Forthcoming in *Mind & Language*, 23 (2008).

Imagination and the *I*[*]

SHAUN NICHOLS

Abstract:

Thought experiments about the self seem to lead to deeply conflicting intuitions about the self. Cases imagined from the 3rd person perspective seem to provoke different responses than cases imagined from the 1st person perspective. This paper argues that recent cognitive theories of the imagination, coupled with standard views about indexical concepts, help explain our reactions in the 1st person cases. The explanation helps identify intuitions that should not be trusted as a guide to the metaphysics of the self.

It would be hard to exaggerate the benefits we reap from our capacity for imagination. The imagination is critical to hypothetical reasoning, planning, and creativity. It is central to science and even more central to philosophy. But my mission here is not to celebrate the imagination; rather, I aim to expose a shortcoming in how the imagination responds to certain thought experiments. While the imagination is plausibly reliable in many thought experiments in both science and philosophy, it is not, I will argue, a good guide in thought experiments that recruit the concept *I*. The problem arises, I suggest, because of the peculiar interplay of indexicals like *I* and the cognitive structure of the propositional imagination.

1. Thought experiments about the self

'Whether we are to live in a future state, as it is the most important question which can possibly be asked, so it is the most intelligible one which can be expressed in language.' Thus begins Bishop Butler in his discourse on personal identity. In hindsight, this was a surprising claim. For the issue of whether we are to persist in a future state has been one of the most contentious issues in metaphysics.

The contemporary agenda in personal identity was largely set in some beautiful early papers by Bernard Williams. He raises a tangle of puzzling phenomena associated with

[*] An early version of this paper was presented at the 2007 *Mind & Language* conference on pretense and imagination and at *Mimesis, Metaphysics, and Make-Believe*, a conference in honor of Ken Walton. A later version was presented at the 2008 *Conference on Imagination*, at Temple University. I'm grateful to the audiences and other colleagues for suggestions, especially Noel Carroll, Emma Cohen, Tim Crane, Greg Currie, Stacie Friend, Tamar Gendler, Samuel Guttenplan, Paul Harris, Sarah-Jane Leslie, Aaron Meskin, Bill Seeley, Deena Skolnik-Weisberg, Kathleen Stock, Ken Walton, Jonathan Weinberg, and Stephen Yablo. I'm also indebted to Claire Cooper, Marga Reimer, and Philip Robbins for extremely helpful comments on this material.

imagining about the self.[1] The first puzzle concerns the possibility of imagining that I am somebody with none of the distinguishing traits that I actually have. Can I imagine that I'm Napoleon at Waterloo? At a first pass it seems that this is possible. Williams writes, 'If we press this hard enough, we readily get the idea that it is not necessary to being *me* that I should have any of the individuating properties that I do have, this body, these memories, etc' (Williams 1966, 41; see also Walton 1990, 32). But if I'm imagining that all my psychological and physical traits are gone, it seems that I wouldn't exist at all – there would only be Napoleon (Williams 1966, 41-42). My apparent ability to imagine that I am Napoleon stands in contrast with attempting to imagine that someone else, say FDR, is Napoleon. It's much harder to make sense of that proposed imagining. Why is it, then, that it seems comparatively easy to imagine that *I* am Napoleon?

The second major puzzle is perhaps the best known example in the literature. Adapting a Lockean thought experiment, Williams gives a case in which two persons, A and B, are to have all of their psychological characteristics swapped by massively reprogramming each brain. Before the swap, they are told that one of the resulting persons will be tortured and the other given a wad of cash. Should this scenario be described as a case of changing bodies? Williams considers a number of possibilities. Suppose A requests that the torture go to the A-body and B requests that the torture goes to the B-body; since the psychology is swapped, the B-body will be the person who remembers A's request, and the A-body will remember B's request. After running through various permutations, Williams writes, "all the results suggest that the only rational thing to do, confronted with such an experiment, would be to identify oneself with one's memories, and so forth, and not with one's body" (167). That is, this thought experiment leads us to describe the scenario as one of changing bodies. Next, however, Williams has us consider the same case from the 1st person perspective. Someone says that he'll torture me tomorrow. But first, he'll remove all of my memories and other distinctive psychological traits, then he'll insert false memories. After all that, he'll begin the torture. How should I react to the prospect of torture in this case?

> Fear, surely, would … be the proper reaction: and not because one did not know what was going to happen, but because in one vital respect one did know 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 (168)

So Williams maintains that the 1st person thought experiment suggests that I survive the destruction of my psychological properties. Williams' 1st person thought experiment thus produces a reaction that is at odds with psychological accounts of the self. According to Williams, it is a critical feature of these cases that one is presented in the 3rd person and the other in the 1st person.

> The first argument, which led to the 'mentalistic' conclusion that A and B would change bodies and that each person should identify himself with the destination of his memories and character, was an argument entirely conducted in third-personal terms. The second

[1] Williams' verdicts about what is intuitive here seem largely to be extrapolations from his own intuitions. It will be important to do careful studies exploring the different factors that influence judgments about self. If we approach this issue as experimentalists, it's clear that there are many possible confounds in Williams' scenarios, and these need to be sorted through. That said, for present purposes I will follow the bulk of the philosophical literature in assuming that Williams' intuitions are representative.

> argument, which suggested the bodily continuity identification, concerned itself with the first-personal issue of what A could expect. That this is so seems to me… of some significance (Williams 1970, 179).

Williams' two scenarios present us with a stark impasse. When given the 3rd person version, our intuitions side with a psychological account of personal identity according to which the self goes where the psychological characteristics go. When given the 1st person scenario, our intuitions run against the psychological account.

