WILLIAMS ON THE SELF AND THE FUTURE*

Dilip Ninan
dilip.ninan@tufts.edu
March 2021

1

Bernard Williams's influential essay 'The Self and the Future" (Williams 1970) focuses on a type of thought experiment frequently discussed in the literature on personal identity over time: so-called 'body-switching' cases. Williams makes a number of intriguing observations about such cases, but here I shall focus only on one, namely his claim that imagining a particular thought experiment from the first-person point of view supports the bodily continuity theory, whereas imagining the same case from the third-person point of view supports the psychological continuity theory. Much of the subsequent discussion of Williams's essay has been concerned with what to make of this fact. Does it show that our judgments about these thought experiments are unreliable, as Rovane (1998) and Szabó Gendler (1998) argue? Or does it show that there is no fact of the matter as to whether personal identity is a matter of psychological continuity as opposed to bodily continuity, as Sider (2001) argues?

But these discussions are premised on a falsehood. Against Williams and subsequent commentators, I argue that imagining the thought experiment from the first-person point of view supports not the bodily continuity theory, but what Parfit (1984) calls the *simple view*, the view that facts about personal identity are independent of facts about bodily and psychological continuity. While this point is arguably an instance of something more general—many personal identity thought experiments, when viewed from the first-person point of view, seem to support the simple view—I shall focus my attention on making the point in the context of Williams's thought experiment.

*Forthcoming in *Analytic Philosophy*. Thanks to an anonymous referee for that journal.

2

Let us begin by characterizing the simple view and its rival, the *complex view*, a bit more precisely. The complex view says that personal identity is essentially a matter of some kind of psychological or physical continuity; the simple view denies this. A natural way to understand this dispute is in terms of the acceptance or rejection of a certain supervenience thesis. The complex view implies that personal identity supervenes on lower-level continuity relations, like psychological, bodily, and brain continuity, whereas the simple view denies this. Let us say that a *person stage* is a pair $\langle \sigma, t \rangle$ of a person σ and a time t at which the person exists. A pair of person stages $\langle \sigma, t \rangle$, $\langle \sigma', t' \rangle$ are *stages of the same person* just in case $\sigma = \sigma'$. Let us say that two pairs of possible person stages, $\langle x, y \rangle$ in possible world w and $\langle x', y' \rangle$ in possible world w', are *the same with respect to a relation R* iff (Rxy in w iff $Rx'y'$ in w').[1] And let us say that two pairs of possible person stages $\langle x, y \rangle$ in w and $\langle x', y' \rangle$ in w' are *the same with respect to continuity* iff $\langle x, y \rangle$ in w and $\langle x', y' \rangle$ in w' are the same with respect to the relations of psychological, bodily, and brain continuity. Then we can state our supervenience thesis as follows.

> Supervenience
>
> For all worlds w, w', and pairs of person stages $\langle x, y \rangle$ in w, $\langle x', y' \rangle$ in w': if $\langle x, y \rangle$ in w and $\langle x', y' \rangle$ in w' are the same with respect to continuity, then x and y are stages of the same person in w just in case x' and y' are stages of the same person in w'.

So we shall assume that the complex view entails Supervenience and the simple view entails its negation. Note that the psychological, bodily, and brain continuity theories of personal identity all entail Supervenience. For example, proponents of the psychological continuity theory argue for the following view of personal identity:

> For all worlds w and pairs of person stages $\langle x, y \rangle$ in w: x and y are stages of the same person in w just in case x and y are related by the relation of psychological continuity in w.[2]

[1] Note here that x is itself a pair $\langle \sigma, t \rangle$ of a person and a time (similarly for y, x', y').

[2] What is the relation of psychological continuity? Two person stages x and y are *psychologically connected* iff x and y have psychological states (beliefs, desires, intentions, apparent memories, character traits, etc.) which are (a) similar in content, and (b) causally connected in the right way, i.e. the psychological states of the later person stage causally depend (in the right way) for their character on the states of the earlier one (Lewis 1976). *Psychological continuity* is then the transitive closure of psychological connectedness.

This theory makes the personal identity relation and the relation of psychological continuity intensionally equivalent. The bodily continuity theory, on the other hand, make the personal identity relation and the relation of bodily continuity intensionally equivalent. Since both these relations are in the relevant supervenience base, both theories entail Supervenience. So Supervenience captures at least part of what these theories have in common.

