Psychological Continuity and Personal Identity

Simon Marcus *May 2009*

The philosophy of personal identity seeks to answer the question, in virtue of what can a person at one time be said to be the same person at another time? Since Locke, this question has resulted in a rich body of literature and a host of fascinating thought experiments. I will draw most of the examples and positions from Williams, Parfit, Olson, Nozick, Shoemaker and Noonan, who jointly comprise a representative sample of the current positions on personal identity. This paper will proceed as follows: (1) I will provide a theoretical backdrop to the problem, (2) I will explicate a representative selection of the thought experiments (3) I will examine the philosophical conclusions that we might draw, and (4) will ultimately conclude that psychological continuity, suitably defined following Nozick, is the criterion of personal identity we should endorse.

What is the question of personal identity?

Locke elicits some of those considerations that arise in questions of personhood and personal identity. He argues that "to find wherein personal identity consists, we must consider what *person* stands for; which I think, is a thinking intelligent being, that has reason and reflection, and can consider itself as itself, the same thinking thing, in different times and places."[1] It is clear from this passage that the notions of continuity and persistence are critically important. Kant's first analogy in the *Critique of Pure Reason* sought to demonstrate, inter alia, that "all appearances contain that which persists (*substance*) as the object itself, and that which can change as its mere determination."[2] Kant's point is that through every alteration in time, there must be some persisting object whose properties undergo that alteration – roughly the view we take today. If I were to paint a hitherto white wall red, we would presumably say that the wall is *now* red. We would not say that in repainting it I had destroyed the white wall and created a new red one – there are not really thought to be two distinct objects. It is a question of *essence and constitution* what it would take to destroy or merely alter the wall (or anything else). While we would say that the wall persists through the colour change (it is the same wall, differently coloured), we might be disinclined to say that the wall would persist through my

[1] Locke, J. (1689). *An Essay Concerning Human Understanding.* (2004 ed.). (R. Woolhouse, Ed.) London: Penguin, Ch XXVII (9-10)

[2] Kant, I. (2000). *Critique of Pure Reason.* (P. Guyer, & A. Wood, Eds.) Cambridge: Cambridge University Press, A182, emphasis in the original.

smashing it into the pile of bricks that comprise it. This is because a wall is defined in some sense *essentially* as a stable or unified barrier. If that essence goes, the wall (or whatever) goes too – it would cease to exist. The thought experiments we will examine attempt to derive that essence in virtue of which we can say that the same person persists through time or is destroyed.

There are two related points before tackling the thought experiments. First, it will do to note with Swinburne[3] the levels of inquiry that figure for questions of personal identity. Swinburne distinguishes, usefully, I think, between the metaphysical question of *what it is* for a person at one time to be the same person at a later time, and the epistemic question of *how we might know* that a person at one time is the same person at a later time. [4] It is uncontroversial that we typically *individuate* persons according to a bodily criterion – we count (e.g.) the number of people in a room by counting how many bodies there are in that room. It is an epistemic question of how or whether we could know that there are more or fewer people in the room than there are bodies. However, I take the question of personal identity to be more importantly interrogated at the metaphysical level. That is, we should ask *in what does personal identity consist*, whatever our epistemic limitations.[5]

Finally, it has been argued that intuition tests are not the correct way to arbitrate philosophical issues, some critics going as far as suggesting that they provide frequently misleading results. They have come under attack with special severity in the philosophy of personal identity, largely because the literature on personal identity makes liberal use of science fiction examples,[6] and it could be suggested that our intuitions are not suitably

---

[3] Swinburne, R. (1973 - 1974). Personal Identity. *Proceedings of the Aristotelian Society , 74*, pp231-247

[4] Cf. Swinburne, R. (1973 - 1974), p231

[5] Cf. Noonan, H. (2003). *Personal Identity* (2nd ed.). New York: Routledge, p2

[6] Johnston (2007) recognises this *modus operandi*: "...when it came to philosophical theorizing about personal identity, the popular methodology—''the method of cases''—had been to collect ''intuitions'' about real and imaginary cases of personal survival and ceasing to be, and then bring those intuitions into some sort of reflective equilibrium that bore on the question of the necessary and sufficient conditions for an arbitrary person's survival. Imagined cases were treated as more or less on a par with real cases; for the then natural idea was that we should not restrict our evidence base to the adventitious experiments of step-motherly nature, when we could also

acclimatised to such spacey conditions. Bernard Williams, himself responsible for some of the more outlandish examples, notes in his influential paper *The Self and the Future* that the way in which the stories are framed can direct our intuitions in divergent directions.[7] I acknowledge this issue though will not attempt to resolve it here.[8] It will suffice to agree in spirit with the contention that intuition tests require supplementation from (and ought not to usurp the place of) critical evaluation and rational argument. I follow Parfit, however, in recognising that, even if intuition tests do not (necessarily) tell us what the truth is, they do tell us *what we believe*. And, on presentation of the thought experiments, most often we are not left speechless: on the contrary, "many of us find that we have certain beliefs about what kind of fact personal identity is."[9] With these theoretical tools in hand, let us turn to the intuition tests.

The Intuition Tests

These tests admit of dozens of constructions in the literature on personal identity (in fact, one would be forgiven for concluding that every philosopher has a private yen to be a science-fiction writer). It is not controversial to assert, however, that only a handful of the tests aim at substantively different points. As such, I will make use of only a suitably representative sample of them. In each case, I will present the experiment, describe the intuitions it might elicit, and will provide reasons for each case to endorse the psychological continuity criterion of personal identity.

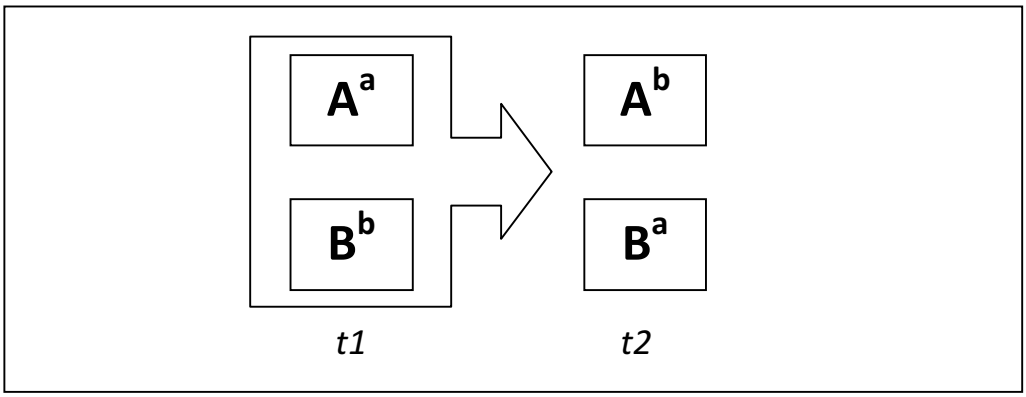
---

avail ourselves of the ingenious thought experiments in the philosophy journals" (p33).

