginative process. Since theoretical assumptions are rarely ever identical for two agents, the fission script is rarely ever identically tokened by two philosophers. Thus, when Gendler (2002) argues that Parfit is mistaken in saying that psychological continuity, not personal identity, is what matters, Gendler's fission script differs from Parfit's.

In addition to being activated conceptually, we have also seen that scripts are also compositional, given the etiology of the debate (Section 4). If so, then the fission script is compositional along the 'prudential concern' etiology of the debate. Where prudential concern, as it is used in the personal identity literature, is the sort of concern we bear towards our future selves, and prudential concern can be understood in both psychological and numerical terms. The fission script, then, is composed of a script for psychological continuity and another script for numerical identity. Since the compositionality of scripts informs different metaphysical modalizing conclusions, it follows that the manner

in which the fission script is composed for Parfit and Gendler explains why they arrive at polar opposite conclusions.

Simply, given Parfit's and Gendler's theoretical assumptions and the compositionality of their fission scripts, the scripts can each unfold in ways that prioritize one of the component scripts. Parfit's theoretical assumptions guide his fission script to prioritize the script for psychological continuity. Hence he says: "In all ordinary cases, personal identity and [psychological continuity] coincide. When they diverge, [psychological continuity] is what matters. That strongly suggests that, in all cases, [psychological continuity] is what matters" (Unpublished paper, but the quote is from Gendler 2002: 44). On the other hand, Gendler's theoretical assumptions guide her fission script to prioritize the script for numerical identity. Hence she says: "The fact that two features coincide in all actual cases may mean that there is no straightforward way for us to determine how we would or should respond to either in isolation" (Gendler 2002: 35).

Now, Gendler does not just say Parfit is wrong; she also says he could not have arrived at his conclusion imaginatively. This seems to be a step too far if what I have said here is correct. As we have just seen, it is consistent with the cognitive architecture of imagination that imagination can lead different agents to different imaginative conclusions, at least insofar as each conclusion follows from the normal unfolding of the agents scripts. Both Parfit's and Gendler's polar opposite conclusions follow from the normal unfolding of their different fission scripts. Everything, so far, is by the book.

We can go a step further, however, by checking whether any of them wrongly used imagination to arrive at their respective conclusions. To do this, we only need to identify in whose fission script interferences lurk. For instance, if Gendler's argument is correct, then some interferences lurk in Parfit's fission script. According to her, Parfit wrongly thinks that because psychological continuity and numerical identity ordinarily coincide, imaginary cases where they diverge show that the former is what matters. Such an illicit move would constitute an interference, blocking the normal unfolding of Parfit's fission script, such that he would have wrongly used imagination to arrive at his conclusion. *Mutatis mutandis* for Gendler if we can isolate the interferences lurking in her fission script. We should not, however, expect that neither Parfit nor Gendler will change their view if the lurking interferences are pointed out. As I have said (Section 4), when interferences are hooked up to theories, theoretical assumptions might, and they often do, make philosophers resolute, even when lurking interferences are pointed out. Thus, if Gendler indeed points out the interferences lurking in Parfit's fission script, we should not expect that he thereby changes his mind.[4]

[4] Gendler thanked Parfit for providing comments on earlier versions of her paper in the acknowledgment section. So, there is no doubt that he read the paper, yet her arguments did not sway him. He still published numerous works between

Interferences can also be psychological, not always conceptual, as in the above, but I do not think psychological interferences pose any threat to non-I-recruiting PITEs or any imaginative exercise for that matter. It has been argued that since the laboratory of thought experiments is the mind, PITEs (as well as other kinds of thought experiments) are subject to a host of psychological biases, like seeing ourselves in positive lights (e.g., Brown 1986, Taylor and Brown 1988). Consequently, Unger (1990) says these psychological biases jeopardize the reliability of PITEs. Put in our terms, the biases would make PITE scripts unfold in different ways than they ordinarily would, and so they are interferences. They are psychological interferences.