Williams' own response to the impasse is to side with the intuitions delivered in the 1st person case. He concludes that the 3rd person version is rife with problems. Perhaps most importantly, Williams notes that if the experimenter had created *two* new individuals with the psychological traits of A, then it would be evidently problematic to identify *the one* that is A. As a result, Williams suggests that we should tentatively follow the intuitions delivered by the 1st person version:

> the principle that one's fears can extend to future pain whatever psychological changes precede it seems positively straightforward. Perhaps, indeed, it is not; but we need to be shown what is wrong with it. Until we are shown what is wrong with it, we should perhaps decide that if we were the person A then, if we were to decide selfishly, we should pass the pain to the B-body-person (1970, 180).

More recently, some have reacted to the impasse by pronouncing a kind of permanent stalemate. For instance, Sider writes:

> It appears that we are capable of having either of two intuitions about the case, one predicted by the psychological theory, the other by the bodily continuity theory. … Perhaps new thought experiments will be devised that tell decisively in favor of one theory or the other. Or perhaps new theoretical distinctions will be made that will make clear that one or the other competing sets of intuitions were confused, or mislabeled…. I doubt these things will occur, but it is impossible to know in advance what future philosophical investigation will reveal (Sider 2001, 198; see also McGinn 1993).

Sider thus doubts that new thought experiments or philosophical distinctions will point the way to an account of personal identity that respects all of our intuitions.

Like Sider, I doubt that this stalemate can be broken by proliferating thought experiments and distinctions, but there is another option – we can explore why the imagination responds as it does to the cases. My thesis is that intuitions about the 1st person scenarios turn on peculiar features of imagining with indexicals, and this will provide some reason to think that the resulting intuitions shouldn't guide our metaphysics of persons. If this is right, it cuts against both those who are skeptical about making progress on the issue (e.g., Sider, McGinn) and those who use the 1st person scenarios to undermine psychological accounts (e.g. Williams). Williams says that we should follow the intuition from the 1st person case 'until we are shown what is wrong with it.' My aim is to show what is wrong with it.[2]

2. Indexical concepts

[2] I will be neutral here about the status of the intuitions that favor the psychological account. But if I'm right that we can't trust the intuitions based on 1st-personal imagining, then this will remove one significant obstacle facing the psychological account.

To proceed, we need to chart several key features of the indexical concepts *I, here*, and *now.*[3] Fortunately, there is a great deal of good work on the topic.

I begin with two familiar observations about indexical terms like '*I*'. First, following Kaplan, it is traditional to identify the *content* of 'I' as the individual picked out by the token of the word. So when Kaplan speaks in the first person singular, the content of his use of 'I' is just him – that particular person. In addition there is the 'linguistic character' of 'I'. The linguistic character is the rule that tells us in general how to determine the content of a term from its particular utterance. In the case of 'I', the character is the rule that the content of an utterance of 'I' is the speaker of the utterance (Kaplan 1989; Perry 1979).

When we move into the psychological realm, it is clear that we need something more than the content/character distinction. The character of 'I' is, crudely, 'the person who is uttering the token "I"'. But as Recanati points out, that is not how I think of myself: 'I think of myself as myself, not as the utterer of such and such a token' (Recanati 1993, 71). To accommodate the indexical in psychology, we need to add another element. Whatever else is required, I think we will need to posit an internal mental symbol or vehicle corresponding to the term 'I'. We deploy this mental symbol when we have thoughts that we express with words like 'I' and 'me'. We can call this mental symbol the 'I-concept', or simply, *I*.[4] This is not, I should stress, a radical proposal. Indeed, Georges Rey presents a version of this view in his textbook on the philosophy of mind. Rey writes:

> Just as in English the speaker is supposed to use 'I' only to refer to his or herself, in [the language of thought] the 'system' uses a certain term to refer only to the receiver of present inputs, the instigator of outputs, and the subject of intervening mental states. Suppose the term were 'i,' and that the system uses 'i' to record automatically that it had certain perceptions and judgments and preferences; and that *its behavior is crucially determined by just those attitudes that do have this 'i' as their subject:* i.e. its actions are standardly caused by beliefs and preferences that are designated as belonging to 'i.' (Rey 1997, p. 291; cf. Perry 2000, Recanati 1993, 88).[5]

Rey's passage might suggest that the psychological system's behavior is only caused by attitudes that implicate the I-concept. Such a claim would likely be too strong. For instance,

---

[3] Interestingly, in the anthropological work on semantic primitives (Goddard & Wierzbicka 2002), these 3 notions – *I, now,* and *here* – are included in the rather small set of primitives (currently 61). For purposes of this paper, the focus will be entirely on these three indexicals, which might be importantly different from other indexicals.

[4] Recanati proposes to supplement the Kaplanian account with *psychological* modes of presentation (Recanati 1993, 72-76, 168). For the case at hand, Recanati proposes the 'egocentric concept' **Ego.** This concept, like other egocentric concepts **Hic** and **Nunc**, serves as a repository for perceptual and descriptive information (1993, 124-5). Since my focus is on the mental *symbols*, the notion of I-concept invoked above might differ from Recanati's notion of **Ego-**concept as repository.