3

Williams (1970) invites us to suppose that there is a device capable of 'extracting' all or most of the information (beliefs, desires, intentions, apparent memories, character traits, etc.) from a person's brain. And we are to suppose that the information can then be 're-inserted' back into that person's brain. Imagine two persons, *A* and *B*, each entering a similar machine, but one which extracts all the psychological information out of each brain and then inserts it into the brain which originally belonged (and may still belong) to the other person, so that, after this procedure, the person in the *A*-body (i.e. *A*'s original body) now has all the apparent memories, thoughts, feelings, etc. that *B* had before the procedure. And similarly, the person in the *B*-body (i.e. *B*'s original body) now has all the apparent memories, thoughts, feelings, etc. that *A* had before the procedure.

The question now is: to whom does each body belong? Does the *B*-body still belong to *B* or does it now belong to *A*? Williams observes that most of us are inclined to say that the *B*-body now belongs to *A*, and that the *A*-body now belongs to *B*. For the person in the *B*-body will have all of *A*'s apparent memories, and none of *B*'s. Furthermore, having *A*'s beliefs, desires, and character traits, the person in the *B*-body will tend to act and talk just as *A* acted and talked. And, of course, this person will think that he is *A*, since he has all of *A*'s beliefs and apparent memories. All this suggests that the person in the *B*-body *is A*. Similar considerations suggest that *B* is the person in the *A*-body after the experiment.

For these reasons, the case, as Williams notes, seems to be one in which two people 'change bodies' (Williams 1970, 51). And this seems to show that who a person will be in the future depends on psychological, rather than on bodily or brain, continuity. Note also that our initial reaction to this case seems to support Supervenience, since Supervenience leads us to expect that if we fix the continuity facts in a given situation, then only one possibility for the personal identity facts will

be compatible with our description of that situation. If the continuity facts did not fix the person facts, there might be two possible situations consistent with Williams's description: one in which the participants switch bodies, another in which they remain in their respective bodies. That we think the latter is not a possibility consistent with the specified continuity facts suggests that we think that once the continuity facts are fixed in the way Williams fixes them, the relevant person facts are also fixed.

So far, the case is not a puzzle—it's simply a case which seems to support the psychological continuity theory. The puzzle is generated by placing the above description of the case next to a description of the case from the perspective of one of its participants. My description is adapted from Williams (1970, 51–52):

> 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. This certainly will not cheer me up. He goes on to tell me that when the moment of torture comes, I shall not remember any of the things I am now in a position to remember. He also tells me that, at the moment of torture, I shall not only not remember the things that I am now in a position to remember, but will have a different set of impressions of my past, quite different from the memories I have now. I do not think that this would cheer me up either. Nor do I see why I should be put into any better frame of mind by the person in charge adding that the impressions of the past with which I shall be equipped on the eve of torture will exactly fit the past of another person now living. And things would be no better if, finally, he adds that something will happen to that other person so that he will wake up tomorrow unable to remember the things he now remembers, and will instead be equipped with impressions of *my* past; and that, far from being tortured, the other person will receive a substantial reward. Fear, surely, would still be the proper reaction: and not because one did not know what was going to happen, but because in one vital respect at least 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.[3]

Williams then writes:

> "If this is right, the whole question seems now to be totally mysterious. For what we have just been through is of course merely one side, differently represented, of the transaction which we considered before... " (Williams 1970, 52–53)

Williams points out that the main difference between the two cases is that the first case is described entirely in third-personal terms, whereas the second case is described from the first-person perspective—as happening to *me* (53). Williams also concludes that while the first case supports the psychological continuity theory, the second case supports the bodily continuity theory, since in the latter, I undergo the procedure and remain in my body. And he writes:

> "It is often recognized that there are 'first-personal' and 'third-personal' aspects of questions about persons, and that there are difficulties about the relations between them. It is also recognized that 'mentalistic' considerations... and considerations of bodily continuity are involved in questions of personal identity... It is tempting to think that the two distinctions run parallel: roughly, that a first-person approach concentrates on mentalistic considerations, while a third-personal approach emphasizes considerations of bodily continuity. *The present discussion is an illustration of exactly the opposite*." (Williams 1970, 62, emphasis added)

That is, we might have thought that a first-person approach to personal identity would support the psychological continuity theory, whereas a third-person approach would support the bodily continuity theory. But, somewhat surprisingly, we find that exactly the opposite is true: the first-person approach supports the bodily continuity theory, whereas the third-person approach supports the psychological continuity theory. Or so Williams argues.