However, unlike their conceptual counterparts, psychological interferences would easily be correctable once pointed out, suggesting that their easy correction is a function of not being hooked up to background theories. If so, then I sincerely doubt that any philosopher would refuse to account for psychological interferences in their imaginative processes once pointed out. In fact, psychological interferences are one way we improve our imaginative processes. I do not take imaginative conclusions at face value anymore; I look out for where I might have overestimated my own abilities. I am confident that this applies to Parfit and Gendler as well. In short, psychological interferences are no threat to the success of non-I-recruiting PITEs.

It might be said, following Wilson and colleagues (1994, 2002), that even if we are aware of psychological interferences, we lack access to the ongoing psychological processes, and so we cannot decontaminate in real-time. It is unclear to me, however, why such access is required, not least because, typically, psychological processes are subpersonal. Take the UpDater. In some ways, its job is to decontaminate, and typically (i.e., in the contexts of belief and practical modalizing, when it works in the involuntary mode), it does this without our awareness. When you hear someone giggling in the living room, and the UpDater updates your belief system—from "my child is sleeping" to "my child is awake"—it brackets out some psychological biases as it does so, e.g., that you are not hallucinating the giggles. If so, then talk of immediate access holds little, if any, weight in talk of decontamination. Decontamination is psychological, not phenomenological.

So far, I have argued that neither Parfit's nor Gendler's conclusion about the self is wrong, although we might be able to say which of them wrongly used imagination to arrive at their conclusion. I want to end this section by saying that there is a deeper sense in which interferences can prove fatal for a philosopher's conclusion about the self. One reason fission is popular is that it aims to show that the non-reductionist, who is committed to identity being always what matters, faces a kind of *reductio ad absurdum*. If identity is always what matters,

2002 and 2017—when he died—that propagate the same idea that what matters is psychological continuity, not personal identity.

then the non-reductionist must describe the outcome of fission in identity terms, yet any such description conflicts with some principle to which they are also committed. Brainy cannot be both Lefty and Righty given the necessity of identity; he cannot be neither, as he survives in the single case, which is no different from each side of the double case; he cannot be Lefty rather than Righty as that would make identity arbitrary.[5] Simply, whatever the non-reductionist say is wrong on their own terms. One might say then that Parfit's argument is meant to show that there is something internally wrong with the non-reductionist's fission script.[6]

I should stress that this deeper sense in which interferences are useful has not taken us too far afield. We are still within the scope of talking about the correct usage of imagination to arrive at metaphysical conclusions about the self; we have not been transported to talking about whether the conclusions themselves are correct. The latter is a metaphysical discussion; the former is a cognitive psychological one. The fact that interferences can be fatal to the success of an imaginative act is part of what "using imagination wrongly" means. Put simply, interferences do not merely reveal the thought experimenter's theoretical commitments; sometimes, they do much more, revealing why some thought experiments work and why some others do not work. If fission succeeds in leading us to what is essential about the self, then its success is at the expense of the non-reductionist. This analysis is compatible with the cognitive architecture of imagination.

What we have come to then is an explanation of non-I-recruiting PITEs with the cognitive architecture of imagination. Put plainly, it is an explanation of why imagination does not fail in non-I-recruiting PITEs. That said, such an explanation does not tell us whether the metaphysical conclusions non-I-recruiting PITEs deliver reveal anything essential about the self. After all, as Nichols points out (Section 3), we can explain I-recruiting PITEs with the cognitive architecture of imagination, but once we do so, we see why we should not infer the nature of the self from them. Thus, we must still ask whether this compatibility between the cognitive architecture of imagination and non-I-recruiting PITEs reveals the same thing about the nature of the self. Does it reveal that we should not infer the nature of the self from non-I-recruiting PITEs? I will argue that it does not.