[7] Cf. Williams, B. (1970). The Self and the Future. *The Philosophical Review , 79* (2), p171

[8] For further discussion on thought experiments in the field, see Coleman, S. (2000). Thought Experiments and Personal Identity. *Philosophical Studies , 98*, 53–69 and Gendler, T. S. (2002). Personal Identity and Thought-Experiments. *The Philosophical Quarterly , 52* (206), 34-54 and Johnston, M. (1987). Human Beings. *The Journal of Philosophy , 84*, 59–83.

[9] Parfit, D. (2007). Is personal identity what matters? *The Ammonius Foundation*, p3



Body Exchange. Williams presents an example admittedly difficult to cast in non-question-begging terms.[10] The idea, roughly, is that two people (A and B) are in the hands of an evil scientist. He tells them that he will perform an operation on their minds and bodies. He explains that he will take the mental contents of A (the beliefs, desires, values, memories etc.), and place them in B's body, and vice versa for B and A's body. Where does person A go (if anywhere)?

This question admits of three apparent answers: (1) person A goes where his psychology goes, into body B, or (2) person A goes where his body goes, and so remains in body A, or (3) person A cannot survive the procedure and is destroyed. Williams thinks that our intuitions should point overwhelmingly in favour of (1), that psychological continuity is the sufficient criterion for survival. However, as mentioned, intuitions ought not be the sole arbitrators in the matter, so do we have further reasons to weigh in in favour of psychological continuity in this case? In order to check our intuitions, Williams supposes that after the procedure the scientist will torture one body and give $100000 to the other. Person A is asked before the procedure occurs to choose (on selfish grounds) which body ought to be given the torture and which the money after the procedure is complete. How should he choose? The typical response to this case is to say that person A should have the torture meted out on the A-body-person and the money given to the B-body-person – that is, people generally take 'changing bodies' to be a good description of the outcome of the procedure. As Williams notes, "this seems to show that to care about what happens to me in the future is not necessarily to care about what happens to this body (the one I now have)."[11] Williams thinks that we

[10] Williams, B. (1970), p163

[11] Williams, B. (1970), p164

might elicit further support for (1) by appealing to questions about the choice that A makes. If the B-body-person has *A*'s psychological states including *A*'s memories, then on receipt of the $100000 that person will recall having elected that the B-body-person should receive the money and feel pleased at his choice. He will say something like "*I* made the right choice." In addition, we might suppose that the B-body-person has some physical impairment, like a wooden leg. When *A*'s psychological states are transferred into the B-body, that person might say "Oh, I don't much like this leg – it's much more uncomfortable than my old one. I suppose I'll have to get used to it now that I'm in this body." It is difficult not to beg the question in such speculations, but I do take this to be roughly how the person might respond. That is, when confronted with the wooden leg, I think that the person would say that *it is he* who would have to do the getting used to thereof. Williams' critical (and I think fair) supposition here is that what the individuals *say* is an indication of what their interests are, and thus in what their persons persist; and here it is clear that it is psychological continuity we prize.

Body Exchange(?) part 2. Williams provides a second example, which he thinks drives our intuitions in the opposite direction. In order to see why, I will not paraphrase the example but rather quote him in full:

> Let us now consider 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. This certainly will not cheer me up, since I know perfectly well that I can forget things, and that there is such a thing as indeed being tortured unexpectedly because I had forgotten or been made to forget a prediction of the torture: that will still be a torture which, so long as I do know about the prediction, I look forward to in fear. He then adds that my forgetting the announcement will be only part of a larger process: when the moment of torture comes, I shall not remember any of the things I am now in a position to remember. This does not cheer me up, either, since I can readily conceive of being involved in an accident, for instance, as a result of which I wake up in a completely amnesiac state and also in great pain; that could certainly happen to me, I should not like it to happen to me, nor to know that it was going to happen to me. He now further adds that at the moment of torture I shall not only not remember the things 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 now have. I do not think that this would cheer me up, either. For I can at least conceive the possibility, if not the concrete reality, of going completely mad, and thinking perhaps that I am George IV or somebody; and being told that something like that was going to happen to me would have no tendency to reduce the terror of being told authoritatively that I was going to be tortured, but would merely compound the horror. Nor do I see why I should be put into any better frame of mind by the person in charge adding lastly that the impressions of my past with which I shall be equipped on the eve of torture will exactly fit the past of another person now living, and that indeed I shall acquire these impressions by (for instance) information now in his brain being copied into mine. 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.[12]

Williams takes this new example to leave us in a quandary, since he takes this second case to be a mere retelling of the first. If we supposed that the best answer to the first example was that psychological continuity is sufficient for personal identity (in a way that bodily continuity is not), then he thinks that this example should incline us in precisely the opposite direction. He offers some guidance as to why this might be so, but thinks that he has in no way constructed the example unfairly. He argues that although that the person's captor refers to the captive individual at each stage as 'you' might appear question-begging, he argues that in fact we follow the example quite clearly and his use of that locution seems fitting. Williams notes also that we are not given much information about the man whose memories are to be inserted into 'my' body. I concede that, in some sense, the two examples of 'body exchange' present the *same* state of affairs, and that we cannot sustain conflicting intuitions about the cases. One option available to refute Williams which I will advance is that our intuitions ought to point in both cases in favour of psychological continuity.