Now I am willing to grant, at least for the sake of argument, that the second of these two claims is true, i.e. that the the third-person approach supports the psychological continuity theory. But I deny the first claim, the claim the first-person approach supports the bodily continuity theory. The reason I deny this emerges when we consider a *third* case, one which Williams doesn't consider:

[3]This follows Williams's text closely, but is not a quotation. I have amended the case to avoid certain irrelevant complications.

> 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 then adds that during the night I will undergo a certain medical procedure. The procedure will leave my psychology intact—I will remember all that he's said to me, and will have my normal memories, feelings, thoughts, dispositions of character, and so forth. This will, of course, do nothing to cheer me up. He also tells me that when I wake up tomorrow, I will no longer be in my present body, nor will I have my present brain, since the procedure will have the result that when I wake up to face my torture tomorrow, I will find myself in a strange and unfamiliar body. Additionally, the person who is currently in the body that I will find myself in tomorrow will awaken in *my* body, and he will be given a reward. None of these facts seem to make anything better. Fear, surely, would still be the proper reaction: and not because one did not know what was going to happen, but because in one vital respect at least one did know what was going to happen—torture, which one can indeed expect to happen to oneself, and to be preceded by certain unwanted medical procedures as well.[4]

While not considered by Williams, this case is a type of thought experiment very familiar from the personal identity literature: imagine you wake up in a different body tomorrow, with all your memories and psychological states intact. Many have used this sort of case to argue for the psychological continuity theory.[5]

Taken together, the second and third cases seem to suggest that the first-person approach supports the simple view, not the bodily continuity theory. For note that if both of those cases are metaphysically possible, then Supervenience fails. An implicit assumption of the our discussion is that the continuity relations between the relevant person stages are the same in both of these cases. Let's call the body I start out in in the second and third cases *the A-body*, and the other relevant body in both cases *the B-body*. Let '*preA*' be an abbreviation for *the person in the A-body before the procedure*, and let '*postA*' be an abbreviation for *the person in the A-body after the procedure*, and define '*preB*' and '*postB*' in an analogous fashion. (Perhaps these bodies have large scarlet letters painted on them.) So in both cases, *preA* (me) and *postA* are related by bodily and brain,

[4] I assume that the medical procedures in all three presentations are exactly the same.
[5] For discussion of this sort of case, see Shoemaker (1963) and Thomson (1997, 217–18).

but not psychological, continuity. But in the second case, I end up as *postA* (I remain in my body) whereas in the third I do not, since I end up as *postB* (I go where my psychology goes). So we have two cases in which *preA* and *postA* are the same with respect to continuity, even though the two cases differ over whether *preA* and *postA* are stages of the same person or not. So if both cases are metaphysically possible, then Supervenience is false.[6]

This, I contend, is the real puzzle raised by Williams's case: when we imagine the case from the third-person point of view, our judgments support Supervenience, and in particular, the psychological continuity theory; but when we imagine it from the first-person point of view, our judgments tell against Supervenience. Note that the majority of commentators on Williams's puzzle seem to have missed what it actually shows. Discussions of Williams's essay have tended to follow Williams in taking it to show that we have intuitions that support the psychological continuity theory, on the one hand, and intuitions that support the bodily continuity theory, on the other. But as our third case shows, Williams's puzzle arises not because we're torn between the psychological and bodily continuity theories, but because we're torn between the complex view (in particular, the psychological continuity theory) and the simple view.

4

I close by considering some objections to the foregoing with the aim of clarifying my view.[7]

Objection 1. "If our judgment about the second case conflicts with our judgment about the third case, why should we accept both of those judgments and conclude that Supervenience is false? Why not accept that one of those judgments is wrong, even if we don't know which one it is?"

Reply. I am not claiming that our judgments about the second and third cases conflict with each other. And I am certainly not advocating that we accept both of pair of conflicting judgments. What the second case supports is a possibility claim: I could undergo the procedure Williams describes and remain in my original body. What the third case supports is another possibility claim: I could undergo the procedure Williams describes and fail to remain in my body. These claims don't conflict with each other; taken together, they conflict with Supervenience.

[6]This argument against Supervenience requires the assumption that I would be a person in both of these cases, and that no two persons are in the same place at the same time. Lewis (1976) denies the latter assumption.

[7]Thanks to an anonymous referee for raising these objections.