[5] Gendler is not a non-reductionist in the sense I am using the term here. Her misgiving with Parfit is just that his explanation for why prudential concern subsists in the absence of identity is wrong: "Nevertheless, as I have maintained throughout, Parfit is right that if Brainy were to undergo fission, the relation of prudential concern he would find himself bearing to Lefty and to Righty would be rational—even if he knew that he was to undergo fission. What Parfit is wrong about is the explanation of this" (2002: 51).

[6] Thanks to an anonymous reviewer for this journal for this stronger sense in which interferences are useful.

6. *Should we descry the nature of the self from non-I-recruiting PITEs?*

Central to answering this question is the challenge that non-I-recruiting PITEs are impoverished in that they lack relevant background information, and so we should be cautious when drawing metaphysical conclusions about the self from them (Wilkes 1988; van Inwagen 1997; Schechtman 2014). For instance, Wilkes says:

> How often [do fission occur]? Is it predictable? Or sometimes predictable and sometimes not, like dying? Can it be prevented? Just as obviously, the background society, against which we set the phenomenon is now mysterious. Does it have such institutions as marriage? How could that work? Or universities? It would be difficult, to say the least, if universities double in size every few days, or weeks, or years. Are pregnant women debarred from splitting? The *entire* background here is incomprehensible (1988: 11, original italics).

The point here is that nouns (common and proper) come with all the descriptive (Russell 1911) and/or causal-historical (Kripke 1980) residuals that characterize them, which non-I-recruiting PITEs leave out. We learn names of places or things at elementary schools, names of people at their christening, or when we come to know/meet them, and we keep updating the descriptions associated with the names throughout life. Not supplying these associated descriptions, therefore, makes non-I-recruiting PITEs incomplete. Being so incomplete, we should take them with the proverbial pinch of salt, in almost the same manner we take their I-recruiting PITEs that are equally impoverished.

I agree that non-I-recruiting PITEs are descriptively impoverished in the above way, but I deny that this poverty of description amounts to anything significant. It does not amount to non-I-recruiting PITEs failing to lead us to metaphysical conclusions about the nature of the self. My reason is that this challenge (hereafter, as the Wilkes-Van Inwagen-Schectman challenge), as evident from the last sentence of the previous paragraph, wants to parallel non-I-recruiting PITEs with I-recruiting PITEs, which cannot work. The Wilkes-Van Inwagen-Schectman challenge wants to say that since I-recruiting PITEs are descriptively impoverished, which is why they are unreliable guides to the nature of the self (Section 3), so too will the descriptive poverty of non-I-recruiting PITEs make them unreliable guides to the nature of the self. This argument does not work.

The reason I-recruiting PITEs are descriptively impoverished is that the mental symbol underwriting their operation (i.e., the I-concept) is also descriptively impoverished (Section 3). This is not the case for non-I-recruiting ones. Though they are descriptively impoverished, their descriptive poverty is not caused by the descriptive poverty of the mental symbol underwriting their operation. Nichols puts this difference in psychological structure between I-recruiting and non-I-recruiting PITEs this way:

> But even if both indexicals and proper names have similarly Kripkean semantics, it would be a mistake to conclude that this means that indexical concepts and proper name concepts are also equivalent in their psychological characteristics. Rather, it's plausible that the processing associated with the I-concept differs in important ways from the processing associated with proper name concepts. To take one example, we often deploy proper names that seem nonunique, as when I think *Michael is meeting me for lunch*. I know which *Michael* I have in mind, and it's plausible that this is because of the information I have associated with that token of *Michael*. By contrast, since there's only one I-concept, I never need to worry about disambiguating it. (2008: 523)

If so, then even though both I-recruiting and non-I-recruiting PITEs are descriptively impoverished, their psychological structures differ. Call the mental symbol underwriting the operation of nouns the 'noun-concept'. Unlike the I-concept, which is flexible (Section 3), the noun-concept is rigid because it contains different mental files for different nouns. For instance, there are separate files for the many *Michaels* I know, and each file keeps getting updated as more historical facts about each of them come to my awareness. If one of them wins a Nobel, that fact will not be stored in the file of a *Michael* who is a soccer player. In short, where the I-concept is poor, the noun-concept is abundantly rich.