---

[12] Williams, B. (1970), pp167-168

The right question to ask is in what sense is the person in A's body *A* after the procedure? It is clear that the person is not *A* psychologically speaking, since *ex hypothesi* A's body has none of A's mental states. But if this is so, in what reasons would we have for thinking that the A-body-person is A, other than that he looks like A? My twin may look just like me, but what tells us apart *in the crucial sense*, is our beliefs, desires, values, memories and projects. *In this sense, the A-body-person is no more similar to A than to anyone else.* The second example turned on the supposition that I should feel fear at the prospect of future torture, irrespective of the mental change which was first to occur. On evaluation, however, the correct response ought to be that I fear torture for that person who is the inhabitant of my body at the time of torture. We might then inquire, who at the time of torture is in fact the inhabitant of A's body. I think we have overwhelming evidence to say (i) that the absence of any of A's mental states rules out that it might be A, and (ii) the presence of B's mental states rules in favour of the proposal that it is *B* who is being tortured. I think that, contrary to Williams' alleged intuition, it is remarkably easy and far more compelling to bite the bullet (as it were) in support of the psychological continuity thesis.

Why the physical criterion is mistaken

Having dealt with some initial hurdles, I wish first to address a few points which direct my thinking about the psychological thesis. Firstly, it is prudent to explain why it is I take bodily continuity to be neither necessary nor sufficient for personal survival.

It is commonplace to suggest that our bodies are significant to our 'sense of self' or 'or identity as persons.' Some eminent psychologists have said, for example, that "the true self is bound up with bodily aliveness,"[13] and "bodily experience, derived from physical sensation, constitutes the first experience of a sense of self."[14] I wish to propose that acceptance of the spirit of these statements need not urge us to endorse the bodily criterion of personal identity at all. It is common cause that I have never had any experiences which weren't in some sense located in my body. Moreover, I am deeply familiar with how it feels to be in my skin, so to speak, and I can only guess at what it might be like to look through someone else's eyes. My body has, to be sure, been pivotal in framing my experience (e.g. if I didn't have asthma, I would not have visited the clinic as on my sixth birthday; if I

---

[13] Winnicott, D. (1965). Ego Distortion in Terms of True and False Self. In *The Maturational Processes and the Facilitating Environment.* London: Hogarth, p147

[14] Bloom, K. (2006). *The Embodied Self.* London: Karnac Books, p6

weren't so tall, I wouldn't have been able to reach the books on my shelf, etc.). However, the bodily continuity theorist has to advance something stronger, the modally more extravagant claim that *to be me I must be in this body*. It is not the same as the materialist supposition (which is more plausible[15]) that in this world I must necessarily be materially embodied;[16] it argues that for me to be me I must necessarily be materially embodied *in this particular body*, and consequently that I couldn't survive without it. I contend that this goes too far.

The thrust of my refutation is derived from experiments involving transplants. When someone, say, acquires a prosthetic hand, we do not think that this person is an entirely new creation, or that to perform such an operation is to destroy the original person. We think that the person persists through that change. In principle we might persist in the replacement of parts, and perform successive transplants, yet the person would persist (and in so doing we would not enter the territory of the paradox of sorites either). To drive this point home, let us imagine a case in the future where a person, Jacob, has sadly been in a terrible accident. He is paralysed utterly, but his mind, housed in his brain, is perfectly functioning. An option exists for Jacob: some neurosurgeons explain that they can transfer his brain (and thus his mental states) into someone else's healthy but brainless body which is similar to Jacob's, and attach the brain to the nervous system in the appropriate way. The person who emerges will have Jacob's brain, and thus all of the attitudes, beliefs, values and memories that Jacob had: in short, that person will have Jacob's *personality*. This result would lead most people to conclude that in some critical sense the person who emerges will be Jacob. That he is not bodily continuous will not matter. We would treat him just as we would treat someone whose body had to be massively reconstructed after an accident, tentatively at first, but with the underlying knowledge that the person inside *really is Jacob*.

Now it is doubtless clear how one will proceed to refute the brain criterion. Suppose that there is some man, Michael, who is otherwise healthy but whose brain develops a strange disease, which will result in death if not attended to. Fortunately, some neurosurgeons explain that they can, using a

---

[15] Cf. Shoemaker (2008): "Contemporary neo-Lockeans have followed Locke in framing their view in a way that is neutral between materialism and dualism and compatible with both, but most of them are in fact materialists. It is entirely compatible with their view, and held by most of them, that the mental states of affairs that they hold to constitute personal identity are physically realized" (p315).

[16] Cf. Parfit, D. (1986). *Reasons and Persons.* Oxford: Oxford University Press, p209

reliable device, extract all of Michael's mental characteristics and transfer them into a healthy brain. This new brain would be then be placed into Michael's body and attached to his nervous system in the appropriate way. After the procedure, that person would think that he is Michael (as he would have all of Michael's memories and projects) and he would behave just like Michael. Furthermore, it is so obvious that this person is Michael that it *would be very difficult and seemingly pointless to convince him or anyone else otherwise.* We endorsed the brain continuity criterion *only insofar as it facilitated psychological continuity*. If another device or distinct brain can do the job just as well as the original brain in that regard we would have no reason to favour it, and thus no reason to think that the continuity of the brain is a necessary criterion for personal identity. This powerful result seems difficult to refute.

A final note: at the level of ordinary discourse, I take it that we are frequently in the business of changing our bodies. We exercise, nip, tuck, splice and graft to amend our bodies all the time. Doubtless after an 'extreme makeover' someone might say 'I feel like a completely new person.' What might she mean? She probably intends something similar to those sentiments of the psychologists mentioned earlier who pointed out correctly that our bodily construction contributes in a significant way to how we perceive the world. What is clear though, is that *it is one and the same person* who 'feels new' after the makeover, because she is suitably *psychologically* continuous with herself prior to going under the knife.