The objector also seems to think that I am *concluding* that Supervenience is false. I am not. I am merely making a claim about *which view is supported by* the judgments we are inclined to make when we imagine Williams's thought experiment from the first-person point of view. Williams took those judgments to support the bodily continuity theory; I deny that, and take them to support the simple view (the denial of Supervenience). But it is a further question whether or not those judgments are actually true, and so a further question whether or not the simple view is actually true. Many other considerations would need to be taken into account in order to arrive at a defensible answer to those questions. For example, I have followed Williams in assuming that our judgments about Williams's first case—the one presented from the third-person point of view—support the psychological continuity view, and so support the complex view. Thus, on my account of the matter, the evidence we get from considering these three versions of Williams's thought experiment is still 'mixed' so to speak: one of those judgments supports the complex view, two of them (taken together) support the simple view. How to resolve the tension between those judgments is a matter I leave for future inquiry.

Objection 2. "You are just *assuming* that the description you give in the third case is correct and that it really is *me* who is going to be tortured tomorrow. This identity-assuming description is perhaps in line with how a defender of the simple view would describe that scenario, but you have not ruled out there being something in virtue of which *I* am the one who is going to be tortured. For example, perhaps the case is one in which my soul moves from one body to another. If that were right, then there may be an explanation for my persistence lurking in the background."

Reply. I did assume in my description of the third case that it was *me* who would be tortured—just as Williams assumed in his description of the second case that all those things were going to happen to *me*. The point is that, in each presentation, the identity-assuming description seems, at least at first glance, to be a coherent one, one we can readily follow and make sense of without much difficulty. That is at least some evidence—defeasible evidence, no doubt, but evidence nonetheless—that both cases are indeed possible.

It may be worth mentioning that, at least since Kripke (1980), this is the standard way of establishing *de re* possibility claims. Pick out an object (Nixon, say) and describe a scenario in which that object satisfies some condition ("imagine that Nixon lost the election"). If we have no

trouble conceiving of that scenario, that is at least some evidence that that object really could have satisfied that condition (Nixon could have lost the election). We do not need to describe a scenario in purely qualitative terms in order to use it to support a *de re* possibility claim.

I described the simple view as the view that personal identity is not simply a matter of psychological, bodily, and brain continuity. That leaves it open that personal identity *is* a matter of 'soul continuity'. So as I've characterized it, the simple view is compatible with (though does not entail) the 'soul continuity theory' that the objector is envisioning.

Objection 3. "How is the body switch happening? You are just assuming in the third case that the person switches bodies. But how? Surely that is relevant to whether or not the judgment that the case is possible can be trusted. If you filled in the details, and we were no longer inclined to judge the case possible, then your case would not support the judgment you need it to."

Reply. Williams describes the relevant procedure in a very schematic fashion:

> "...suppose it were possible to extract information from a man's brain and store it in a device while his brain was repaired, or even renewed, the information then being replaced...
>
> ...we can imagine the case we are concerned with in terms of information extracted into such devices from *A*'s and *B*'s brains and replaced in the other brain..." (Williams 1970, 47).

Williams emphasizes that, after the procedure, *postB*'s psychological states causally depend on *preA*'s psychological states, and likewise for *postA* and *preB*.

Williams then claims that when one imagines his thought experiment—which involves imagining undergoing this schematically-described procedure—from the first-person point of view, our judgments support the bodily continuity theory. That is what I object to. I claim that there are (at least) two ways of imagining this case from the first-person point of view, one of these ways supports the claim that I could undergo this procedure and remain in my body, while the other supports the claim that I could undergo this procedure and fail to remain in my body. Taken together, those judgments support the simple view, not the bodily continuity theory.

So it is consistent with what I have claimed so far that, once the relevant procedure is described

in more detail, we would not accept the judgment that I could undergo this procedure and fail to remain in my body. But note that the objector has not actually provided any evidence for thinking that once the relevant procedure is described in more detail, we *will* no longer accept that judgment; rather, they are simply pointing out that this *might* be so. But this is not at all surprising, and the same point applies to pretty much every non-trivial judgment about an interesting hypothetical case in philosophy. It applies, for example, to the judgments that Williams reaches in discussing his two cases as well. Since we lack the power to describe hypothetical thought experiments in maximally specific detail, it is almost always going to be the case that filling in more details in a case *might* lead us to alter our initial judgments about the case. But it is not clear what to make of that fact.

Having said all that, I can also offer a more direct response. Given a broadly functionalist theory of mind, the states that are 'extracted' from the *A*-body and 'inserted' into the *B*-body are functional states, states that can be characterized by their causal relations to perception, other mental states, and action. Given a non-skeptical epistemology of mind, there is some way of coming to know what mental states a given agent is in given full knowledge of the relevant perceptual inputs, brain states, and action outputs. So we may imagine a machine that is given full information concerning the *A*-body person's relevant inputs, states, and outputs and uses this information to produce a complete record of the functionally characterized mental states that the *A*-body person is in at that time.[8] The machine then has the power to take this record and 're-program' the *B*-body brain, with the result that, after the re-programming is complete, the person in the *B*-body is in all of the same functionally characterized mental states that the *A*-body person was in. At the same time, the machine re-programs the *A*-body brain so that it is now in all of the same functionally characterized mental states that the *B*-body person was in.