[5] A similar view is presented in Maite Ezcurdia's entry on indexicals and demonstratives in the *Encyclopedia of Cognitive Science*: 'What "I" "now" and "here" thoughts do for the subject is locate the subject, the time, and the place in a way that connects with the subject's ability to perceive, think, and move…. Upon having an "I" thought, one is presented to oneself in a special way' (Ezcurdia 2002, 502). See Jenann Ismael (2007) for a subtle treatment of the idea that indexicals like 'I' have a locating function.

many behaviors are produced by automatic processes and these might well proceed without implicating the I-concept. Nonetheless, the important point to draw from Rey's passage is that a key function of the I-concept is to pick out the self for purposes of action production. To adapt an example from Perry (1977, 494), if I learn that SN is about to be attacked by a bear, this won't motivate the appropriate action (slowly retreating without turning my back on the bear) unless I also think that *I* am SN.[6]

For our purposes, the other important feature of the I-concept is that it is descriptively exceedingly thin. I take this observation to be in concert with Kant's claim that the *I* is the 'the poorest of all representations' B408.[7] We can illustrate this by considering the possibility of thinking *I*-thoughts under amnesia (cf. Perry 1977). A person can wake up in darkness with total amnesia yet think *I have a headache*. The I-concept is functioning normally here, but the agent has no distinctive descriptive content associated with *I*. Indeed, there is a bit of relevant clinical evidence. People with Alzheimer's disease use the first-person indexical (and presumably the associated concept *I*) frequently and appropriately, even in late stages of the disease (Tappen et al. 1999).

This point about the poverty of the I-concept concerns the *psychological* profile of the I-concept. When it comes to the *semantics*, as we've already seen, the standard view is that the referent of 'I' is determined not by associated descriptions, but rather by the sparse character ('the speaker of this token of "I"') plus the context. This aligns in important respects with the standard Kripkean account of the semantics for proper names, according to which the reference of a proper name is determined not by associated descriptions but rather by a causal chain of uses stretching back to a naming ceremony. There are reasons to be skeptical of a univocally Kripkean approach to proper names (e.g. Machery et al. 2004; Reimer forthcoming). 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. (See also Recanati 1993.)

Although Kant claims that the *I* is the poorest representation, it's important to recognize that there are other indexical concepts that are almost equally poor, e.g., *now* and *here*. To steal yet another example from Perry (1979), I can be oblivious to the time, but find out from a colleague that the meeting starts *now*. *Now* is functioning normally in this case, even though it is descriptively impoverished. Indeed, it's plausible that the poverty of these indexical concepts facilitates action production in absence of descriptive information. It's an important feature of

---

[6] The I-concept is plausibly also distinctively connected with introspection and perhaps proprioception (see Nichols 2001; Robbins 2002).

[7] We find a related claim about the poverty of the *I* in Anscombe's 'The First Person'. Anscombe takes the radical view that 'I' is not a referring expression (p. 32). But in light of Kaplan's work, we can allow that 'I' refers perfectly well – just as much as 'now', so Anscombe's proposal seems an overreaction. Of course, this leaves open what the self really is to which 'I' refers. But it would be too much to ask a theory of indexicals to answer that question.

the *I* concept that, even if I don't know any distinguished features of myself, I can still engage in appropriate retreating behavior given the representation *I am being threatened by a bear*. Similarly, and less dramatically, it's an important feature of the *now* concept that, even if I have no idea of the day or time, I can do my departmental duty given the representation *the meeting starts now.* The fact that these concepts do not depend on descriptive content allows them to work even under conditions of great ignorance.

Finally, it's worth noting that perhaps none of these indexicals is *entirely* empty of categorical information. The thought *I am a curtain* seems somehow defective. Perhaps *I* carries the descriptive content that the bearer is an *agent.* But such categorization information is much less rich than the categorization information associated with, say, *human.* Similarly, *now* plausibly carries *some* categorical information. There is something wrong with the thought *Now is the cup. Now* carries the descriptive content that the referent has to be *temporal*.

3. Cognitive accounts of the imagination

Turning now to cognitive accounts of the imagination, I want to set out what I take to be three points of consensus in recent work on cognitive accounts of the imagination.

First, recent accounts adopt a representationalist approach to the imagination. To believe that *Charles is the Prince of Wales* is to have a "belief" representation with the content *Charles is the Prince of Wales.* Analogously, to imagine that *Keith Richards is the Prince of Wales* is to have an "imagination" representation with the content *Keith Richards is the Prince of Wales.*

The second point is that imagination states are distinguished from beliefs by their *functional roles* not by their contents. This is nicely illustrated in an experiment by Alan Leslie (1994). Leslie had young children watch as he pretended to pour tea into two (empty) cups. Then he picked up one of the cups, turned it over and shook it, turned it back right side up and placed it next to the other cup. The children were then asked to point at the 'full cup' and at the 'empty cup'. Both cups were really empty throughout the entire procedure, but two-year-olds reliably indicated that the 'empty cup' was the one that had been turned upside down and the 'full cup' was the other one. On the most natural interpretation of this, the child is *imagining that the cup is empty*. But the child also, of course, *believes that the cup is empty*. This suggests that the crucial difference between imagination representations and beliefs is not given by the *content* of the representation. Rather, contemporary accounts of the imagination maintain that imagination representations differ from belief representations by their *function.* Just as desires are distinguished from beliefs by their characteristic functional roles, so too pretenses are distinguished from beliefs.

There are, of course, some obvious functional differences between believing and imagining. For one thing, the inputs to the imagination are at the whim of intention. We typically decide when to engage in an imaginative episode, and in many ways we can also control the particular contents that we imagine. As a result, we can fill out an imaginative episode in all kinds of surprising ways. Beliefs are much less whimsical. A second functional difference turns on the consequences of believing and imagining. When children engage in pretend play, they carry out behavioral sequences that conform in important ways to the actions they would perform if they really had the beliefs. Nonetheless, there are important behavioral discontinuities. When pretending that mud globs are delicious pies, even hungry children don't eat the "pies". Moreover, as adults, when we consume fiction, daydream, or fantasize, we don't typically produce actions that would be produced if we believed what we are imagining.

