

Forthcoming in *Philosophical Psychology*, Special issue on experimental philosophy

## **Intuitions about Personal Identity: An Empirical Study**

Shaun Nichols

Department of Philosophy

University of Arizona

sbn@email.arizona.edu

Michael Bruno

Department of Philosophy

Lewis & Clark College

michael.bruno@gmail.com

*Williams (1970) argues that our intuitions about personal identity vary depending on how a given thought experiment is framed. Some frames lead us to think that persistence of self requires persistence of one's psychological characteristics; other frames lead us to think that the self persists even after the loss of one's distinctive psychological characteristics. The current paper takes an empirical approach to these issues. We find that framing does affect whether or not people judge that persistence of psychological characteristics is required for persistence of self. This difference is not explained by whether the case is framed in first or third person. By contrast, open-ended, abstract questions about what is required for survival tend to elicit responses that appeal to the importance of psychological characteristics. This emphasis on psychological characteristics is largely preserved even when participants are exposed to a concrete case that yields conflicting intuitions over whether memory must be*

*preserved in order for a person is to persist. Insofar as our philosophical theory of personal identity should be based on our intuitions, the results provide some support for the view that psychological characteristics really are critical for persistence of self.*

When is someone the same person across space and time? That is, when do two individuals at different places on different occasions count as quantitatively identical?<sup>1</sup> Following philosophical tradition, we can understand a *person* to be a more-or-less autonomous agent that thinks, acts on the basis of reasons, and is subject of perceptual experiences. When attempting to determine the conditions under which persons persist, ordinary reflection quickly leads to conflict. This suggests that devising an adequate account of *personal identity* is likely to be a highly non-trivial endeavor.

To illustrate the tension, suppose that Fiona is a 35-year old visual artist living in Portland. It seems obvious, nothing more than a piece of common sense, to say that Fiona is the same person as someone who had been born 35 years earlier, i.e. someone who had been named 'Fiona' by Fiona's parents and whose infant body is a physical precursor of Fiona's current body. It seems, that is, that Fiona *just is* her body and hence that she persists just when her body does. And yet, it also seems natural that Fiona would most readily identify herself with some constellation of values, beliefs, experiences and memories, i.e. with her current psychological make up. Fiona's psychological make up at 35, however, would be quite distinct from the psychological characteristics of the infant named 'Fiona' born 35 years ago. In fact, Fiona's psychology probably has much more in common with other 35-year old visual artists living in Portland than it does with the infant. Moreover, upon death, it seems that Fiona either ceases to exist, even though her body persists for some time, or continues to exist without her body, e.g. in some afterlife or as reincarnated into a new body. In either case, contrary to where we started, it seems that Fiona is not identical to her body. The

conditions under which she persists and the conditions under which her body persists come apart.<sup>ii</sup>

This apparent intuitive conflict is enshrined in philosophical thought experiments. Some thought experiments apparently lead to the intuition that psychology is what matters for the persistence of self; others apparently lead to the intuition that psychology is not what matters. Yet, academic philosophers did not invent this problem – its seeds are within us.

### **1. Williams' Personal Identity Thought Experiments**

The conflict that arises from these different ways of understanding identity receives a stark and well-known formulation in Bernard Williams' (1970) "The Self and the Future."

Williams' discussion of personal identity has generated enormous discussion over the last several decades.<sup>iii</sup> Adapting one of Locke's thought experiments, Williams initially presents a case in which two persons, A and B, will soon have their brains altered by some medical procedure. After the procedure, all of the psychological characteristics (e.g. memories and personality traits) that had been associated with A's brain will be associated with brain B; a parallel shift occurs for the psychological characteristics initially associated with B's brain.<sup>iv</sup> Before the procedure, A and B are told that one of the resulting persons will be tortured while the other will be given a large sum of money.

Williams claims that when we consider how the resulting individuals would regard the situation after the procedure, it seems that they would say that they have swapped bodies. After all, the person in the original B-body will remember things that happened to the original A, not the original B, and vice versa. Moreover, Williams has us imagine various requests that A and B might have made before the procedure. If A requests that the A-body gets tortured and B requests that the B-body gets tortured, then, after the procedure, the B-body person will recall directing the torture to the A-body. Correspondingly, the B-body

person will likely think that the preference he had stated prior to the procedure has been satisfied if the A-body person (now with B's psychological characteristics) is tortured, while the B-body person (now with A's psychological characteristics) receives the money. After considering alternative explanations of this case, Williams concludes of this case that "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" (1970, p. 167). The scenario presented in this thought experiment, Williams allows, seems to be a case in which A and B switch bodies. <sup>v</sup>

Williams then has us imagine something "apparently different":

Someone in whose power I am tells me that I am going to be tortured tomorrow. I am frightened, and look forward to tomorrow in great apprehension. He adds that when the time comes, I shall not remember being told that this was going to happen to me, since shortly before the torture something else will be done to me which will make me forget the announcement.... when the moment of torture comes, I shall not remember any of the things I am now in a position to remember (pp. 167-8).

In this case, you (the reader) are asked to imagine that you are before a captor who has informed you that you will be tortured tomorrow and that he will induce complete amnesia beforehand. In addition, your captor informs you that he will also extinguish all of your other distinctive psychological traits and insert false memories. After all that, he'll begin the torture. How should you react to the prospect of torture in this case? Williams says, "Fear, surely, would ... be the proper reaction" (p. 168). So, in this case, it seems like you'll presume that you will persist to feel the pain, despite the annihilation of your psychology.

Williams goes on to claim that any differences between these scenarios are superficial and philosophically insignificant. Indeed, they seem to be essentially the same case; all that differs is how each is framed. In the first thought experiment, which we'll label 'the *Lockean*

frame', amnesia will be induced and radical psychological alterations will be made before one of the bodies is tortured, and it seems like the appropriate reaction is to say that after those alterations, the original person would no longer be in his original body. It would be irrational for the person to fear the torture that would befall his (soon-to-be) former body. However, in the second thought experiment, which we'll label 'the *Pain* frame', a completely different reaction is elicited. In *Pain*, it seems quite sensible to fear the pain that will be experienced by the person with your original body, despite the amnesia.

## 2. Theories of Personal identity

The intuitions elicited by Williams' two frames point in opposite directions for developing a theory of personal identity. The *Lockean* frame suggests that the persistence of one's psychological characteristics is necessary and sufficient for the persistence of self. Under that frame, it seems that the original person A is transported into the B-body, since that is the body that has all of A's psychological characteristics after the procedure. The *Pain* frame suggests that persistence of one's psychological characteristics is *not* necessary for a person to persist. For under that frame, it seems that I persist (to feel the pain) even after my distinctive psychological characteristics have been eliminated.