A brief note on memory and externalism about mental content

I have thus far contended that psychological continuity is both a necessary and sufficient criterion for personal identity, in a way that bodily continuity is neither. It will do briefly to mention a few surrounding philosophical aspects in virtue of which I take this to be the correct view to advance. Firstly, regarding mental states, the functionalist model of multiple realisation has received widespread support. The idea is that to be in a particular mental state *M* is to be in some functional state. This specification allows that in principle the same mental state could be realised in a suitably *configured* but differently *embodied* creature. Part of this response is to explain the distinct notions intended by "identical with," "realised by" and "constituted in." In this respect, Lynne Rudder Baker argues for "'the Constitution View, according to which persons are

constituted by bodies without being identical to the bodies that constitute them."[17]

To borrow a well-known sort of example, if a Martian (who happens to look and speak quite like you and me) were to trip and fall, and exclaim and wince in a way consistent with a pain role, I would attribute to him the *mental state* of pain. It would not matter that he does not share human biology, because the sufficient criterion for the mental is the functional state he (it) is in. Furthermore, if the Martian expressed views on, say, global politics, the well-being of animals and the advancement of the rights of previously oppressed people, I would take him to be making genuinely *moral* claims. And, in virtue of this, if I saw him kick a dog later, I would consider him *really* blameworthy. The best way of making sense of this, I contend, is that the Martian is, in the correct sense of the word, a *person*. Eric Olson[18] has advanced the claim that we have radically misconceived the notion of personal identity because we have paid far too little attention to that most salient fact, that *humans are animals*. What I take to be the correct mode of rebuttal of Olson's position is to grant him his claim insofar as the relation between humans and animals goes, but to deny that there is a *necessary* relation between humans and *persons*. Though more could be said here, it is my contention in brief that the multiple realisation thesis (within a materialist framework) is apt to defuse Olson's animalistic claims in this regard.[19]

Finally it will do to note why memory has been such a successful criterion for personal identity. It is argued that memory captures most of what we want from our mental life, as it is specially tied to what it is like to be a particular person, and to her point of view. This tight tethering suggests that it would provide just the sort of criterion for identity we are after. Recent considerations regarding the externalism of mental content seem to provide even further reason for supposing this to be true.

Externalism about mental content argues (roughly) that the content of a person's mental state depends in some way on his relationship to the environment, and the causal ancestry of that mental state. Suppose my body

---

[17] Baker, L. R. (1999). What am I? *Philosophy and Phenomenological Research , 59* (1), 154

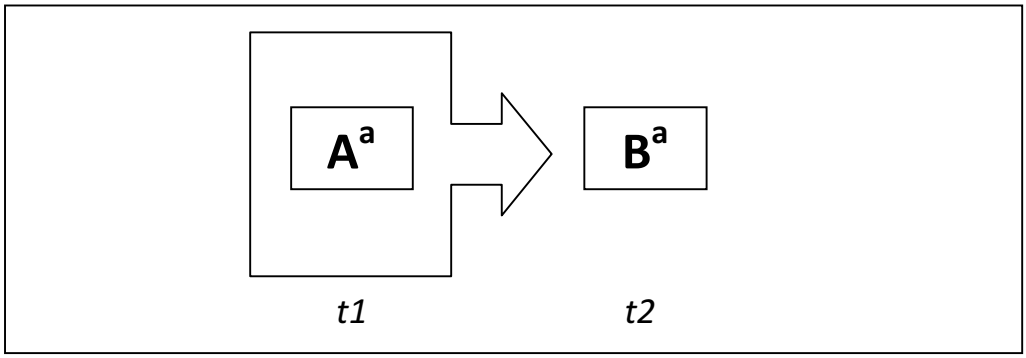
[18] Olson, E. (1995). Human People or Human Animals? *Philosophical Studies , 80*, 159-181.

[19] Shoemaker (2008), in "Persons, animals, and identity" appears to take a similar approach to defuse Olson's claim.

and brain are about to die, and so I elect to have my mental states scanned and transferred into a new, healthy body that looks like mine. Consequently, that body doesn't have thoughts or memories 'of his own' when he is created. Rather, an externalist analysis of his memories reveals that they are *my* memories. Later when that person affirms to my wife that 'I love you,' that person will be evincing that *I* love her. This gives some weight to those intuitions that we have in such a case, that the resulting person really would be me: he would be me because his mental life is causally tethered to mine in a stable way. (Whether we endorse externalism about mental content or not will not make the psychological thesis of identity stand or fall: I mention it here because it captures the causal component to mental states which secures a tight form of mental continuity, but we could retain this causal component without necessarily endorsing the other claims of externalism.)

Revising Psychological Continuity

I have mentioned that even if we are persuaded by the psychological continuity thesis, there are yet some points which bear upon how we articulate the position. One of the more pressing thought experiments, that of fission, seems to some to throw a metaphysical spanner in the works. I will introduce the experiment via it's seemingly more benign cousin, Teletransportation.



Teletransportation. This now-familiar sort of case calls to mind Star Trek episodes and 'beaming up.' Parfit[20] tells the story with some literary flair, and gets the reader to imagine that a time when the process of Teletransportation is commonplace. A person who undertakes the process

[20] Cf. Parfit, D. (1986), p199

would know that one steps into a cubicle and presses a button, which makes a machine capture your 'blueprint' – exhaustive information about your mental and physical constitution. A moment after pressing that button, you are now on Mars (or wherever else you want to go); at least (not to beg the question) the person who is now on Mars thinks that he is *you* – the same *you* who pressed the button a moment earlier. Why ought we to think that it is you? Well, the machine reliably and instantly recreates a molecule-for-molecule replica of you, and implants your entire mental history. Simultaneously, the body that stepped into the machine on earth is destroyed. The mind that emerges on Mars will think, 'This is *my* body – and, darn, even the scar from my bike accident last year is still here. But what a nifty way to *travel!*'

A similar case raised elsewhere by Parfit is that of resurrection.[21] Suppose that I die, and my body perishes. God, however, sees fit later to resurrect me (His ways are mysterious indeed). He thus creates a perfect living replica of me as I was at the time of my death (and by 'perfect,' I mean that the replica and I would be qualitatively exactly alike in mental and bodily constitution). Parfit asks how we should assess this resurrection. He acknowledges that the replica will not be *unambiguously* me. That is, insofar as the replica is not bodily continuous with me, there is a prima facie concern that it might not be *me* (my friend Thomas regards me with a dubious aspect, say). What is established, however, is that there is psychological continuity between me and my replica. It is similar to the case of Teletransportation or transplant in that at time *t1* there is a person with mental states *a* in body *A*, and at *t2* mental states *a* are newly located in a numerically distinct body *B*. On this thought experiment, as above, we ought to reject the bodily criterion of identity in favour of the psychological one, because we suppose that the discontinuity of the body is irrelevant to my continued identity through resurrection (or Teletransportation). It would be *me* that is resurrected (or Teletransported).