The third important point of consensus about the imagination is that imagination representations interact with some of the same mental mechanisms that belief representations interact with, and these shared mechanisms treat imagination representations and belief representations in much the same ways. That is, imagining and believing have shared pathways in the mind, and those pathways process imagination input and belief input in similar ways. For instance, most theorists maintain that imagination representations are processed by inferential mechanisms which also process belief representations. Consider again Leslie's experiment. Virtually all of the children in Leslie's experiment responded the same way when asked to point to the 'empty cup'. How are these orderly patterns to be explained? The prevailing cognitivist view is that the imagination representations are processed by the same inference mechanisms that operate over real beliefs.

To adopt a computational locution for this aspect of imagination, we say that imagination representations and beliefs are in the same "code" (Nichols 2004a). Of course, it's far from clear what the code is for belief representations, so it's not possible to be specific about the details or the nature of the putatively shared code. But the important point for present purposes can be made without giving further detail about what the code is. The key point is just that, if imagination representations and beliefs are in the same code, then mechanisms that take input from the "imagination box" and from the "belief box" will treat parallel representations much the same way.[8] For instance, if a mechanism takes imagination representations as input, the single code hypothesis maintains that if that mechanism is activated by the occurrent belief that *p*, it will also be activated by the occurrent imagination representation that *p*. More generally, for any mechanism that takes input from both the imagination box and the belief box, the imagination representation *p* will be processed much the same way as the belief representation *p*. Construed in this way, "single code" theories dominate the landscape in cognitive accounts of the imagination. The single code hypothesis is shared by theorists of quite different allegiances (e.g., Currie 1995, Doggett & Egan forthcoming, Gordon & Barker 1996, Harris 2000, Leslie 1987, Nichols & Stich 2000, Schroeder & Matheson 2006, Weinberg & Meskin 2006). Most prominently, off-line simulation theorists count as single code theorists. For they maintain that several mental mechanisms process 'pretend beliefs' just like real beliefs (e.g., Gordon 1986; Goldman 1989; Harris 1992) Off-line simulation theorists often have additional commitments of course. For instance, many prominent versions of off-line simulation theory explicitly invoke pretend desires in addition to pretend beliefs, and also maintain that the practical reasoning system takes as input pretend beliefs and pretend desires (e.g. Gordon 1986, Currie 1995). Those additional stipulations are consistent with, but not required by, the single code hypothesis.

One of the theoretical virtues of the single-code hypothesis is that it makes sense of the fact that the propositional imagination is not free from constraint. While the imagination is enormously flexible in the kinds of scenarios that can be entertained, there are clear and explicable limits. As is illustrated in Leslie's experiments, the imagination exhibits a kind of inferential orderliness (see also Harris & Kavanaugh 1993). This also explains why certain imaginings are rejected outright, like the suggestion that 1+1=7 (cf. Craig 1975; see also Gendler 2000; Moran 1994). The inferential mechanisms will balk at such a representation whether it comes from imagination or belief.

[8] Of course the claim is not that the mental processing of imagination representations will be *exactly* parallel to the processing of isomorphic belief representations (for discussion, see Nichols 2006).

4. Indexicals in imagination

It's a familiar feature of representationalist theories of mind that inference mechanisms are sensitive to the *format* of a representation, not the denotation of the representation (Fodor 1987). If John has a belief that we capture with the sentence "chickpeas are required for hummus" but he also has a (false) belief that corresponds to "garbanzo beans aren't chickpeas" then John is likely to infer that he's missing an ingredient for hummus when his pantry only contains cans that say "garbanzo beans". This is because what matters for inferential processing is the representation's format ("chickpeas") not, or not simply, the denotation. To take a case closer to hand, it matters whether I think that my wife's meeting starts at noon or whether I think it starts *now*. For if I don't know that it's now noon, then I won't draw the inference that my wife is currently at a meeting. In this case, the inference mechanisms are sensitive to whether the representation has the format of an indexical or not.

Now we have the pieces in place to begin exploring Williams' puzzle cases. Indexical concepts have a distinctive representational format, as indicated by the familiar points rehearsed in section 2. According to the single code hypothesis, indexicals in the imagination should get processed by the inferential mechanisms much as indexicals in belief get processed. And indeed, we find that the indexicals do still work in their impoverished way. Imagine the following: It's sometime in the future when you're cooking dinner and the pasta is done *now*. It's easy to imagine this, even though there is no descriptive content associated with *now*. *Now* can be deployed in the imagination without any further descriptive content. Again, this is what we'd expect on the single code theory. Something similar happens when I imagine that *I* am a rock star. The *I* doesn't come with all the descriptive (or causal-historical) residuals that characterize *Shaun Nichols*. Just as with the case of *now* in imagination, this is to be expected on the single code account, given the assumption that inferential processing is sensitive to representational format.