Considered on its own, Williams' *Lockean* frame provides intuitive support for a *psychological criterion* for personal identity, which might be roughly put as follows:

Two person-slices, P1 at t1 and P2 at t2 are the same person (i.e. are quantitatively identical) if and only if P1 and P2 are psychologically continuous with each other.

Such theories have been widely discussed and defended in the philosophical literature.<sup>vi</sup> Our characterization of the view is deliberately vague. Psychological theorists can and have developed their theories in quite different ways, i.e. by specifying which kinds of psychological features they take to be critical and what kind of continuity they take to be

required.

What is most striking about Williams' *Pain* frame is that it seems to weigh against *any* psychological account. At a minimum, the psychological approach says that in order to survive, some psychological characteristics must be preserved. Yet, the *Pain* frame elicits intuitive resistance to just this claim. Rejecting a psychological criterion does not, of course, count as providing a positive account of personal identity. An obvious alternative, however, is to view the persistence of a person as instead dependent on some measure of physical, rather than psychological, connectedness. Roughly:

Two person-slices, P1 at t1 and P2 at t2 are the same person (i.e. quantitatively or numerically identical) if and only if P1 and P2 are physically continuous with each other.

Williams himself takes *Pain* to support this kind of alternative to the psychological approaches. In the contemporary literature, bodily or organismic approaches to personal identity are by far the most common alternative to psychological approaches.<sup>vii</sup> A proponent of a *bodily criterion* may, for instance, maintain that persons persist so long as their associated living organisms survive.

Another option, already alluded to in our discussion of Fiona, is that a person is identical to a *soul* or *immaterial ego*. Roughly:

Two person-slices, P1 at t1 and P2 at t2 are the same person (i.e. quantitatively or numerically identical) if and only if P1 and P2 have the same soul.

Such self-as-soul views of personal identity dominated philosophical thinking on the issue from Plato through the Early Modern period.<sup>viii</sup> Furthermore, given the prevalence of religious belief to this day, such views likely resonate and influence intuitions about personal identity, even amongst those who eschew belief in anything supernatural. Soul-based views of personal identity were the primary target of Locke's critical attention in his chapter on

personal identity. Soul-based approaches have fallen out of favor in contemporary philosophy. However, since our primary interest here concerns how the *folk* view of personal identity and how it bears on the problem of personal identity, we would do well not to exclude theories on the grounds that they are at odds with philosophical orthodoxy. As a result, we focus on the simple contrast between *psychological* views of personal identity and *non-psychological* views.

### **3. An Intractable Debate or a Job for Experimental Philosophy?**

In light of these conflicting intuitions to differently framed personal identity thought experiments, some have lamented that we cannot rely on such thought experiments for discerning the conditions under which persons are quantitatively identical across space and time. Carol Rovane, for instance, writes: “The real lesson of Williams’s discussion is that the Lockean thought experimental method should serve in the end to shake all conviction about the condition of personal identity” (1998, p. 44). That our intuitions about personal persistence seem so fickle provides good reason, Rovane suggests, not to trust our reaction to *either* of Williams’ cases. Ted Sider makes a similar point:

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, p. 198).

Rovane and Sider are probably right that no theory of personal identity will satisfy all intuitions about all devisable scenarios, and we’re sympathetic with the suggestion that the

solution is not to be found by proliferating thought experiments or drawing new distinctions. But, we are more sanguine that progress can be made by investigating *why* we have the intuitions we do. For if we can discredit or find additional support for the source of just one of the intuitions, then this might help us in determining whether or not persistence depends on our psychology. More generally, until we know more about what influences these responses, such pessimism seems premature. What factors generate the different intuitions? Why do we get pulled in different directions? What influences which direction we go? Answering these questions may well illuminate which candidate theory of personal identity is more plausible.<sup>ix</sup>

Williams himself credits this strategy by offering a tentative explanation for why we should distrust our responses to the *Lockean* frame (pp. 179-180). He suggests that the *Lockean* frame is guilty of something like leading the witness or experimenter demand in eliciting the description that the persons have “changed bodies” (p. 179). According to Williams, the problem with the case is that it has been artificially constructed in order to get the desired response: “it is the product of the will of the experimenter to produce a situation which would naturally elicit, with minimum hesitation, that description” (p. 179). Williams maintains that the case illicitly favors the psychological criterion by artificially stopping at the juncture most favorable to the intuition favoring the psychological theory. If the story ended before the original person’s memories were recreated in another’s brain, Williams contends it’s less clear that we would respond in accordance with the psychological theory. Furthermore, if the original person’s psychological characteristics were recreated in multiple donors, it’s also not clear that our responses would converge on a psychological theory (p. 180).

A further hypothesis is suggested in “The Self and the Future.” Williams observes that the asymmetry in responses corresponds with an interesting asymmetry in his

presentation: whereas the *Lockean* frame is offered in the third-person – it’s about A and B, the *Pain* frame is in the first-person – it’s about *me*. Williams writes:

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 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, p. 179).

What is surprising, Williams claims, is not that there is a self-other difference between first-person and third-person presentations. Rather, what’s surprising is that the difference trends in the way that it does – when we focus on ourselves in the *Pain* frame, our intuitions cut *against* the psychological theory.

These observations from Williams are worth serious consideration and further investigation. But the *kind* of attention they most need, we think, is *empirical* attention. Williams’ appreciation that framing can affect our intuitions anticipated a thriving research program in cognitive psychology on framing effects.<sup>x</sup> To determine whether Williams is right about how the frames influence our judgment requires controlled empirically study. Yet there is virtually no evidence on whether Williams is right even about the intuitiveness of the responses. Clearly this is a job for experimental philosophy.

#### **4. Experimental Evidence on Williams’ *Lockean* Frame**

Drawing inspiration from Williams’ work, we ran several pilot studies on versions of his cases. In early pilot studies, we found that Williams’ *Lockean* frame was very difficult for the participants to comprehend. This is perhaps not surprising, since Williams’ first presentation depends on the idea that psychological properties are entirely a function of neural properties,

which is not something that is fully accepted by the folk (to say the least). Given the difficulty the participants in our study had with these cases, we abandoned the attempt to test Williams' *Lockean* frame in its full form. In particular, we gave up the attempt to determine whether the folk regard persistence of psychological characteristics as *sufficient* for persistence of self.<sup>xi</sup> Rather, we focused our studies entirely on whether persistence of psychological factors was regarded as *necessary* for persistence of self. For ease of exposition, we will count any theory as a *psychological* theory of personal identity if it maintains that persistence or continuity of distinctive psychological factors is necessary for persistence of self.<sup>xii</sup>

In light of the difficulties with testing Williams' *Lockean* frame, we began to explore whether people would think that memory is at least *necessary* for survival in a *Lockean* style thought experiment. We were pleased to discover that the work had already been done for us. Sergey Blok, George Newman, and Lance Rips (2005) investigated whether people regard that memory as required for survival in a scenario in which a person's brain is transplanted into another body. In the relevant part of the experiment, participants were given the following scenario:

Jim is an accountant living in Chicago. One day, he is severely injured in a tragic car accident. His only chance for survival is participation in an advanced medical experiment called a "Type 2 transplant" procedure. Jim agrees.