Insofar as this is the case, and we endorse psychological continuity as the correct criterion of personal identity, we will say that my replica really is me. Now, some have objected that cases like Teletransportation or resurrection just cannot culminate in *identity*. Where the bodies are discontinuous and thus numerically distinct (even if they are perfectly alike) we are not presented with an *identical* person. Parfit's point in *Reasons and Persons* is that this lack of strict identity should not concern us,

---

[21] Parfit, D. (1971). On "The Importance of Self-Identity". *The Journal of Philosophy , 68* (20), p689

provided we have serviceable survivors related to ourselves in the correct way. Parfit explains what this 'correct way' might be through his contention that our interest in identity is in some sense a *derivative* interest: we desire identity only because of something that we take identity to secure, namely the psychological continuity constituted by the persistence of our memories, values, and projects in the future. Parfit's approach is to grant the incorrigible sceptic what he wants, and concede that the person after undergoing Teletransportation[22] is in a sense not identical with person who pressed the button moments before. Yet Parfit maintains that such 'Parfitian' survival really is *as good as* personal identity, and that we should have no reason to favour the one over the other (other than a propensity for sceptical incorrigibility). I take this point to be generally satisfying, because it reveals what we value in our personal persistence, and shows identity to be nothing more than one usual way of preserving those features.

Our incorrigible sceptic (having missed the point *again*) might protest that it is significant *somehow* that the psychological continuity is not secured through its normal cause. Should this concern us? Parfit concludes, and I agree with him, that it matters not a jot. He asks whether someone with artificial lungs can be said to be breathing[23] or artificial eyes can be said to be seeing.[24] In such cases, we wish to say that, insofar as the artificial eyes or lungs are reliably performing the same function as their natural counterparts, then that person really is seeing or breathing, again revealing our merely derivative interest in these features. If the sceptic insists that breathing requires 'the normal cause,' I would suggest that he relies on an arbitrarily narrow definition of breathing. However, even granting the person his dogmatic definition, I would contend that such non-usual 'breathing' is certainly as good as normal breathing. This would lead me to concur with Parfit that even if psychological continuity does not have its normal cause (and thus does not provide personal identity on a narrow conception), what it provides is *as good as* personal identity.[25]

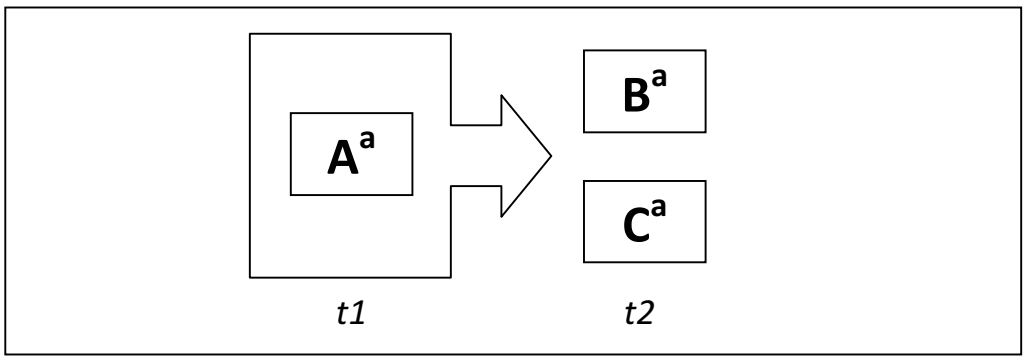
---

[22] Or resurrection, or whatever example matches the criterion of bodily discontinuity while retaining psychological continuity (ceteris paribus)

[23] Cf. Parfit, D. (1971) , p690

[24] Cf. Parfit, D. (1986), p209

[25] Cf. Parfit, D. (1986), p209



Division (Fission). Now that we have established Parfit's plausible take on what matters in personal identity, let's turn to a more radical example of fission to see if the intuition holds up. In the previous case we imagined a situation in which, like a brain transplant, personal survival was preserved because my mental states remained intact through the transfer into a new body. However, there is reason to suggest (in fact) that the two hemispheres in one's brain in some sense duplicate each other's work. This is revealed in cases of stroke sufferers, where one hemisphere of the brain loses all functionality, yet the person retains all of his mental states because of the persisting functionality of the remaining half. Let's suppose that a person, Alfred, has a body which will soon collapse under the strain of an awful disease, even though his brain is perfectly functional. One option would be to place his brain in someone else's body, and thus ensure Alfred's survival. We would agree that he would survive even if only one of Alfred's two brain hemispheres was still functional and sustained Alfred's mental states. Suppose that Alfred is one of three identical triplets. Now let's suppose that Alfred's whole brain is functional, and that one half is transplanted into the (brainless) body of brother B and the other half into the (brainless) body of brother C.

What happens to Alfred? Does he survive as both of these people, as one of them, or as neither of them? We committed originally to the supposition that if one hemisphere were transferred into just one other body, that the person would really be Alfred. We have supposed that this happened twice over, but now appear reluctant to say that they *are both* Alfred. We are in the throes of an antinomy – "how could a double success be a failure?"[26] We should note also that it seems prima facie implausible that Alfred survives as only one of them (although Swinburne seems to argue for this), since B

[26] Parfit, D. (1971a). Personal Identity. *The Philosophical Review , 80* (1), p5

and C are alike in their claim to being Alfred in having his mental states, and have them through the same causal mechanism. The psychological continuity thesis, as I have thus far characterised it, is committed then to the claim that they *both* survive as Alfred or that Alfred does not survive.

The major objection to the psychological continuity account at this point derives its strength from the transitive nature of identity. If we agree that B=A and C=A, we must transitively affirm that B=C. This, however, strikes some as an untenable result. In the case of fission, they might say, what could be clearer than that B and C are numerically distinct, qualitatively distinct, and that we have two people rather than one?

Initial Intuitions: A = B = C

Regarding the fission into two new bodies, I admit to having no clear intuitions other than these: (i) that I would, if there were *only one* recipient, say that identity is preserved, (ii) that the two resulting individuals have an equal stake in the identity in question because they came about through the same causal process, (iii) that each would feel from the inside as though he were the original person, although in a slightly different body, and (iv) that two numerically distinct objects cannot be one and the same object.