The imagination is not, it should be emphasized again at this point, an unfettered fantasy generator. As we saw with the case of Leslie's tea party, the imagination follows logical constraints. Moreover, some representations meet obstacles in the imagination. For instance, there is an obstacle to imagining that *2+2=6* or that *I am the Hope diamond*. But what is striking about Williams' Napoleon case is that the imagination allows it, despite its apparent incoherence. That is, there seems to be no obstacle to imagining that *I am Napoleon*; by contrast, there is an obstacle to imagining that *FDR is Napoleon*. Why then, does the imagination allow the thought that *I am Napoleon*? Well, we have the answer before us. There's no obstacle to imagining *I am Napoleon* because the I-concept is descriptively impoverished. Apart from minimal categorical information (like, *I* has to be an agent), there are no inferential restrictions on the I-concept. Since the imagination engages the same inferential device as belief, there will be no inferential restriction on imagining *I am Napoleon.*[9] While the inferential mechanisms pose numerous substantive constraints on what is allowed in the imagination, none of these constraints restrict imagining that *I am Napoleon*, and this is because of the poverty of the I-concept.

The explanation for why we can imagine being Napoleon also helps explain the other important case. In Williams' 1st person scenario, I imagine that all of my distinctive

[9] Note that this holds for belief as well. It's possible for a person other than Napoleon to have a belief of the form *I am Napoleon.* Amnesia plus delusions might generate such a belief.

psychological traits are removed, but it still seems intuitive that I persist to experience the torture. Part of the reason it's natural to conclude that I will be there for the torture is because of the supreme flexibility of the I-concept. Because of the flexibility of the *I*, there is no restriction against imagining that I persist in this case. In particular, the fact that all of my distinctive psychological properties are gone is no obstacle whatsoever. Given the poverty of *I*, there is no constraint against the representation *I exist in this location with completely different psychological properties*.

Obviously this is not a full explanation of the intuition in Williams' 1st person scenario. For it's not just that we *can* imagine persistence, we *do* imagine that we persist rather than expire in this case. I think that an additional part of the explanation of the intuition in Williams' first-person thought experiment hangs on another peculiar feature of imagining with indexicals. I argue elsewhere (Nichols 2007) that the single code account of imagination predicts that it should be difficult to have an imagining of the form *I don't exist*. Such an obstacle to imagining one's own nonexistence would further contribute to the intuition that I survive in Williams' scenario.[10]

5. Thought experiments and the self

Often thought experiments are an excellent technique for assessing what is and isn't possible. Different theorists have different accounts for why (and when) thought experiments succeed at this (e.g. Cooper 2005; Sorensen 1992). However, in light of the way imagination interacts with indexicals, I think we have reason to be deeply suspect of a certain class of thought experiments regarding the self.

Concerning the Napoleon case, Williams, like most subsequent commentators, maintains that it is really not possible for me to be Napoleon, despite the fact that it seems that I can imagine being Napoleon. I couldn't really be Napoleon because all of the available candidates for my self – memories, character, body, brain – are absent in that scenario. Nothing is preserved that could be the self. Hence, Williams maintains that something goes badly wrong in the Napoleon case. According to Williams, I can't really imagine that I am Napoleon in the way it might seem. What I actually imagine is something like this: I (i.e., *Napoleon*) am Napoleon. But I might confusedly think that what I am imagining is: I (i.e., *the real me*) am Napoleon (Williams 41-44; see also Blackburn 1997, 196). Williams thus explains away the Napoleon case by proposing a more respectable analysis of the *content* of my Napoleonic imaginings.

Even if Williams is right about the most charitable interpretation of the content of the Napoleonic imagining, I think that the Napoleon case provides an important and underappreciated lesson about the psychology of certain thought experiments. Williams maintains that the Napoleon puzzle arises because of a confusion. By contrast, in 'The Self and the Future,' Williams (tentatively) sides with the intuition delivered in his 1st person case (1970,

[10] This is still not a complete explanation of the Williams intuition, as Steve Yablo pointed out to me. For there is more to Williams' case. There will also be a second body that will have psychological properties that match mine before the operation. So why do we think that the self stays with the brain rather than move to the other body? At this point, I think the situation is quite unclear. There are significant empirical questions about what people's intuitions really are in these cases and about the factors that influence their intuitions. I leave this as a job for experimental philosophy.

180). The intuition that I would survive to feel the pain in this scenario is used to bolster the idea that the self is really the body. But if I'm right, the 1st-person case and the Napoleon case share a deep psychological root. Both cases depend on how the imagination interacts with the impoverished *I*. If that's correct, then the Napoleon case should be taken as a cautionary tale about the perils of imagining with the *I*. The diagnosis of what goes wrong, psychologically, in the Napoleon case suggests that the 1st person case has the same infection. Williams recommends following his 1st person case 'until we are shown what is wrong with it' (1970, 180). What is wrong with it is, in part, the very thing that is wrong with the Napoleon case.

In the context of beliefs, the poverty of the I-concept doesn't pose an especially serious problem. For when the I-concept is tokened in belief states (as in the belief that *I have a headache*), there will be a plausible referent for the I-concept. Of course there might still be difficulties about exactly what the right referent is in these cases.[11] But there will typically be something that will serve as a plausible referent of 'I' in I-beliefs. By contrast, the poverty of the *I* generates disastrous results in thought experiments, because it allows for imagining that *I* persist even while none of my distinguishing psychological or physical properties persist. Thus, it is dangerous to draw any metaphysical conclusions from these imaginative exercises with the *I*. Just as I am unwarranted in concluding that I *might have been* Napoleon, so too I'm unwarranted in concluding that I will persist to feel the pain in Williams' 1st person case.