Here, then, is the psychological difference between I-recruiting and non-I-recruiting PITEs. Since the I-concept is poor, it is functioning normally in I-recruiting PITEs, which are also descriptively poor. This is why it is easy to imagine that *I am someone else*: the I-concept has no descriptive content, so it works anyway. I-recruiting PITEs inherit the descriptive poverty of the I-concept. Contrariwise, since the noun-concept is abundantly rich, it is not functioning normally in non-I-recruiting PITEs, which are descriptively poor. This is why it is difficult to imagine that *Obama is Napoleon*. My noun-concept has separate files for *Obama* and *Napoleon*, which contain all the historical facts I associate with them, and so the noun-concept finds it difficult to combine or crisscross data from both files. Non-I-recruiting PITEs do not inherit the descriptive wealth of the noun-concept.

The Wilkes-Van Inwagen-Schectman challenger may respond that all that this talk of malfunctioning of the noun-concept shows is that imagination also fails in non-I-recruiting PITEs, just as it fails in I-recruiting ones, such that their parallelism stands. Put differently, they would say that we are being asked to imagine a world where the mental files we have for nouns are different from the ones we currently have, but we are not told what data they contain, and this is troubling because we are using our current concepts for the nouns in the imagined world. Thus, when Parfit talks about fission, the Wilkes-Van Inwagen-Schectman challenger would retort that he skips relevant details about *brains*, *triplets*, *splitting*, and so on. As we saw, the complaint is that details like "How often do fission occur? Is it predictable? Can it be prevented?" (Wilkes 1988: 11) are skipped.

To start with, imagination does not thereby fail because the noun-concept is malfunctioning in non-I-recruiting PITEs. This is because the architecture of imagination can supplement the shortcomings of the noun-concept. As we saw (Section 4), scripts are unrestrictive in that the script elaborator can tease out details that are neither informed by scripts nor the combination of the UpDater-unfiltered-out beliefs and the imagination premise. If so, then notwithstanding the malfunctioning of the noun-concept, the script elaborator will supply the details needed to ensure the success of imagination in non-I-recruiting PITEs. Put differently, the noun-concept cannot be descriptively rich to such an extent that the script elaborator becomes superfluous. In short, the script elaborator ensures the success of imagination even though the noun-concept is malfunctioning in non-I-recruiting PITEs.

In addition, the details the Wilkes-Van Inwagen-Schectman challenger demands are, contrary to what they say, irrelevant to non-I-recruiting PITEs. Earlier, we saw that a script is generated for an event on account of the event's repeatedness: e.g., by repeatedly engaging in PITEs, a PITE script is generated (Section 4). This, I said, is why Parfit has a fission script and a first-year philosophy student may not. If so, then Parfit could have fleshed out fission with more details than he did—even along the lines Wilkes (1988: 11) enumerates. He presumably did not because such details were irrelevant to the points he wanted to make. Here is why.

First, we do not live in a splitting world, so it is unclear why *sociological* facts about splitting worlds should be important to us. As Kripke (1980) complains similarly about Lewis' (1971) counterpart theoretic framework of possibilia: what our counterparts in possible worlds do is irrelevant to what happens to us in the actual world. Second, we are after metaphysical, not sociological, conclusions, and we can draw them from hypothetical situations that are sociologically under-described. After all, not only do we not live in a world where cats are both dead and alive but the world is also sociologically under-described, yet we infer the relativistic nature of time from such a world. Simply, Parfit is licensed to draw metaphysical conclusions from fission even though it is under-described.