It is the year 2020 and scientists are able to grow all parts of the human body, except for the brain. A stock of bodies is kept cryogenically frozen to be used as spare parts in the event of an emergency. In a "Type 2 transplant procedure," a team of doctors removes Jim's brain and carefully places it in a stock body. Jim's original body is destroyed in the operation. After the operation, all the right neural connections between the brain and the body have been made. The doctors test all physiological

responses and determine that the transplant recipient is alive and functioning. The doctors scan the brain of the transplant recipient and note that the memories in it are the same as those that were in the brain before the operation.

Participants were also presented with a parallel scenario in which the last sentence was altered to say that when the doctors scanned the brain they discovered that it had *no memories*, and that “something must have happened during the transplant”.

The results were overwhelmingly clear. People tended to agree with the claim that it was “still Jim” when the memories were preserved, but disagreed with the claim that it was “still Jim” when the memories were erased.<sup>xiii</sup>

This result conforms to the intuitive reaction anticipated from Williams’ *Lockean* frame. People in this experiment opt for a view that fits with a psychological approach to personal identity. Recall that one difference between Williams’ frames is that the *Lockean* frame is in the third-person and *Pain* is in the first-person. This, he intimates, might be part of the reason for the different reactions. There are several other differences between his two cases, of course. We can, however, run a self-oriented case closely matching the case from Blok et al (2005). That would give us a much better test of Williams’ suggested difference. To do this, we first replicated the results from Blok et al’s (2005) third-person version. Just as they had, we found that on the third-person presentation of these cases, participants tended to agree that the person persisted when the memories were preserved, but not when the memories failed to be preserved. The difference was statistically significant ( $t(28) = 3.17$ ,  $p < .01$ ).<sup>xiv</sup> To create a first-person version of the Blok cases, we simply replaced the proper names with the first-person pronoun. We found that once again, people tended to say that the person persisted in the cases where the memory was preserved, but not in the case where memory it wasn’t, and the difference was again statistically significant ( $t(42) = 3.33$ ,  $p = .01$ ).<sup>xv</sup> Furthermore, there was no significant differences when we compare responses on the first-

person vs. third-person conditions, either for the memory case ( $t(70)=.69$ ,  $p = .493$ , n.s.) or the no-memory case ( $t(70)=1.09$ ,  $p = .28$ , n.s.).<sup>xvi</sup> These results thus suggest that the basic response to the *Lockean* frame is not simply a function of whether the case is presented in the first-person or the third-person.

The results from these studies also bear on Williams' suggestion that we shouldn't trust our response to the *Lockean* frame because of the experimenter demand. Recall that Williams proposes that part of the reason we find the psychological-theory response so natural in the *Lockean* frame is because the psychological characteristics are recreated in another brain; if not for that feature, he suggests, we would be less inclined to favor a psychological response (p. 180). However, the results from the foregoing studies suggest that this does not explain our tendency to respond in concert with the psychological approach. For in the Blok et al (2005) study, as well as in our follow-up, all that is specified is that the memories are erased. There is no hint that the memories are recreated somewhere else. Yet the responses participants gave accords with a psychological criterion that holds that memory-preservation is necessary for persistence.<sup>xvii</sup>

## **5. Experimental Evidence on Williams' *Pain* Frame**

Although Blok et al (2005) provide some evidence that bears on Williams' *Lockean* frame, there is no evidence at all pertaining to Williams' *Pain* frame. As a result, we explored this in our own study. Participants were drawn from an introductory philosophy class at the University of Arizona. Personal identity had not been discussed in class at the time the survey was taken.

Given the difficulty earlier participants had in understanding Williams' *Lockean* frame, we developed a stripped down version of Williams' *Pain* frame. In place of the mad scientist bent on torturing me, we had helpful doctors trying to deal with a brain infection.

The scenario presented goes as follows:

Imagine that some time in the future your brain has developed a lethal infection and will stop functioning within a few hours. In the emergency room, you are alert and listening as the doctors explain to you that the only thing they can do is the following:

Render you completely unconscious, and then shave your head so that they can place electrodes on your scalp and shock your infected brain.

Unfortunately, this procedure will permanently eliminate your distinctive mental states (including your thoughts, memories, and personality traits).

You slip into unconsciousness before the doctors can discuss the matter further with you, and they elect to perform the procedure. It works exactly as expected. Several days after the procedure, the doctors perform some follow up brain scans and administer a series of painful shots.

After being presented with the scenario, subjects were asked a series of questions.

**Assume this happened as described. With that assumption in mind, please indicate which of the following claims you agree with.**

1. After the procedure, your distinctive mental states will be eliminated.

YES (AGREE)      NO (DISAGREE)

2. After the procedure, the infected brain will continue to function.

YES (AGREE)      NO (DISAGREE)

3. While they shave your head, you will see the doctors.

YES (AGREE)      NO (DISAGREE)

4. When the doctors administer the series of shots, *you* will feel the pain.

YES (AGREE)      NO (DISAGREE)

Obviously the last question is the crucial one. It parallels Williams' question "do you fear the pain?" We added the first three questions in an attempt to disguise somewhat the point of the

experiment. In addition, we wanted to have the option to use question 1 as a screening question (see below). After participants answered the Yes/No questions, they were asked to explain their answers.

As mentioned in section 4, our pilot studies showed that participants had difficulty understanding some of the personal identity thought experiments. To ameliorate this worry, we made our scenario as simple as we could. But we also wanted to be able to identify participants who we could be fairly confident *did* understand the scenario. This motivated the inclusion of the screening question (question 1) as well as the demand for an explanation. This allowed us to do a focused analysis on participants who passed the screening question (i.e., those who agreed with the statement in #1) and also gave adequate explanations for their answers. We independently coded the explanations for whether they reflected an understanding of the scenario (inter-rater agreement was very high, disagreements were resolved by discussion). Here are two examples that counted as adequate explanations: (i) “If the procedure went as planned, and all your thoughts, memories, experiences, and so on were eliminated then you would no longer be you”; (ii) “The doctors stated that you will lose your thoughts, memories, and personal thoughts, not your sense of feeling.” (The first participant disagreed with the statement “you will feel the pain”, the second participant agreed with it.) And here is an example of an explanation that did not count as adequate: “I liked it”.