It's worth considering how this plays out with other indexicals. Just as there is no obstacle to imagining that *I am Napoleon*, there is no obstacle to imagining that *it is now 1815*. *Now* is just as flexible and impoverished as *I*. However, this doesn't generate a problem in the case of *now* because we aren't even tempted to draw the defective metaphysical conclusions from the imaginative exercise. It would be preposterous to think that there is some particular time that is both *now* (the current time) and 1815. The error with *now* is obvious because we have a firm enough grip on time. We lack any such firm grip on the self. But that is no reason to trust the problematic interplay of imagination and indexicals when we try to plumb the nature of the self. Just as it would be a mistake to think that there could be a time that is both *now* and 1815, so it is a mistake to think that there could be a self that is both *me* and Napoleon.

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. 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.[12]

6. Folk dualism and the imagination

Recently a number of cognitive scientists have embraced the idea that dualism is the default view

[11] An enthusiast of evolutionary psychology might maintain that natural selection installed the *I* as an innate concept with the function of tracking the *organism*. But even if that's right, it would be hasty to conclude from this that the self is really the organism. Part of what constrains an account of the self is what we care about. And what we care about can diverge from what natural selection cares about.

[12] The poverty of the I might also facilitates the related view of the self as simple and identical across time, as suggested again by Kant on the paralogisms. (Johnston [1987] draws similar connections to Kant for similar purposes, see especially fn 11 and fn 16.)

for the folk. For instance, Paul Bloom writes

> we are dualists who have two ways of looking at the world: in terms of bodies and in terms of souls. A direct consequence of this dualism is the idea that bodies and souls are separate. And from this follow certain notions that we hold dear, including the concepts of self, identity, and life after death (Bloom 2004, 191).

In support of the thesis that the folk are dualists, Bloom notes that the implications for afterlife are in fact confirmed by delightful studies by Jesse Bering and colleagues. In these studies, children are told about Brown Mouse and Mr. Alligator, represented by puppets. After introducing the characters, the alligator eats the mouse and children are told, "Well, it looks like Brown Mouse got eaten by Mr. Alligator. Brown Mouse is not alive anymore." (Bering & Bjorklund 2004, 220). The children were subsequently asked several questions about the continuity of a biological trait and a related psychological trait. For example they were asked, "Now that the mouse is no longer alive, will he ever need to *drink water* again?" and "Is he still thirsty?". Another pairing was: "Does his brain still work?" and "Is he still *thinking* about Mr. Alligator?" For every single pairing, children were more likely to say that the psychological trait continued than they were to say that the biological trait continued (Bering & Bjorklund 2004, 224).

In addition to Bering's work on afterlife beliefs, recent work on reincarnation beliefs also fits with the idea that people are dualists. One idea associated with dualism is that the persistence of the self does not require preservation of distinguishing psychological traits (Reid 1785/1969). This kind of view is found in the doctrine of reincarnation without memory. Claire Cooper found that a majority of (largely Western) adults maintained that *if* they were reincarnated, the best indicator of their future self would be an autobiographical memory; but these participants also tended to think that the person would still be their reincarnation even if this last remaining vestige of their distinguishing psychological traits was subsequently lost (Cooper 2008).

People also have explicit beliefs about the soul (Richert & Harris 2006, 2008). In fact, people distinguish the soul from both body and mind. For instance, Richert & Harris (2006) found that children tended to think that the soul (but not the mind) changes upon baptism. And Western adults are more likely to maintain that the soul persists after death than that the mind does (2008, 107).[13]

Thus, a variety of different approaches have produced evidence that people have beliefs that fit will with the idea of a soul. Bloom intimates that people's views about the afterlife and personal identity flow from a prior notion of the soul: "A direct consequence of this dualism is the idea that bodies and souls are separate…. [F]rom this follow…the concepts of self, identity, and life after death" (Bloom 2004, 191; also 207). The notion of the soul itself might be innately specified on this account, as suggested by Bloom's claim that babies are "natural born dualists" (2004, xiii).

[13] It should be noted that there is considerable variability in explicit beliefs about the soul, including beliefs directly relevant to issues about self and identity. Richert & Harris (2008) found that while 52% of their undergraduate sample said that the soul changes, 28% maintained that the soul remains invariant throughout the biological lifespan (2008, 105). Furthermore, participants regard the soul as different from the mind on some dimensions (e.g., cessation at death) but not on others (e.g., importance for emotions) (2008, 107).

If babies are natural born dualists, that would also help explain the familiar anthropological claim that the notion of a soul is common across cultures (for some recent work, see Astuti & Harris forthcoming and Hardmann 2000). A very different, and much older, explanation for the prevalence of folk dualism comes from the 19th century anthropologist E. B. Tylor, who maintained that the notion of a soul emerges cross-culturally because it is a rational inference based on the available evidence (Tylor 1871).

Williams offers what seems to be yet another explanation for the common belief in the soul, one that draws on our imaginative activities. Williams suggests that in the exercise of imagining being Napoleon, 'If we press … hard enough, we readily get the idea that it is not necessary to being me that I should have any of the individuating properties that I do have, this body, these memories, etc… The limiting state of this progress is the Cartesian consciousness: an 'I' without body, past, or character' (Williams 1966, 41; see also Blackburn 1997). It's not entirely clear what Williams has in mind here. One interpretation is that we arrive at the idea of a soul by virtue of our imaginative activities. That is, we imagine existing without any of our current traits, and this leads us to conclude that there is something – a Cartesian 'I' – that really does get preserved across all the changes.