The Wilkes-Van Inwagen-Schectman challenger may say that I have missed the point of their challenge, which is that providing the details would have made imagining the scenario easier. Though some theorists have caved to this line of response—"the details simply go to making the scenario more easily imaginable" (Beck 2016: 124)[7]—I

[7] I am unsure why Beck concedes this point, however. I read him as saying the details the Wilkes-Van Inwagen-Schectman challenge demands are irrelevant to the imagined scenario. His view, which I agree with and discuss below in the main text, is that the challenge mistakes which belief system is integral to imagining the scenario. The Wilkes-Van Inwagen-Schectman challenge thinks it is some non-actual belief system that's actualized for the imagined scenario, whereas what is needed is non-actualizing our actual belief system.

want to dig in my heels. I do not think I have missed the point because I doubt that any of Wilkes, Van Inwagen, and Schectman would agree with this interpretation. What they are saying is rather that once the details they demand are provided, it becomes clear that we are not imagining what we think we are imagining at all, i.e., the details would make imagining the scenario more difficult, not easier.

I contend, however, that they only say this because they wrongly think that the details are relevant to fission, such that the relevance justifies why not providing them is fatal for fission. Having seen that the details are, in fact, irrelevant to fission, it follows that the Wilkes-Van Inwagen-Schectman challenge is unfounded, and so non-I-recruiting PITEs are not descriptively impoverished. They have just the right amount of background details they need, and we are imagining what we think we are imagining with them. This calls to mind Berto and Jago's clarification about how imagination operates: "It's important, however, not to treat agents as importing too much background information into acts of imagination. We do not indiscriminately import arbitrary, unrelated contents into imagined scenarios […] exercises of imagination must obey some constraint of relevance" (2019: 144). Put simply, imagination does not work in the way the Wilkes-Van Inwagen-Schectman challenge wants.

There is more. We saw that one reason the Wilkes-Van Inwagen-Schectman challenge is plausible is that we are supposed to employ our current concepts in the imaginative process even though our noun-concepts have different mental files. Since everyone agrees that different nomological laws hold in possible worlds such that we cannot observationally test the accuracy of our concepts, the Wilkes-Van Inwagen-Schectman challenger would add that we cannot know what we would say, and "what we would say" is the fulcrum on which PITE scenarios turn (Fodor 1964, Ricœur 1992, Wagner 2016). It is unclear to me, however, why what we would say *in* the described possible world matters—as I have said, we do not live there, so why should we worry about some putative belief system that we would hold there? Simply, the issue is not "'What would our beliefs *in the context* be if such-and-such were the case?' [But] 'What do we say *in our context* if such-and-such were the case'" (Beck 2006: 43, original italics). The issue is not actualizing some non-actual belief system for non-I-recruiting PITEs but non-actualizing our actual belief system.

This correction, of course, is backed by the cognitive architecture of imagination. As we have seen, imagination operates solely by manipulating our actual beliefs (Sections 2 and 4). This is why the content of the belief box is copied into the imagination box once the imagination premise is generated by the script elaborator: the agent's actual web of beliefs (occurrent and dispositional) is used as premises during imagination. The belief box only contains actual beliefs. Even scripts, which supply details that are not inferable from our background knowledge,

are components of actual beliefs. In short, a scenario is imaginable if and only if the agent has either occurrent (conscious and unconscious) or dispositional beliefs about it. Priest can imagine 1+1=7 because he has at least dispositional beliefs about it (inferring from his paraconsistent logical theoretical assumptions). Whereas because I lack both occurrent and dispositional beliefs about it, given that I am no paraconsistent logician, I cannot imagine it. Unlike Priest, I cannot maintain a coherent reasoning process if I manually hijack my UpDater, telling it to override any belief that would block me from imagining 1+1=7.

Lastly, the Wilkes-Van Inwagen-Schectman challenger may say that even if our actual belief system is at work, non-I-recruiting PITEs cannot show what ought to matter to everyone. That is, since there is no universal belief system that applies to everyone, even if imagination works with the imaginer's actual belief system, only subjective, not objective, normative conclusions can be drawn from it (Martin 1997; Rovane 1997; Baker 2000). This residual challenge does not say that we should not draw metaphysical conclusions from non-I-recruiting PITEs, but that the drawn metaphysical conclusions would lack the dispositive force they *ought* to have because they would only apply to individuals, not everyone. Simply, what follows from non-I-recruiting PITEs is not indicative of what obtains in real life in that the normative conclusions are not factual. It is unclear to me, however, why normative conclusions must be factual.