Fortunately, the results of the study are basically the same regardless of whether we include all participants or only the participants who showed adequate comprehension. With all participants included, 68% agree with the statement “*you* will feel the pain,” which is significantly different from what one would expect by chance alone ( $\chi^2$  goodness-of-fit (1,  $N=50$ ) = 6.48,  $p = 0.01$ , 2-tailed). Given our antecedent worries about comprehension, however, we will focus on the participants who demonstrated comprehension of the scenario.

Among those participants 75% agree with the claim “*you* will feel the pain”. This is significantly different from what one would expect by chance alone ( $\chi^2$  goodness-of-fit (1,  $N=28$ ) = 7.000,  $p = 0.0082$ , 2-tailed). This suggests that Williams’ intuition about the case was in fact representative of our population of participants.

With this result in hand, we can now turn to examine Williams’ speculation that our responses to these scenarios might depend on whether the cases are framed in the first-person or third-person. We designed the above scenario to be in the first-person, since that followed Williams. But it’s easy enough to adapt it to a third-person scenario. So we designed a second survey along these lines:<sup>xviii</sup>

Imagine that some time in the future Jerry’s brain has developed a lethal infection and will stop functioning within a few hours. In the emergency room, Jerry is alert and listening as the doctors explain to him that the only thing they can do is the following:

Render him completely unconscious, and then shave his head so that they can place electrodes on his scalp and shock his infected brain. Unfortunately, this procedure will permanently eliminate Jerry’s distinctive mental states (including his thoughts, memories, and personality traits).

Jerry slips into unconsciousness before the doctors can discuss the matter further with him, and they elect to perform the procedure. It works exactly as expected. Several days after the procedure, the doctors perform some follow up brain scans and administer a series of painful shots.

As before, we had participants read this scenario and answer questions about it. The questions are matched to the earlier questions, except for the switch from first-person to third-person.

**Assume this happened as described. With that assumption in mind, please indicate which of the following claims you agree with.**

1. After the procedure, Jerry's distinctive mental states will be eliminated.

YES (AGREE)      NO (DISAGREE)

2. After the procedure, the infected brain will continue to function.

YES (AGREE)      NO (DISAGREE)

3. While they shave Jerry's head, he will see the doctors.

YES (AGREE)      NO (DISAGREE)

4. When the doctors administer the series of shots, *Jerry* will feel the pain.

YES (AGREE)      NO (DISAGREE)

As in the first-person condition, participants were asked to explain their answers. And as before, we identified the participants who gave adequate explanations for their answers and passed the screening question. Again, the results are the same whether we look at the entire sample or restrict our analysis to those who demonstrate adequate comprehension. With all participants included, 72% agree with the statement "Jerry will feel the pain" ( $\chi^2$  goodness-of-fit (1,  $N=54$ ) = 10.67,  $p < 0.01$ , 2-tailed). Focusing just on the group that demonstrated adequate comprehension, we also found that 72% agree that Jerry will feel the pain. This is significantly different from what one would expect by chance alone ( $\chi^2$  goodness-of-fit (1,  $N=36$ ) = 7.111,  $p = 0.0077$ , 2-tailed).

Thus, our version of Williams' *Pain* produces responses that run against the psychological approach to personal identity. Participants tend to say that they would feel the pain of the shots even after their distinctive psychology has been eradicated. But this is not restricted, as Williams suggested, to the first-person rendering of the case. We found an equally strong tendency to give answers inconsistent with a psychological theory in the third-person rendering as well. Indeed, there was no significant difference at all between the first-person and third-person conditions ( $\chi^2$  (1,  $N=64$ ) = 0.062,  $p = 0.8029$ , 2-tailed, n.s.). (See figure 1).

\*\*\*Figure 1 about here\*\*\*

## **6. Abstract/Concrete: A New Frame?**

Our results suggest that Williams was indeed on to something. By using different frames, modeled on his work, people can be led to give responses that either oppose or conform to the idea that persistence of self requires persistence of psychological characteristics. But *why* do the different frames produce their different effects? We have already encountered Williams' own speculative suggestion that the different responses might be a result of presenting in either first-or-third-person. But that proposal is not supported by the data.

Let's return to Williams' other suggestion about why we get the difference. Williams proposes that the *Lockean* frame works its magic by a kind of experimental demand, what we might dub *thought experimenter demand*. The case elicits the response that fits with psychological approaches to personal identity because it leads the witness down a path to that conclusion. Of course, this kind of charge is easy to make and correspondingly easy to disregard, especially in this context. After all, Williams' other frame seems to explicitly introduce a contrary thought experimental demand, since it begins by stating that someone "tells me I am going to be tortured tomorrow". Nevertheless, Williams makes specific suggestions about what changes to the case would make a difference to our intuitions. In particular, he suggests that we would be less inclined to give a response consistent with a psychological criterion if the thought experiment didn't stipulate that there would be a person with our psychological characteristics at the end of the process. Yet, as we saw, the results from Blok et al (2005) indicate that this doesn't make a difference. Even when there is no mention of another recipient of the person's psychological characteristics, people tended to say that if Jim's memories were gone, then the resulting person wouldn't be Jim.

Although Williams' particular suggestion looks to be mistaken, he was right to call attention to the possibility that framing might introduce thought experimenter bias. We suggest that there is such a bias in the frames, but that the *Pain* frame is the one at fault.<sup>xix</sup> In that frame, there seems to be a demand to respond that I would feel the pain. After all, if *I* am not going to feel it then *who is?* Similarly for the third-person version, if *Jerry* isn't going to feel the pain, then *who?* There is plausibly pressure here to give a persistence response. If this is right, then if we remove or decrease any thought experimenter demand, we should find less inclination to give the persistence response.<sup>xx</sup>

How, then, can we decrease the demands of the thought experiment? One simple idea is just to ask a very open ended question about what is required for personal identity. If we put such an *abstractly* framed question to participants, at a minimum this should reduce the particular kind of demand evinced by the *Pain* frame. In addition, if we ask the question in a way that makes no mention of psychology, we will have removed yet another demand characteristic. This simple idea is what we tried. We presented participants, again drawn from an introductory philosophy class, with an *abstract* question about personal identity and asked for a free response to the following probe:

One problem that philosophers wonder about is what makes a person the *same person* from one time to another. For instance, what is required for some person in the future to be the *same person* as you? What do you think is required for that? (Please pitch your explanation at a very simple level – don't use any words that might be unclear.)<sup>xxi</sup>

After answering this question, we probed participants specifically about the significance of memory to personal identity. Participants were asked whether they agreed or disagreed with the following statement:

In order for some person in the future to be *you*, that person doesn't need to have any of your memories.