On the foregoing proposal, we would explain the ubiquity of dualism by maintaining that the imagination guides most people to the same Cartesian conclusion. But there is a more modest, and more plausible, way that the imagination might contribute to the ubiquity of folk dualism. This requires a brief diversion into cognitive anthropology. In trying to determine why we have the supernatural ideas we do, many anthropologists think it is more promising to set aside the question of origin (how did people come up with supernatural ideas?) in favor of the question of transmission – why did certain supernatural ideas catch on? (e.g. Sperber 1996; Boyer 1994). Facts about our psychology will play a central role in this kind of 'epidemiological' investigation. To know why certain ideas caught on, we need to know what the mind is like. Motivation and affect surely played an important role (e.g. Nichols 2004b). More interestingly for present purposes, some supernatural ideas are *cognitively* attractive, in ways that have been emphasized by Dan Sperber (1996) and Pascal Boyer (1994, 1999). Sperber has emphasized the role of cognitive modules in cultural transmission: "Mental modules… are crucial factors in cultural attraction. They tend to fix a lot of cultural content in and around the cognitive domain the processing of which they specialize in" (Sperber 1996, 113). And Boyer has argued that we can understand a great deal about the success of certain supernatural ideas by adverting to the role of modules in cultural transmission (Boyer 1994; 1999).

Sperber and Boyer focus on modules, but the basic approach can be applied without invoking modules. Some ideas can be cognitively attractive because they resonate with aspects of our psychology, regardless of whether the psychological resonance has a modular source. Given certain of our intellectual tendencies, some ideas will have greater cultural uptake than others. This epidemiological approach offers an alternative way to explain the prevalence of the idea of the soul.[14] We may never know the origin(s) of the idea of the soul. But it's plausible that the idea is a cultural achievement, rather than an innate idea or a rational inference. We can nonetheless explain the cross-cultural prominence of the idea of the soul by adverting to features that make the idea cognitively attractive. Imaginative tendencies constitute one important vector of cultural attraction. Williams' Napoleon example illustrates a kind of imaginative propensity

[14] As P.F. Strawson notes (1974, 175), the idea of a soul likely has several different sources. Here I'll focus on just one factor.

that fits neatly with the idea of a soul. His example suggests that I can imagine existing while having completely different distinguishing traits. The fact that we can imagine persisting even after losing our distinguishing traits would contribute to the cultural traction of the notion of a soul – a bare self that can persist through dramatic changes in traits. The role of the imagination here is indirect. I'm not suggesting that most people come to believe in a soul through their own self-directed imaginative activities. Rather, I'm suggesting that most people find the idea of a soul intuitive partly because of their imaginative propensities.

The way we think of the self is, of course, partly a function of what we can and can't imagine about the self. As Williams brought out, I can imagine persisting without my psychological or bodily traits. This imaginative flexibility shapes how we react to thought experiments regarding personal identity, and it also likely contributes to the intuitiveness of the idea of a soul. This imaginative flexibility regarding the self can be explained, I've suggested, by adverting to the poverty of the *I* and structure of the cognitive imagination. With this explanation in hand, we have good reason not to trust the intuitions that are delivered by this unruly form of imagining, imagining with the *I*.

Department of Philosophy
University of Arizona

References

Anscombe, E. 1975: The first person. In S. Guttenplan (ed.) *Mind and Language*. Oxford University Press, 45-65.

Astuti, R. and Harris, P. L. forthcoming. Understanding mortality and the life of the ancestors in rural Madagascar. *Cognitive Science.*

Bering, J. 2006: The folk psychology of souls. *Behavioural and Brain Sciences*, 29: 453-462.

Bering, J. and Bjorklund, D. 2004: The natural emergence of afterlife reasoning as a developmental regularity. *Developmental Psychology, 40*, 217-33.

Bering, J., Hernández-Blasi, C., and Bjorklund, D. 2005: The development of "afterlife" beliefs in religiously and secularly schooled children. *British Journal of Developmental Psychology*, 23, 587-607.

Blackburn, S. 1997: Has Kant refuted Parfit? In *Reading Parfit*, J. Dancy (ed.). Oxford: Blackwell.

Bloom, P. 2004: *Descartes' Baby*. New York: Basic Books.

Boyer, P. 1994: Cognitive constraints on cultural representations: natural ontologies and religious ideas. In L. Hirshfeld and S. Gelman (eds.), *Mapping the mind*. Cambridge, UK: Cambridge University Press, 391-411.

Boyer, P. 1999: Cognitive tracks of cultural inheritance: how evolved intuitive ontology governs cultural transmission. *American Anthropologist*, 100, 876-889.

Carruthers, P. 2006: *The Architecture of the Mind.* Oxford University Press.

Castañeda, H-N. 1999: *The Phenomeno-Logic of the I: Essays on Self-Consciousness.* Indiana University Press.

Cooper, C. 2008: *The Natural Foundations of Reincarnation Beliefs.* PhD. Dissertation, Queens' University, Belfast.

Cooper, R. 2005. Thought experiments. *Metaphilosophy,* 36, 328-347.