Why must what "*ought* to matter" matter to everyone? Answer: it must not. It is not a requirement for normative conclusions that they apply to everyone; there is room for disagreements. I may say, "you ought to be friendly with your neighbors," and you may counter, "what if they are nosy and annoying?" In short, normative conclusions, either physical (as with being friendly with your neighbors) or metaphysical (as with PITEs), are contested, so they need not apply to everyone.

But that's not all: the oughtness of normative claims seems to override these disagreements. What I mean is that we often admit differences in what ought to matter to different people, respect their choices, and still say, "even so, what ought to matter to you is so-and-so." Simply, the oughtness of a normative claim overrides whatever differences of opinion there may be among different agents. You may say, "what matters to me when identity does not obtain is numerical identity," and someone else may respond, "that's okay, but what ought to matter to you is psychological continuity." This, in part, is what Parfit aims to demonstrate with fission, which is that regardless of whether you think numerical identity is what matters in the absence of identity, fission shows that what ought to matter to you is psychological continuity.

This view applies to thought experiments even outside philosophy. For example, the Einstein-Bohr disagreement about entangled particles,[8] which asks whether physical reality exists independent of

[8] Quantum entanglement occurs when two or more particles interact in a way

our ability to observe it. Einstein said yes; Bohr said the question is meaningless. We now know, thanks to John Bell some 30 years after the debate, that Einstein was wrong: there are indeed limits on the predicted correlations between entangled particles. The diagnosis of this resolution in cognitive psychological terms is now clear given what I have said in this paper: the interferences in Einstein's script are the fatal kinds á la those in the non-reductionist's fission script.

In conclusion, to the extent to which the metaphysical conclusion drawn from non-I-recruiting PITEs is normative, the cognitive architecture of imagination allows a plurality of them, leaving room for how one can trump another. If so, then Nichols is right: unlike I-recruiting PITEs, there are no dangers to descrying the nature of the self from non-I-recruiting PITEs.[9]

References

Baker, L. R. 2000. *Persons and Bodies: A Constitution View*. Cambridge: Cambridge University Press.

Beck, S. 2006. "These Bizarre Fictions: Thought-Experiments, Our Psychology and Our Selves." *Philosophical Papers* 35 (1): 29–54.

Beck, S. 2016. "Technological Fictions and Personal Identity: On Ricoeur, Schechtman and Analytic Thought Experiments." *Journal of the British Society for Phenomenology* 47 (2): 117–132.

Berto, F., and Jago, M. 2019. *Impossible Worlds*. Oxford: Oxford University Press.

Brown, J. 1986. "Evaluations of Self and Others: Self-enhancement Biases in Social Judgments." *Social Cognition* 4 (4): 353–376.

Chalmers, D. 1996. *The Conscious Mind*. Oxford: Oxford University Press.

Connors, M., and Halligan, P. 2015. "A Cognitive Account of Belief: A Tentative Road Map." *Frontiers in Psychology* 5: https://doi.org/10.3389/fpsyg.2014.01588

Fodor, J. 1964. "On Knowing What We Would Say." *Philosophical Review* 73 (2): 198–212.

Fodor, J. 1987. *Psychosemantics*. Denver: A Bradford Book.

Gendler, T. 2002. "Personal Identity and Thought-Experiments." *Philosophical Quarterly* 52 (206): 34–54.

Kripke, S. 1980. *Naming and Necessity*. Cambridge: Harvard University Press.

Lewis, D. 1971. "Counterparts of Persons and Their Bodies." *Journal of Philosophy* 68 (7): 203–211.