At the level of *intuition,* I don't feel in a position to say who the *real* Alfred is, and if presented with the two survivor-candidates I expect they would not either. They presumably would think it a thoroughly perplexing state of affairs! After the shouting of "it's me!," "No, it's me!," "Well then what's Alfred's favourite colour, eh?," and "Who does Alfred secretly want to marry?" have been resolved in a predictable tie, they presumably would reiterate the point supposed earlier: that a double success *really isn't* a failure. Thus I allow as *at least plausible* that they are both Alfred. But what I take to be most important is that, however we resolve this issue, it does not make the physical criterion any more plausible. Furthermore, it is difficult to see how the fission example would really *weaken* the psychological account, if properly evaluated. My point is this: that there is a conceivable state of affairs like division (for which we have *very few* clear intuitions) should not dissuade us from that criterion which ably captures everything we desire from those situations where our intuitions are clearest.[27] This last point is an epistemic one, not a metaphysical one. I

[27] I take it to be correct that we should not be dissuaded from a philosophically satisfactory doctrine on presentation of a thought experiment with minimal intuitive or instructive content. As Noonan (2003) points out, it is "often said, the concepts

endorse the claim that if the division of Alfred were to occur, there would be a *de re* state of affairs, where it would be really determinate who (if anyone) survived as Alfred.[28] My concern goes to what it is we should take the thought experiments to show, and my conclusion is that the division example is in some sense *epistemically* dissatisfying – it does not elicit from us a response, a way of telling, which would give us a clue as to what we want from a metaphysical thesis of personal identity.[29]

It is pre-theoretically plausible and in some sense intuitive that both A = B and A = C.[30] How then shall we treat the logical transitivity entailment, namely that B = C? The objection is that we violate a principle of identity (including the relations of necessary identity between rigid designators and Leibniz's law) to assert that B = C because B and C are numerically distinct. A few responses are available. One option is to give up the language of identity following Parfit. We could insist that psychological continuity is nevertheless the correct criterion for what matters in personal identity: when there is only one survivor, psychological continuity constitutes identity, and when there is more than one survivor what it provides is *as good as* identity.[31] More on this later. One other option exists which appears

---

we have serve our practical needs in the situations in which we find ourselves, but there is no reason to suppose that they must have a determinate application in every bizarre situation that philosophic ingenuity can conceive" (p104).

[28] Cf. Salmon, N. (1982). *Reference and Essence.* Oxford: Blackwell, pp 244ff

[29] Cf. Garrett (2002): "[Evan's and Kripke's] conclusion may help mitigate Bernard Williams' observation that it is hard to know how I should react, upon hearing of the unfortunate fate of some future person, where I know that it is vague whether I am that person. As he points out, we have no model for such expectation... However, this situation may be rendered less problematic once we appreciate that the vagueness is linguistic and not ontological" (p81).

[30] As Wright (2006) notes: "I have discovered that, when the fission problem is explained to persons who are *not* professional philosophers, such as undergraduates, the most common response is to assert, or even insist, that B is the same person as A *and* that C is the same person as A. It is only *after* it is pointed out that since B and C are numerically distinct, and that this leads to a contradiction, do they start to retract from this position. But the position that appears to be *pre-theoretically attractive* is that A=B and that A=C and that, in consequence, B=C" (p133).

[31] Parfit, D. (1971a): "Psychological continuity is a ground for speaking of identity when it is one-one... I have suggested that if psychological continuity took a branching form, we ought to speak in a new way, regarding what we describe as having the same significance as identity" (pp12-14).

prima facie intriguing and potentially satisfying: it is to suggest that despite appearances B and C are *numerically identical*.

Wright[32] and Ehring[33] expound this claim of *numerical* identity through examples involving time travel. Suppose some years from now time travel is actually possible,[34] and I have a desire to have a conversation with myself when I was ten years old. I step into a time machine, and find myself face-to-face with my earlier self. Of course, many features are different – I didn't have a beard at that age, or the scar from the bike accident when I turned thirteen. But these needn't be incompatible with identity: we would say that Simon-2020 has a beard but Simon-1995 does not, thus not culminating in contradiction. It would be just like saying that Simon-2020 remembers what he had for breakfast on his 30th birthday, while obviously the ten year old Simon-1995 could not 'remember' (or foresee) any such thing, even though they are numerically identical. So are there two people or one person in the room? We profit from the host of science fiction films involving time travel in judging that there is in fact *one* person. Regarding fission, Wright and Ehring conclude that though B and C might appear distinct, they bear the same relation as I would to my former self in the case of time travel: the bodily disjunction is compatible with a relation of strong identity, that is, B and C are numerically identical. Furthermore, the claim thus constructed fares well against Leibniz' account of the indiscernibility of identicals: both B and C have identical (relevant) psychological properties as a result of a *causally* stable process.

When Alfred divides, as it were, into two distinct bodies, I mentioned that I take it as *plausible* that they are both him. In the absence of strong intuitions, I take this to be one serviceable view, and one which is in line with our earlier considerations which affirmed the psychological thesis of identity. However, it will do to point out that I do not think this 'divided identity' must hold indefinitely; on the contrary, one of the few intuitions I have about the matter suggests that 'divided identity' is a plausible view of the situation only for the briefest time after the division has taken place. Brian Garrett points out that though they are perfectly similar at the time of fission, "B and C possess distinct loci of mental life and occupy different

---

[32] Wright, J. (2006). Personal Identity, Fission and Time Travel. *Philosophia , 34*, 129–142.

[33] Ehring, D. (1987). Personal Identity and Time Travel. *Philosophical Studies , 52*, 427-433.

[34] It is a live question as to whether time travel is conceptually coherent – can an event precede its cause? But let us set that aside for the moment.

bodies at different, perhaps causally unconnected, spatial locations."[35] They will very soon continue their lives (Alfred's life, to an extent) in some clear sense as distinct individuals, largely because of their physical distinctness. They will, say, enjoy rights which accrue to them *as individuals*, and should one break his arm, the other would not feel it, and there would doubtless be some difficulty in mediating those personal relationships which Alfred had, as those relations now need to take account of two distinct bodies even if there are not two completely distinct minds. *The psychological causal ties which grounded their initial identity as one person will swiftly erode.*

I contend then that numerical identity has something of the ring of truth, but seems plausible only at the moment of fission. This was because the two bodies, and thus minds, will soon take on quite distinct lives, which are *causally* disconnected. The consequence, it is clear, is that B and C will quickly lose their relation of identity as the causal bonds are loosed. It would seem that the account of identity needs to do better by the idea that if C had not existed, B would be really *identical* with A. How best can we make sense of the relations between A, B and C while preserving the psychological account of identity?