YES (AGREE)                      NO (DISAGREE)

In their free responses, over 70% of the participants explicitly mentioned psychological factors like memory or personality traits as necessary for persistence. One representative answer was “That you have the same beliefs, values, and memories.” If we do the statistics on the free response questions, we find that the number of participants who invoked psychological factors was significantly greater than what would be expected by chance alone. ( $\chi^2$  goodness-of-fit (1,  $N=53$ ) = 9.98,  $p < .01$ ). When specifically asked about memory, the result was even stronger: more than 80% of participants disagreed with the claim that they could be identical to someone in the future who didn't have any of their memories. This is significantly different from chance ( $\chi^2$  goodness-of-fit (1,  $N=54$ ) = 21.41,  $p < .001$ ). It is also radically different from what we found in our version of Williams' *Pain* frame ( $\chi^2$  (1,  $N=88$ ) = 25.018,  $p < .001$ ). (See figure 2)

\*\*\*Figure 2 about here\*\*\*

Thus, when given an abstract query about personal identity participants respond in ways that cohere with a psychological approach to personal identity. This, we think, is an important outcome. Nevertheless, we cannot simply assume that people's responses to abstractly framed questions are invariably the most informative or reflective responses.<sup>xxii</sup> In some cases, it is likely that responses to abstract descriptions are less susceptible to errors. In other cases, it seems best to favor responses to concrete cases. The notorious Linda example from Tversky & Kahneman (1983) provides an example in which it seems like the response to the abstract question would be less error prone than the concrete version. In their study,

Linda is described as single, outspoken woman who was a philosophy major in college and was concerned with issues of discrimination. After reading this description, people tended to rate as more probable the statement “Linda is a feminist bank teller” than the independent statement “Linda is a bank teller”. But of course if Linda is a feminist bank teller she is also a bank teller. The Linda case is designed to trigger a certain kind of reasoning that leads to the wrong answer in this case. On Tversky & Kahneman’s model, the case triggers the “representativeness heuristic” (Tversky & Kahneman, 1974) because the description of Linda is representative of feminists. Obviously in order for this heuristic to get triggered, content details are required. As a result, by providing a more abstract formulation of the question, fewer errors of this sort can be expected.

In some cases, then, a more abstract description will likely lead to more reliable responses. In other cases, however, concrete cases likely generate more reliable responses. An obvious example comes from judgments about grammaticality. Consider the following abstract question: “Does a proper name have to come before any pronoun that is linked to it?” It’s likely that many English speakers would say “yes” to this abstract question. But this reflects a superficial thought about English grammar. Asking whether the following sentence is grammatical would reveal a deeper grasp of grammar: “Before his band broke up, Milo went to college.” It is clear that this sentence is well-formed, and the response to the abstract question should be disregarded.

## **7. The Folk on Reflective Equilibrium**

The foregoing suggests that there is no simple recipe for favoring abstract or concrete formulations. Perhaps it counts for something that the responses to the abstract question *do* converge with responses to some concrete questions, namely, those that follow Williams’ *Lockean* frame. But it would be better if there were another way of getting past the conflict.

One thing to note about the Linda case is that if people are given both the Linda case and an abstract version of the question, people will likely relinquish their initial intuition that it's more likely that Linda is a feminist bank teller. That is, when people are given both an abstractly formulated question and the concrete question about Linda, people will likely recognize that their answer to the abstractly formulated question is the right answer. This isn't a general advantage for abstractly formulated questions. For a similar phenomenon likely holds for cases like the grammar case. If people were asked both an abstract question about pronouns and a concrete question about the grammaticality of the sentence about Milo, people would presumably settle on their response to the sentence about Milo.

We ran one final experiment to investigate this issue. Again, participants were drawn from an introductory philosophy class, and the surveys were completed in class. For the task, all participants received both the case based on Williams' *Pain* frame (see section 5) as well as the abstract question (see section 6). The order of those questions was counterbalanced. These questions were then followed by a general question that effectively asks participants to engage in an exercise of reflective equilibrium:

Now that you have answered these questions, we want to call your attention to the fact that it wouldn't really be consistent to say both that you would feel the pain of the shots and also that in order for a person to be you, that person must have some of your memories. In light of this, which one are you more inclined to agree with? (check one please)

\_\_\_ More inclined to say that you would feel the pain in case #1.

\_\_\_ More inclined to say that in order for some person in the future to be you, that person must have some of your memories.

The point was to see whether people would show a preference for one judgment over the other. And we did find a preference. 64% of participants sided with the psychological

response under these reflective conditions, greater than what one would predict by chance alone ( $\chi^2$  goodness-of-fit (1,  $N=45$ ) = 3.76,  $p = .053$ , two-tailed).<sup>xxiii</sup> (see Figure 3)

\*\*\*Figure 3 about here\*\*\*

Once again, people's judgments favor the view that persistence of psychological features is required for persistence of self.<sup>xxiv</sup>

## 8. Conclusion

Much experimental philosophy attempts to debunk the philosophical appeal to intuitions.<sup>xxv</sup> Here we have opted for a more constructive strategy. We have not attempted to defend the role of intuitions in assessing theories of personal identity. Rather, we have tried to show that *if* it is appropriate for philosophers to rely on intuitions in assessing theories of personal identity, then it will help to identify which intuitions are especially robust. If intuitions are supposed to guide theory building, then, *ceteris paribus*, intuitions that are more robust ought to be given more weight.

Over the course of our experiments, we found that there that intuitions favoring a psychological approach to personal identity are resilient across significant changes in the cases. Those intuitions also converge with the judgments people make when simply asked an abstract question about what is required for persistence. Ordinary folk are inclined to continue to hold on to this psychological-based view even when it is presented alongside a case that elicits anti-psychological intuitions.

Obviously we don't think that in general the way to solve philosophical problems is to ask undergraduates. However, at the same time, it would be reckless to disregard their considered judgments. As individual philosophers we can persevere on our responses to

certain thought experiments, losing a broader perspective. Collective responses can bring that broader perspective back to us. So while Williams was right that his case elicits intuitions at odds with psychological approaches to personal identity, those intuitions don't seem to stand to scrutiny when examined in a broader and aggregative context.