The question now is whether we could make sense of an expanded version of the third case that includes these details concerning how the machine works. I believe we can:

> 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 then adds that during the night I will undergo a certain medical procedure. As a result of this procedure, when I wake up tomorrow, I will no longer be in my present body, nor will

[8]For an estimate on the computational power needed to produce such a record, see Bostrom (2003).

> I have my present brain, since the procedure will have the result that when I wake up to face my torture tomorrow, I will find myself in a strange and unfamiliar body. The procedure involves a machine that will be given full information concerning my relevant perceptual inputs, brain states, and action outputs. The machine will then produce a complete record of my mental life, and will proceed to re-program the brain of another body so that that brain/body will then be in all of the mental states that I was in previously. As a result of this procedure, I will awake to find myself *in* that body with my psychology wholly intact. I will remember all that he's said to me, and will have my normal memories, feelings, thoughts, dispositions of character, and so forth. This will, of course, do nothing to cheer me up....

It seems that I can make sense of what is being proposed here, even when the details are filled in in this way. I may wonder how the experimenter can be so sure that this is indeed what will happen to me. But that is not really the issue—the issue is whether I can envision the possibility he is proposing. And indeed I can; it seems that what he is proposing is something that I can readily imagine happening to me.

The general point here is one that has been made before in the literature on personal identity over time, though the point has not, to my mind, been sufficiently appreciated. A number of philosophers have pointed out that, when one imagines personal identity thought experiments from the inside, one can imagine undergoing and surviving any number of vicissitudes. For example, as Johnston observes:

> ...we can imagine many sorts of cases that seem to involve one's ceasing to be associated with a particular human body and human personality. These cases are particularly compelling when imagined "from the inside." So I am to imagine undergoing a radical change in my form... and perhaps concurrently a wild change in my psychology. There seems to be nothing internally incoherent about such imaginings. (1987, 70)

In a similar vein, Nagel writes that:

> When I consider my own individual life from inside, it seems that my existence in the future or the past—the existence of the same 'I' as this one—depends on nothing but

> itself... My nature then appears to be at least conceptually independent not only of bodily continuity but of all other subjective mental conditions, such as memory and psychological similarity. It can seem, in this frame of mind, that whether a past or future mental state is mine or not is a fact not analyzable in terms of any relations of continuity, psychological or physical, between that state and my present state. (1986, 33)

Of course, what to make of these observations at the end of the day is a further question, as I have been emphasizing. But if we want to understand the bearing of thought experiments like Williams's on questions of personal identity, these observations about how things look from the first-person point of view are ones that we need to contend with.

References

Bostrom, N. (2003). Are we living in a computer simulation? *The Philosophical Quarterly*, **53**(211), 243–255.

Johnston, M. (1987). Human beings. *The Journal of Philosophy*, **84**(2), 59 – 83.

Kripke, S. (1980). *Naming and Necessity*. Blackwell, Oxford.

Lewis, D. K. (1976). Survival and identity. In A. O. Rorty, editor, *The Identities of Persons*, pages 55–77. University of California Press, Berkeley.

Nagel, T. (1986). *The View from Nowhere*. Oxford University Press, New York.

Parfit, D. (1984). *Reasons and Persons*. Clarendon Press, Oxford.

Rovane, C. (1998). *The Bounds of Agency*. Princeton University Press, Princeton, NJ.

Shoemaker, S. (1963). *Self-Knowledge and Self-Identity*. Cornell University Press, Ithaca.

Sider, T. (2001). Criteria of personal identity and the limits of conceptual analysis. *Philosophical Perspectives*, **15**, 189–209.

Szabó Gendler, T. (1998). Exceptional persons: On the limits of imaginary cases. *Journal of Consciousness Studies*, **5**(5–6), 592–610.

Thomson, J. J. (1997). People and their bodies. In J. Dancy, editor, *Reading Parfit*. Blackwell, Oxford.

Williams, B. (1970). The self and the future. *Philosophical Review*, **79**(2), 161–180. Reprinted in Williams (1973, 46–63). Pages references are to the 1973 reprint.

Williams, B. (1973). *Problems of the Self: Philosophical Papers 1956-1972*. Cambridge University Press, Cambridge.