Intrinsic and Extrinsic Conditions

Properly to evaluate this point requires a consideration which has informed various theses of personal identity, that of the intrinsic or extrinsic criteria of identity. According to the Intrinsic Criterion (elsewhere called "the only a and b"[36] or "only x and y"[37] principle), "whether a later individual *y* is identical with an earlier individual *x* can depend only on facts about *x* and *y* and the relationships between them: it cannot depend upon facts about any individuals other than *x* or *y*."[38] This principle suggests that the criterion of identity we endorse between A and B must not be affected by any further person C.

On the main line case of brain transplant we were happy to adopt psychological continuity as the correct criterion for personal identity: the B-body-person was really A because he was psychologically continuous with A. However, if the process were duplicated (fission), some critics are inclined to say that A is not identical with either B or C because identity is a

---

[35] Garrett, B. (1990). Personal Identity and Extrinsicness. *Philosophical Studies , 59*, p182

[36] Cf. Wiggins, D. (1980). *Sameness and Substance.* Oxford: Blackwell.

[37] Cf. Noonan, H. (1985). The Only x and y principle. *Analysis , 45*, 79–83.

[38] Noonan, H. (2003), p127

one-one relation, and the presence of a competitor means that identity cannot obtain. This rests upon *extrinsic* criteria of identity, namely that identity in some way depends not merely on the intrinsic relations between A and B, but on the existence of any further competitors for identity with A. Where Wright and Ehring championed the thesis that A, B, and C are numerically identical, their result was a consequence of following through on the Intrinsic Criterion: relations of identity depend only on intrinsic features of the relationship, so on pains of consistency they affirmed transitively that B = C. However, many take this conclusion not merely to be counterintuitive but conceptually unstable, and favour the Extrinsic Criterion as the way successfully to avoid such pitfalls.

Parfit recently argues that we should endorse the extrinsic model of identity. His suggestion is that this extrinsic view lends clear support to his thesis that identity is not what matters for survival. His thesis has some compelling features, and runs as follows. In the single case of my brain transplant, the resulting person would be me, and my relation to myself in the future would here contain what prudentially matters. That *this relation* contains what matters depends only on its intrinsic features. My relation to myself tomorrow, in the *Single Case*, is intrinsically the same as my relation, in *My Division*, to each of the two resulting people. Therefore my relation to each of these people must contain what matters. However, it is not true that each of these people would be identical to me. Therefore personal identity would not, here, be what matters.[39]

Parfit's point is that in the single case my identity is preserved *and* the relation contains what matters, thus people are inclined to conflate the two, or to misinterpret the derivative relations between them. This is clarified by the division case, as it retains what matters but simply cannot be identity. This is intuitively accurate: it would be wholly irrational for me to bribe the nurse to destroy the left portion of my brain to secure strict identity after the procedure, since *what matters is preserved anyhow*. Thus identity cannot be what matters for me. His argument concludes that the extrinsic criterion for identity is incompatible with the suggestion that we have a non-derivative interest in identity.

This argument takes for granted that we endorse the Extrinsic Criterion of identity, but it is worth evaluating why or whether we should. The major argument for the Extrinsic Criterion rests on the intuition that A = B when there is a single survivor, and A ≠ B when there is more than one survivor.

---

[39] Parfit, D. (2007), pp46-47

As Garrett puts it, in the case of division "the best explanation of why A is not B is simply that there is another, equally good, candidate for identity with A. Consequently, the most plausible theory of personal identity over time must incorporate a non-branching or no-competitors clause, and is therefore a best-candidate theory."[40] Garrett here introduces a claim at the heart of Parfit's and Nozick's accounts, that the Extrinsic Criterion should compel us to adopt a 'best candidate' approach to personal identity. Furthermore, the notion of candidacy we endorse should provide an explication of the attitudes I have towards that person (those people) who is (are) to be my future survivor(s).

The Close, Closer, and Closest Continuer Thesis

Parfit explains that the question to ask is "what is it rational to care about, in our concern about our own future?"[41] Here Parfit's and Nozick's accounts reach somewhat divergent conclusions. I wish not to arbitrate their dispute, since it does not go to my purposes here, and in any event seems to turn often on misunderstandings of each other's claims.[42] However, I will suggest that Nozick's account seems prima facie more satisfactory and reveals a helpful way of drawing the psychological account into epistemically clearer territory.

Nozick's suggestion is that we have a special concern which does not waver for that person who is our *closest* continuer. The intuition is drawn out through considering how we might feel *about ourselves* in the future, if there were no threat of branching, transplant or duplication. It is clear that we have a special care for our future selves, and frame this sentiment in a way which is distinct from my caring about some other person. Although I have a kind of terror at the thought of my friend's being hurt in the future, it is a distinct kind of terror from knowing that I am later to be hurt. Part of this involves the feelings which I (my closest continuer) would experience *first-hand*. My continuer would be causally connected to me in a strong way, namely through sharing my mental set. He has my hopes, memories, projects, desires and can achieve my ends as well as I can. As Nozick puts it,

---

[40] Garrett, B. (1990), p182

[41] Parfit, D. (1986), p282

[42] Nozick, for example, probably does not consider Parfit's claims regarding the combined spectrum in his account. Noonan (2003), on the other hand, agrees with me that they seem at cross-purposes, though he ultimately disagrees with them both. He writes, "this argument against Parfit is, as I said, Nozick's. But in presenting it Nozick makes a mistake which the statement just given avoids, and causes Parfit, in his discussion of Nozick, to miss the point" (p167).