We think all of this points in favor of resolving Williams' intuitive conflict in favor of a psychological view of personal identity (assuming that intuitions should inform our theories of personal identity). But this is hardly a decisive finale to the issue. For one thing, there are other considerations that might move people's intuitions away from the psychological approach. For instance, it is plausible that the psychological approach to personal identity may be less intuitive for parents thinking about their children. In addition, our own experiments focus entirely on Western undergraduates. This sample is homogenous on several important factors, including age, culture, and socioeconomic status. It's quite possible that people in different cultures or age or socioeconomic groups will respond differently from the population we studied. And even *within* the population we studied, there was far from uniform agreement about the questions. Other work on intuitions suggests that there are systematic individual differences in people's intuitions about philosophical issues and that might well be the case for the problem at hand.<sup>xxvi</sup> Nevertheless, we've tried to illustrate how more careful attention to the influence of framing effects can advance philosophical debate, rather than simply debunk the presumptions upon which such debates depend. In particular, we suspect that a more nuanced understanding of the distinction between abstract and concrete framing may well shed light not just on the problem of personal identity, but also on many other, seemingly intractable, philosophical problems.

## **Acknowledgements**

We'd like to thank Dan Bartels, Brian Detweiler-Bedell, Emily Esch, Danielle Fagre, Brian Fiala, Tamar Gendler, Clara Laurence, Sarah Nichols, Hannah Tierney and Jen Zamzow for discussion and comments on this work.

## References

- Alexander, J., Mallon, R., & Weinberg, J. (forthcoming). Accentuate the negative. *European Review of Philosophy*.
- Ayer, A. J., (1936). *Language, truth, and logic*. London: Gollancz.
- Ayers, M. (1990). *Locke*, Vol.2. London: Routledge.
- Bartels, D. and Rips, L. (forthcoming). Psychological connectedness and intertemporal choice. *Journal of Experimental Psychology: General*.
- Bartels, D. and Urminsky, O. (forthcoming). On intemporal selfishness: The perceived instability of identity underlies impatient consumption.
- Blackburn, S. (1997). Has Kant refuted Parfit? In J. Dancy (Ed.). *Reading Parfit*, Oxford: Blackwell.
- Blok, S., Newman, G., & Rips, L. J. (2005). Individuals and their concepts. In W.-K. Ahn, R. L. Goldstone, B. C. Love, A. B. Markman, & P. Wolff (Eds.), *Categorization Inside and Outside the Laboratory*. Washington, DC: American Psychological Association, 127–149.
- Butler, J. (1736). *The Analogy of Religion*. London: J., J. and P. Knapton.
- Carter, W. (1990). Why personal identity is animal identity. *LOGOS*, 11, 71-81.
- Cokely, E. & Feltz, A. (forthcoming). Understanding variation in folk judgment and intuition. *Consciousness and Cognition*.
- Feltz, A., & Cokely, E.T. (forthcoming). Do judgments about freedom and responsibility depend on who you are? Personality differences in intuitions about compatibilism and incompatibilism. *Consciousness and Cognition*.

- Freiman, C. & Nichols, S. (forthcoming). Is desert in the details? *Philosophy and Phenomenological Research*.
- Gallie, I. (1936). Is the self a substance? *Mind*, 45, 28-44.
- Gendler, T. (1998). Exceptional persons: On the limits of imaginary cases. *Journal of Consciousness Studies*, 5, 592–610.
- Gendler, T. (2007). Philosophical thought experiments, intuitions and cognitive equilibrium. *Midwest Studies in Philosophy*.
- Grice, H.P. (1941). Personal identity. *Mind*, 50, 330-350.
- Lewis, D. (1976). Survival and identity. In *The Identities of Persons*, A. Rorty (ed.), Berkeley, CA: University of California Press.
- Locke, J. (1710/1975). *An essay concerning human understanding*. Oxford: Clarendon Press.
- Nichols, S. (2008). Imagination and the I. *Mind & Language*, 23, 518-535.
- Nichols, S. and Knobe, J. (2007). Moral responsibility and determinism: The cognitive science of folk intuitions. *Noûs*, 41, 663-685.
- Nichols, S. and Ulatowski, J. (2007). Intuitions and individual differences: The Knobe effect revisited. *Mind & Language*, 22, 346-365.
- Olson, E. (1997). *The human animal: Personal identity without psychology*. Oxford University Press.
- Parfit, D. (1984/1987). *Reasons and persons* Oxford: Clarendon Press.
- Piattelli-Palmarini, M. (1996). *Inevitable illusions: how mistakes of reason rule our minds*. Wiley.
- Perry, J. (1972). Can the self divide? *The Journal of Philosophy*, 69, 463-88
- Pollock, J. & Ismael, J. (2006). So you think you exist? – in defense of nolipsism. In T. Crisp, M. Davidson, & D. Vander Laan (Eds.), *Knowledge and reality: essay in honor of Alvin Plantinga*. Springer Verlag.

- Reid, T. (1785). *Essays on the intellectual powers of man*. Edinburgh: J. Bell.
- Rips, L., Blok, S. & Newman, G. (2006). Tracing the identity of objects. *Psychological Review*, 113, 1–30.
- Rovane, C. (1998). *The bounds of agency*. Princeton: Princeton University Press.
- Shoemaker, S. (1963). *Self-knowledge and self-identity*. Ithaca: Cornell University Press.
- Shoemaker, S. (1984). Personal identity: A materialist's account. In S. Shoemaker and R. Swinburne, *Personal Identity: Great Debates in Philosophy*. Oxford: Blackwell.
- Sider, T. (2001). Criteria of personal identity and the limits of conceptual analysis. *Philosophical Perspectives*, 15, 189-209.
- Sinnott-Armstrong, W. (2008). Abstract + Concrete = Paradox. In J. Knobe and S. Nichols, *Experimental Philosophy*. New York: Oxford University Press.
- Snowdon, P., (1990). Persons, animals, and ourselves. In C. Gill (Ed.), *The Person and the Human Mind*. Oxford: Clarendon Press.
- Swain, S., Alexander, J., & Weinberg, J.M. 2008. "The Instability of Philosophical Intuitions. *Philosophy and Phenomenological Research*, 76, 138-155.
- Swinburne, R. (1984). Personal identity: The dualist theory. In S. Shoemaker and R. Swinburne, *Personal Identity: Great Debates in Philosophy*. Oxford: Blackwell.
- Thomson, J. (1987). Ruminations on an account of personal identity. In J. Thomson (Ed.) *On Being and Saying: Essays for Richard Cartwright*. The MIT Press.
- Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, 185, 1124-1131.
- Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*, 211, 453-458.
- Tversky, A., & Kahneman, D. (1983). Extensional vs. intuitive reasoning: the conjunction fallacy in probability judgment. *Psychological Review*, 90, 293–315.