Martin, R. 1997. *Self-Concern*. Cambridge: Cambridge University Press.

that each particle's quantum state (i.e., the probability distribution for the outcomes of each possible measurement) cannot be described independently of the quantum state of the others, however large the distance between the particles.

[9] Special thanks to Simon Beck for reading an earlier version of this paper and for his useful comments. An anonymous reviewer for this journal also provided significant comments, which helped to solidify the paper's contribution.

Miyazono, K., and Liao, S. 2016. "The Cognitive Architecture of Imaginative Resistance." In A. Kind (ed.). *The Routledge Handbook of Philosophy of Imagination*. New York: Routledge, 233–246.

Nichols, S. 2004. "Imagining and Believing: The Promise of a Single Code." *The Journal of Aesthetics and Art Criticism* 62 (2): 129–139.

Nichols, S. 2006a. "Just the Imagination: Why Imagining Doesn't Behave Like Believing." *Mind and Language* 21 (4): 459–474.

Nichols, S. 2006b. *The Architecture of the Imagination*. Oxford: Clarendon Press.

Nichols, S. 2007. "Imagination and Immortality: Thinking of Me." *Synthese* 159 (2): 215–233.

Nichols, S. 2008. "Imagination and the I." *Mind and Language* 23 (5): 518–535.

Nichols, S., and Stich, S. 2003. *Mindreading*. Oxford: Oxford University Press.

Omoge, M. 2021. "Imagination, Metaphysical Modality, and Modal Psychology." In C. Badura and A. Kind (eds.). *Epistemic Uses of Imagination*. New York: Routledge, 79–99.

Omoge, M. Forthcoming. "On the Place of Imagination in the Architecture of the Mind." In E. Sullivan-Bissett (ed.). *Belief, Imagination, and Delusion*. Oxford: Oxford University Press.

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

Priest, G. 2016. "Thinking the Impossible." *Philosophical Studies* 173 (10): 2649–2662.

Ricœur, P. 1992. *Oneself as Another*. Chicago: University of Chicago Press.

Rovane, C. 1997. *The Bounds of Agency*. Princeton: Princeton University Press.

Russell, B. 1911. "Knowledge by Acquaintance and Knowledge by Description." *Proceedings of the Aristotelian Society* 11: 108–128.

Schank, R., and Abelson, R. 1977. *Scripts, Plans, Goals, and Understanding*. New York: Psychology Press.

Schechtman, M. 2014. *Staying Alive*. Oxford: Oxford University Press.

Shoemaker, S. 1999. "On David Chalmers's The Conscious Mind." *Philosophy and Phenomenological Research* 59 (2): 439–444.

Taylor, S., and Brown, J. 1988. "Illusion and Well-being: A Social Psychological Perspective on Mental Health." *Psychological Bulletin* 103 (2): 193–210.

Unger, P. 1990. *Identity, Consciousness, and Value*. Oxford: Oxford University Press.

Van Inwagen, P. 1997. "Materialism and the Psychological-Continuity Account of Personal Identity." *Philosophical Perspectives* 11: 305–319.

Wagner, N. 2016. "Transplanting Brains?" *South African Journal of Philosophy* 35 (1): 18–27.

Wilkes, K. 1988. *Real People*. Oxford: Oxford University Press.

Williams, B. 1970. "The Self and the Future." *The Philosophical Review* 79 (2): 161–180.

Williams, B. 1973. "Imagination and the Self." In B. Williams. *Problems of the Self*. Cambridge: Cambridge University Press, 26–45.

Wilson, T., and Brekke, N. 1994. "Mental Contamination and Mental Correction: Unwanted Influences on Judgments and Evaluations." *Psychological Bulletin* 116 (1): 117–142.

Wilson, T., Centerbar, D., and Brekke, N. 2002. "Mental Contamination and the Debiasing Problem. " In D. Griffin, D. Kahneman, and T. Gilovich (eds.). *Heuristics and Biases*. Cambridge: Cambridge University Press, 185–200.