"the closest continuer relation is the best instantiated realisation of the relation of identity... and we care about it as identity."[43]

It is clear from this conception that Nozick agrees that identity in some sense is not what matters to us – Like Parfit, we care about identity only in a derivative manner – and that the closest continuer will provide us with what we care about for our future qua *our* future. Briefly, where Nozick and Parfit diverge is regarding how to theorise the relations between the care I have for a continuer, the number of competing continuers, and the closeness of the continuer to a stage of my life. The relation of closeness I bear to my continuer is intrinsically defined – it depends on no one other than me and my continuer. The extrinsic component of the thesis, noted best by Nozick, is that how much I care about my continuer is more than a function of the closeness of the continuer to me, it is also a function of whether that continuer is *closest*.

An example which neatly captures this point[44] is that of Parfit's simple Teletransportation and branching instances. In the simple case, I have a relation *R* of closeness to my continuer on Mars, which is a function merely of my relationship with him. Since he is my *closest* continuer in this instance, I am him. However, in the branching case, my perfect replica is created on Mars but my body (and mind) are not destroyed on Earth: the machine malfunctioned and I will live on earth for just a week until my heart collapses from the strain of the procedure. That is, there will be a week for which my life and that of my replica on Mars will overlap. In such an instance, I will be just as *closely connected* to my replica on Mars as in the simple case, since closeness is intrinsically determined. However, the *care* I have for the replica that week is quite different in the two cases, because in the branching case he is not my *closest* continuer – at least not yet. I care specially about my closest continuer in a way that is disproportional to the other less-close continuers. It is thus argued (successfully, I think) that the most suitable account of psychological continuity must satisfactorily emphasise the 'best candidate' nature of what we care about.

---

[43] Nozick, R. (1981). *Philosophical Explanations.* Oxford: Clarendon Press, p67

[44] Here I am indebted to Noonan (2003) for neatly spelling out this point – Cf. pp165-167.

Closing Up: Nozick's Account Evaluated and Endorsed

I have presented several reasons why we ought to favour the psychological account as providing us with the best way of capturing what we want in personal identity. I take it that Nozick's thesis, combined with my contentions regarding the multiple realisation of the mind and the causal nature of mental content, give us the answers we intuitively desire from the thought experiments. In closing I will spell out the way in which our considerations successfully and satisfyingly arbitrate that trickiest contender, the fission debate.

Suppose that we have the case above involving Alfred's brain divided into the cavities of his triplet siblings, B and C. Here there are two competing candidates for the claim to identity. As we have seen, the most plausible way of evaluating this state of affairs is to consider the matter using the Extrinsic Criterion – the 'best candidate' approach. We note then that in the absence of competitor B (or C), C (or B) would really be Alfred. The presence of the competitor and the rules of identity forbid that the relations between these candidates could be identity after fission has occurred. However, the Extrinsic Criterion entails that our desire for identity is in some sense derived from our desire for psychological continuity. In this case the appeal of this point is clearest – it is this psychological continuity that *matters*, and it is this that is preserved. We will have to judge that in some sense they are not both Alfred – they couldn't be. However, Nozick's (and Parfit's) highly intuitive point is that we would be thoroughly mistaken in supposing that Alfred is as good as dead: rather we should conclude that Alfred is *as good as alive*.

The model is that of the best candidate, but here we have a tie. As Nozick claims, what we care about is who the *closest* continuer is, but who is that here? In this dispute, there is some sense in which (i) there is no closest continuer since there is not one who is closer than the other, and another sense in which (ii) they are both the closest since each continuer can claim that no one is closer than he is. Nozick gives effect to our intuition that a double success just isn't a failure: "what I care about is that there remains something that continues me closely enough to be me if it were my sole continuer."[45] This point, which satisfies all we desire for our future selves, should incline us straightforwardly to adopt the second horn of the dispute: they *both* are closest. Our care for them is equal, and it is the care we have for ourselves. Nozick's point can be summed up neatly: in the case of a tie, it is preposterous to say that there is no winner.

---

[45] Nozick, R. (1981), p67

Bibliography

Baker, L. R. (1999). What am I? *Philosophy and Phenomenological Research , 59* (1), 151-159.

Bloom, K. (2006). *The Embodied Self.* London: Karnac Books.

Coleman, S. (2000). Thought Experiments and Personal Identity. *Philosophical Studies , 98*, 53–69.

Ehring, D. (1987). Personal Identity and Time Travel. *Philosophical Studies , 52*, 427-433.

Garrett, B. (1990). Personal Identity and Extrinsicness. *Philosophical Studies , 59*, 177-194.

Garrett, B. (2002). *Personal Identity and Self-Consciousness.* New York: Routledge.

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

Johnston, M. (2007). "Human Beings" Revisited: My Body is Not an Animal. In D. W. Zimmerman (Ed.), *Oxford Studies in Metaphysics* (Vol. 3). Oxford: Oxford University Press.

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

Kant, I. (2000). *Critique of Pure Reason.* (P. Guyer, & A. Wood, Eds.) Cambridge: Cambridge University Press.

Locke, J. (1689). *An Essay Concerning Human Understanding.* (2004 ed.). (R. Woolhouse, Ed.) London: Penguin.

Noonan, H. (2003). *Personal Identity* (2nd ed.). New York: Routledge.

Noonan, H. (1985). The Only x and y principle. *Analysis , 45*, 79–83.

Nozick, R. (1981). *Philosophical Explanations.* Oxford: Clarendon Press.

Olson, E. (1995). Human People or Human Animals? *Philosophical Studies , 80*, 159-181.

Parfit, D. (2007). Is personal identity what matters? *The Ammonius Foundation* .

Parfit, D. (1971). On "The Importance of Self-Identity". *The Journal of Philosophy , 68* (20), 683-690.

Parfit, D. (1971a). Personal Identity. *The Philosophical Review , 80* (1), 3-27.

Parfit, D. (1986). *Reasons and Persons.* Oxford: Oxford University Press.

Salmon, N. (1982). *Reference and Essence.* Oxford: Blackwell.

Shoemaker, S. (2008). Persons, animals, and identity. *Synthese , 162*, 313–324.

Swinburne, R. (1973 - 1974). Personal Identity. *Proceedings of the Aristotelian Society , 74*, 231-247.

Wiggins, D. (1980). *Sameness and Substance.* Oxford: Blackwell.

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

Winnicott, D. (1965). Ego Distortion in Terms of True and False Self. In *The Maturational Processes and the Facilitating Environment.* London: Hogarth.

Wright, J. (2006). Personal Identity, Fission and Time Travel. *Philosophia , 34*, 129–142.