Unger, P. (1990). *Identity, consciousness and value*. New York: Oxford U.P.

van Inwagen, P. (1980). Philosophers and the words 'human body'. In P. van Inwagen (Ed.)

*Time and Cause*, Reidel.

Weinberg, J., Nichols, S., & Stich, S. 2001. "Normativity and Epistemic Intuitions."

*Philosophical Topics*, 29, 429-460.

Williams, B. (1956-7). Personal identity and individuation. *Proceedings of the Aristotelian*

*Society*, 57, 229-252.

Williams, B. (1970). The self and the future. *The Philosophical Review*, 79, 161-180.

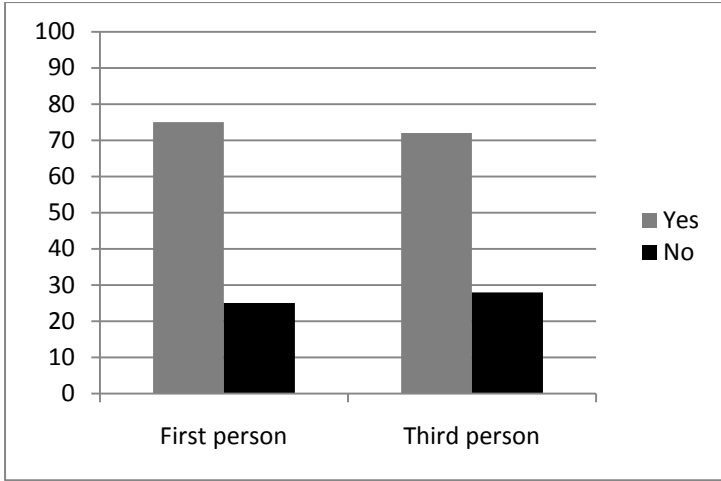


Figure 1: Experiments on Williams' *Pain* frame: Will you/Jerry feel the pain?

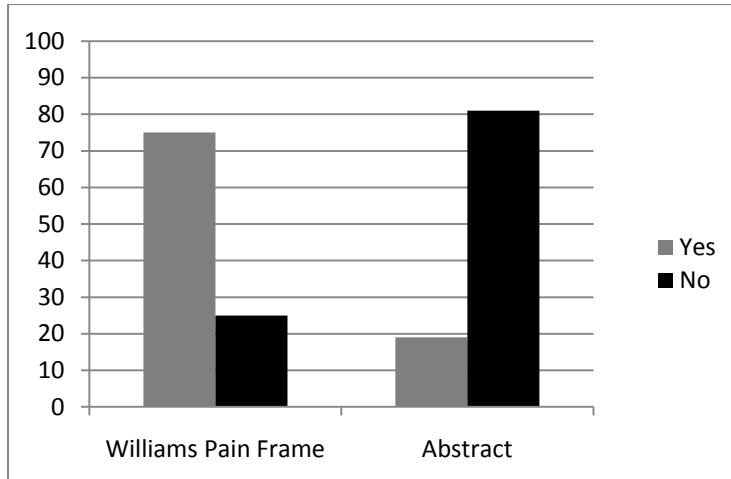


Figure 2: Can I persist if my memories are gone?

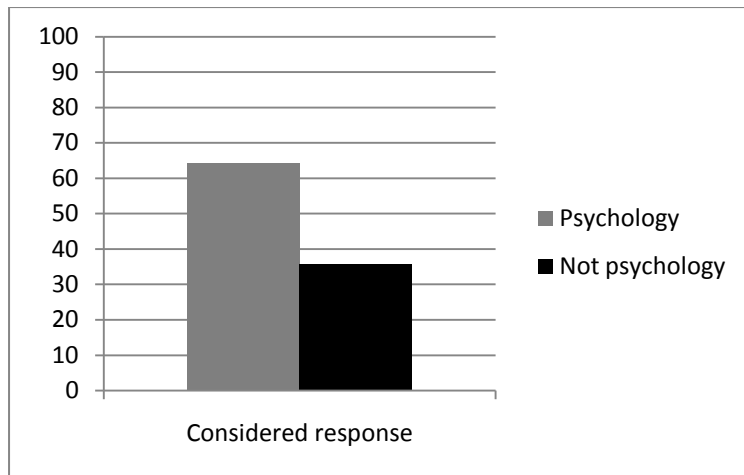


Figure 3: On reflection, can I persist if my memories are gone?

## Notes

---

<sup>i</sup> Quantitative (or numerical) identity is a transitive, symmetric and reflexive relation. It is important to keep in mind that it is different that qualitative identity. Two cars produced by the same auto manufacturer may well be qualitatively identical (e.g. same model, color, etc.), but they are clearly not quantitatively identical. A similar remark applies for so-called ‘identical’ twins. Philosophical debates over personal identity concern quantitative, not qualitative, identity.

---

<sup>ii</sup> On this way of setting up the philosophical issue about persistence, see Williams (1956-7, p. 229): “There is a special problem about personal identity for two reasons. The first is self-consciousness – the fact that there seems to be a peculiar sense in which a man is conscious of his own identity... The second reason is that a question of personal identity is evidently not answered merely by deciding the identity of a certain physical body.”

<sup>iii</sup> E.g., see Blackburn (1997), Gendler (1998 & 2007), Nichols (2008), Parfit (1984), Rovane (1998), and Sider (2001).

<sup>iv</sup> More precisely, the case specifies that the psychological characteristics of the person with A's body *after* the procedure will be *qualitatively identical* to psychological characteristics of the person with B's body *before* the procedure, and vice versa. In this paper, nothing of philosophical significance will turn on being imprecise on this matter, so we will not attempt to be more precise in the text.

<sup>v</sup> The most famous body swapping case in the personal identity literature is Locke's (II.xxviii.xv) tale of the prince and the cobbler: “Should the soul of a prince, carrying with it the consciousness of the prince's past life, enter and inform the body of a cobbler, as soon as deserted by his own soul, every one sees he would be the same person with the prince, accountable only for the prince's actions: but who would say it was the same man?” The cases involving ‘body-swapping’ are familiar to non-academics from works fiction, e.g. in the films *Big* and *Freaky Friday*. Shoemaker (1963) employed a version of Locke's case involving the swapping of brains between bodies to argue for psychological account of personal persistence. Williams (1970) is largely intended to undermine ‘body-swapping’ arguments.

<sup>vi</sup> E.g., see Lewis (1976), Locke (1710), Parfit (1984), Perry (1972), Shoemaker (1963 & 1984), and Unger (1990).