Craig, E. 1975: The problem of necessary truth. In *Meaning, Reference, and Necessity,* S. Blackburn (ed.). Cambridge: Cambridge University Press.
Currie, G. 1995: Imagination and simulation: Aesthetics meets cognitive science. In *Mental Simulation: Evaluations and Applications*, A. Stone and M. Davies (eds.). Oxford: Basil Blackwell.
Currie, G. and Ravenscroft, I. 2002: *Recreative Imagination*. Oxford: Oxford University Press.
Doggett, T. and Egan, A. forthcoming. Wanting things you don't want. *Philosophers' Imprint.*
Ezcurdia, M. 2002: Indexicals and demonstratives. *Encyclopedia of Cognitive Science*. MacMillan Publishing Company, 500-503.
Fodor, J. 1987: *Psychosemantics.* MIT Press.
Gendler, T. 2000: The puzzle of imaginative resistance. *Journal of Philosophy,* 97, 55-81.
Goddard, C. & Wierzbicka, A. 2002: *Meaning and Universal Grammar*. Philadelphia: John Benjamins.
Gordon, R. and Barker, J. 1994: Autism and the 'theory of mind' debate. In *Philosophical Psychopathology: A Book of Readings*, G. Graham & G. L. Stephens (eds.). Cambridge, MA: MIT Press.
Hardman, C. 2000: *Other worlds: Notions of self and emotion among the Lohorung Rai.* Berg, Oxford.
Harris, P. 2000: *The Work of the Imagination*. Blackwell.
Harris, P. and Kavanaugh, R. 1993: *Young children's understanding of pretense.* Monographs of the Society for Research in Child Development, (Serial No. 231).
Ismael, J. 2007: *The Situated Self*. New York: Oxford University Press.
Johnston, M. 1987: Human beings. *Journal of Philosophy*, 84, 59-83.
Kant, I. 1968: *Kant's Critique of Pure Reason.* Translated by N. Kemp Smith. New York: St. Martin's.
Kaplan, D. 1989: Demonstratives. In *Themes from Kaplan,* J. Almog, H. Wettstein, and J. Perry (eds.). Oxford University Press.
Leslie, A. 1987: Pretense and representation: The origins of 'theory of mind'. *Psychological Review*, 94, 412-426.
Leslie, A. 1994: Pretending and believing: Issues in the theory of ToMM. *Cognition*, 50, 211-238.
Machery, E., Mallon, R., Nichols, S., and Stich, S. 2004: Semantics, Cross-Cultural Style. *Cognition*, 92, B1-B12.
McGinn, C. 1993: *Problems in Philosophy: The Limits of Inquiry.* Blackwell.
Meskin, A. and Weinberg, J. 2003: Emotions, fiction, and cognitive architecture. *The British Journal of Aesthetics.*
Moran, R. 1994: The expression of feeling in the imagination. *Philosophical Review*, 103, 75-106.
Nichols, S. 2001: The Mind's 'I' and the Theory of Mind's 'I': Introspection and Two Concepts of Self. *Philosophical Topics*, 28, 171-199.
Nichols, S. 2004a: Imagining and believing: The promise of a single code. *Journal of Aesthetics and Art Criticism*, 62, 129-139.
Nichols, S. 2004b: Is Religion What We Want? Motivation and the Cultural Transmission of Religious Representations. *Journal of Cognition and Culture*, 4, 347-371.
Nichols, S. 2006: Just the Imagination: Why Imagining Doesn't Behave Like Believing. *Mind & Language* 21, 459-474.

Nichols, S. 2007: Imagination and immortality: Thinking of me. *Synthese*, 159, 215-233.
Nichols, S. and Stich, S. 2000: A cognitive theory of pretense. *Cognition*, 74, 115-147.
Perry, J. 1977: Frege on demonstratives. *Philosophical Review,* 86, 474-497.
Perry, J. 1979: The problem of the essential indexical. *Nous* 13, 3-21.
Recanati, F. 1993: *Direct Reference: From Language to Thought.* Cambridge, MA: MIT Press.
Reid, T. 1785/1969: *Essays on the Intellectual Powers of Man.* Edited by B. Brody. Cambridge, MA: MIT Press.
Reimer, M. forthcoming: Jonah cases. In *Empty Names,* A. Everett (ed.). Oxford University Press.
Rey, G. 1997: *Contemporary Philosophy of Mind.* Oxford: Blackwell.
Richert, R. and Harris, P. 2006: The ghost in my body: Children's developing concept of the soul. *Journal of Cognition and Culture* 6, 409-427.
Richert, R. and Harris, P. 2008: Dualism revisited: Body vs. Mind vs. Soul. *Journal of Cognition and Culture* 8, 99-115.
Robbins, P. 2002: The paradox of self-consciousness revisited. *Pacific Philosophical Quarterly,* 84, 424-443.
Schroeder, T. & Matheson, C. 2006: Imagination and emotion. In *The Architecture of the Imagination,* S. Nichols (ed.). Oxford: Oxford University Press.
Sider, T. 2001: Criteria of personal identity and the limits of conceptual analysis. *Philosophical Perspectives*, 15, 189-209.
Sorensen, R. 1992: *Thought Experiments*. New York: Oxford University Press.
Sperber, D. 1996: *Explaining culture*. Cambridge, Mass: Blackwell.
Strawson, P. 1974: *Freedom And Resentment And Other Essays*. London: Methuen & Co. Ltd.
Tappen, J., Williams, C., Fishman, S., & Touhy, T. 1999: Persistence of self in advanced Alzheimer's disease. *IMAGE Journal of Nursing Scholarship*, 31, 121–125.
Tylor, E. 1871: *Primitive Culture.*
Walton, K. 1990: *Mimesis as Make-Believe.* Harvard University Press.
Weinberg & Meskin 2006: Puzzling over the imagination: Philosophical problems, architectural solutions. In *The Architecture of the Imagination: New Essays on Pretense, Possibility, and Fiction,* S. Nichols (ed.). Oxford: Oxford University Press.
Williams, B. 1966: Imagination and the self. British Academy Annual Philosophical Lecture, reprinted in *Problems of the Self.* Cambridge: Cambridge University Press, 1973. Page references to reprinted version.
Williams, B. 1970: The Self and the Future. *The Philosophical Review*, 79, 161-180.