<sup>vii</sup> E.g., see Ayer (1936), Ayers (1990), Carter (1990), Olson (1997), Snowdon (1990), Thomson (1987), van Inwagen (1980), and Williams (1956-7 & 1970).

<sup>viii</sup> Butler (1736) and Reid (1785) provide influential defenses of the soul-based view. For fascinating discussion of the debate over personal identity in the 18<sup>th</sup> and 19<sup>th</sup> Centuries, and how some of the forgotten maneuvers made in the course of these debates were recapitulated in second half of the 20<sup>th</sup> Century, see Martin & Barresi (2000). Soul-based approaches have fallen out of favor in contemporary philosophy (though see Swinburne, 1984). Most objections to soul-based views rely on the fact that those views require understanding persons as substances (cf. Locke's prince and the cobbler case). For an insightful, but neglected, debate over

---

whether selves are substances from the first half of the 20<sup>th</sup> Century, see Gallie's (1936) defense of a substantialist view and Grice's (1941) subsequent criticism and defense of a Lockean view.

<sup>ix</sup> Pollock & Ismael (2006) employ a similar general strategy, viz. they attempt shed light on the problem of personal identity explaining how *de se* representations, i.e. non-descriptive, reflexive, representations of one's self, are both required for cognition and are responsible for the conflicting intuitions people have to various personal identity thought experiments. Unlike Rovane (1998) and Sider (2001), their investigation of why we have these intuitions leads them to adopt what they call "Nolipsism", basically the view that, strictly speaking, persons do not exist. Nichols (2008) offers a related explanation for why Williams' *Pain* frame generates an untrustworthy response. The claim there is that the concept <I> is excessively flexible in imaginative activities, allowing it to seem applicable in an imaginative scenario even when no distinctive features are preserved.

<sup>x</sup> See Tversky & Kahneman (1981). For a popular treatment of these issues, see Piatelli-Palmarini (1996).

<sup>xi</sup> This means that we could not test Williams' hypothesis (p. 180) that we would be less inclined toward psychological theory responses if the psychological states were recreated in multiple donors. Similarly, we do not explore intuitions about fission or teletransportation, since those too involve judgments about what is sufficient to ensure persistence of self.

<sup>xii</sup> It is typical to identify theories as psychological when the theories maintain that psychological factors are both necessary and sufficient for personal identity. So we are adopting a more liberal category.

<sup>xiii</sup> See also Rips et al. (2006, pp. 7-8).

<sup>xiv</sup> Although we replicated their effect, we would note that our means were generally closer to midline than theirs. In the case of preserved memory, the mean response was 5.45 (on 0-9 scale); in the no-memory case, the mean response was 3.24.

<sup>xv</sup> Again the mean responses were closer to the midline than in Blok et al.'s (2005) study. When memory was preserved, the mean response was 5.95 (on 0-9 scale); when memory wasn't preserved, mean response was 4.09.

<sup>xvi</sup> Although we did not find a significant difference, we think that the Blok-style cases deserve much more study. There *was* a difference in the mean responses between first-person and third-person conditions for the no memory case, but it didn't come close to being a statistically significant difference. Nonetheless, it would be worth exploring whether changing the experiment would lead to significant differences along this dimension. Perhaps the most immediate question is whether the results would be different if the memory/no-memory conditions were run between participants. We followed Blok et al. in having all participants evaluate both cases,

---

and hence used a within-subject design. Different responses might well emerge in a between-subject design. Although we hope to see empirical work along these lines in the future, the present point stands that memory matters even in the first-person presentation.

<sup>xvii</sup> Recent work by Dan Bartels and colleagues provides further evidence that at least under certain conditions; people seem to think of the self in terms of psychological continuity (Bartels & Rips, forthcoming; Bartels & Urminsky forthcoming). Their ingenious experiments exploited the well-documented phenomenon of temporal discounting – people tend to prefer smaller rewards sooner rather than larger rewards later. What Bartels and colleagues have found is that degree of expected psychological continuity is a critical factor in the extent to which participants are willing to wait for a bigger reward. People are less willing to wait for a bigger reward if they expect to undergo significant psychological changes before the time of the reward.

<sup>xviii</sup> Again, the participants were students in an introductory philosophy class at the University of Arizona in which the topic of personal identity had not been discussed.

<sup>xix</sup> In Nichols 2008 (see note *ix*), it was assumed that Williams was right that the reaction to the *Pain* frame is largely driven by the fact that the case is presented in the first-person. The data reported in the current paper undercuts that assumption since we found that participants responded the same way to Williams' *Pain* frame even when it was presented in the 3<sup>rd</sup> person. Nonetheless, the <I> concept might still play a critical role in these cases. For one of the most important resources for understanding others is projecting from one's own case. So it's possible that our response to the 3<sup>rd</sup> person version of the pain frame is still partly a function of how the <I> concept interacts with the imagination (see Nichols 2008).

<sup>xx</sup> Another possibility, suggested by Emily Esch (personal communication), is that the responses to Williams' *Pain* Frame might be driven by the fact that the mental state is *pain*. A less vivid mental state, like *itch*, might generate different responses to the case.

<sup>xxi</sup> In pilot studies, we found that the open-ended question often triggered religious or spiritual responses. This prompted us to include the parenthetical request that the answer be written in plain and simple terms.

<sup>xxii</sup> See, e.g., Freiman & Nichols (forthcoming), Gendler (2007), Nichols & Knobe (2007), and Sinnott-Armstrong (2008).

<sup>xxiii</sup> There was a sizable minority – 36% did not respond in accordance with the psychological approach – and there are important questions to explore regarding this diversity. One key question is whether the responses of the minority reflect some stable pattern of individual differences (see, e.g., Nichols & Ulatowski, 2007). Do personality traits correlate with differences in responses here? (see, e.g. Cokely & Feltz, forthcoming). Are

---

there cultural or religious factors that drive these differences? Are the responses just noise generated by the strange nature of the question? We have no serious hypotheses at this point, but the questions seem quite tractable.

<sup>xxiv</sup> By contrast, when Nichols & Knobe (2007) had participants respond in a similar reflective equilibrium condition concerning intuitions about compatibilism vs. incompatibilism, they found that the philosophical disagreement was simply recapitulated in the sample of participants. Participants split roughly 50/50 on their considered responses.

<sup>xxv</sup> E.g., see Weinberg et al. (2001), Swain et al. (2008), and Alexander et al. (forthcoming).

<sup>xxvi</sup> See, e.g. Feltz & Cokely (forthcoming) and Nichols & Ulatowski (2007).