The Memory Criterion and the Problem of Backward Causation

Locke famously wrote “And as far as this consciousness can be extended backwards to any past action or thought, so far reaches the identity of that person, it is the same self now with this present one that now reflects on it, that this action was done.”[1] This and similar passages have been interpreted as providing a memory criterion for personal identity. Lockeans, as well as their critics, have pointed out that the memory criterion is likely to mean that none of us were ever fetuses or even infants due to the lack of direct psychological connections between then and now. But what has been overlooked is that the memory criterion leads to either backward causation and a violation of Locke’s own very plausible principle that we can have only one origin, or backward causation and a number of overlapping people where we thought there was just one. I will argue that such problems cannot be avoided by replacing direct psychological connections with overlapping chains of connectedness – what has been called “psychological continuity.”[2] The most famous account of psychological continuity, that of Derek Parfit, will still fall prey to these problems for he understands psychological continuity to consist of overlapping chains of strong psychological connectedness, the latter defined as involving “at least half the number of direct connections that hold, over every day, in the lives of nearly every actual person.”[3] Moreover, even if these problems can be avoided by some revamped account of psychological continuity, it will not do justice to what is Locke’s insight - recognized by David Lewis as well as Parfit - about the importance to our identity of our consciousness being directly extended into the past.

* * *

Assume you have memories extending back to your early childhood. Then through either a natural process of forgetting (or a minor stroke or a blow to your head), you lose your earliest memory of something that happened to you. Let’s say that this memory was of an

experience of an event at T_1 (1937). Your earliest memory is now of a later time T_2 (1938). That means you are not identical to a being that existed in 1937– at least according to the unreconstructed Lockean memory criterion. Locke wrote: "For whatever any substance has thought or done which I cannot recollect and by my consciousness make my own thought and action, it will no longer belong to me.[4] If the earliest experience you can recall is *now* 1938, and you are not identical to any person that existed earlier, then that actually means you have changed your origins! You have come into existence at a later time than was true before. Thus an event in the present, a memory loss, *causes* your first moment of existing in the past to change. Even if that is not incoherent, it sounds like a very unwelcome sort of backward causation.

Someone might protest that the alleged backward causation is as benign as the arrival of the Second World War making the First World War become just that – the first of the world wars. But the case stated above seems to be more than an acquisition of an unproblematic relational property. It isn't that something which existed acquired another relational property (as in the case of the first of two world wars), but that something which presently exists obtained a new and different origin. The more appropriate comparison is World War I ceasing to be the first world war because of a later event. Imagine the date of the beginning of the First World War *changing* from one time to another because of later events. So what has happened in the case of your memory loss - according to Locke's memory criterion - is that you have ceased to be as old as you were for your first moment of existence on this planet has been changed by an event long after the time of your origins. You had existed at T_1 (1937), but that is no longer true.[5] You now existed no earlier than 1938. Not only does Locke's memory

criterion turn out to imply that you have two origins but it violates his own principle that:

> "When therefore we demand, whether any thing be the same or no, it refers always to something that existed such a time and in such a place, which 'twas certain, at that instant, was the same with its self and no other: From whence it follows, *that one thing cannot have two beginnings of Existence*, nor two things one beginning, it being impossible for two things of the same kind, to be or exist in the same instant, in the very same place; or one and the same thing in different places. That therefore that had one beginning is the same thing, and that which had a different beginning in time and place from that, is not the same but divers."[6]

An alternative to claiming that someone can come into existence *twice* is to instead describe the memory loss as the introduction of a new person.[7] As a result, there is now a person existing from 2004 to 1938. But that means the other person who originated in 1937 has ceased to exist when the memory of the 1937 experience was lost. So the memory loss would result in a new person coming into existence while another person going out of existence. That is quite bizarre. And there still seems to be a backward causation in that a contemporary mental event *now* determines exactly what moment in the past was a person's origins. It seems obvious that something should not happen to a brain in 2004 which results in someone coming into existence years earlier in 1938. That is not the harmless sense of a world war becoming the First World War when there occurrs a second.

It might be thought the backward causation problem can be eliminated just as Reid's transitivity puzzle was by adopting psychological continuity rather than direct psychological connections as the criterion for personal identity.[8] A number of philosophers have sought to patch up various problems in Locke by appealing to psychological continuity. All that is supposedly needed are overlapping chains of memory: at T_N (now) one can recall T_2 (1938) and at T_2 one can recall T_1 (1937) even though at T_N one can't recall the events of T_1. Overlapping chains of memory (or intentions, desires etc.) would seem to imply that there would be no loss of a person, no new origins, and no present event changing your first moment on the planet.

But it isn't clear that such a move is in the spirit of Locke for it lacks the intuitive appeal that one goes back in time as far as one's consciousness extends. Mayra Shectman makes this point well:

> "Certainly a view that places identity in the ancestral relation of psychological connection rather than in direct connection does not have Reid's *transitivity* problem, but it is also not clear that it captures the relation we take to underlie the importance of personal identity. Locke's observation is, roughly speaking, that it is my direct conscious access to experience makes it *mine*. This is not, however, the relation in terms of which psychological continuity theorists define identity. With Reid's objection in mind, these theorists place identity in a weaker relation that does not demand direct conscious access to the actions and experiences that are ours – the ancestral relation of direct access. It is not obvious, however, that this weaker relation can rightfully claim to have all the intuitive appeal as the bearer of identity that the original relation had. In fact

> psychological continuity theorists make it clear that they attach much more importance to direct connections than to the weaker relation of continuity."[9]

The importance of direct psychological connections rather than the overlapping chains of psychological continuity is evidenced in the claims of modern day neo-Lockeans like Parfit and Lewis. They stress psychological connectedness more than continuity. Parfit writes "of these two general relations, connectedness is more important (than continuity) in both theory and practice."[10] Lewis makes a similar point in his account of Methuselah. He writes that

> "We sometimes say: in later life I will be a different person. For us short-lived creatures, such remarks are an extravagance. A philosophical study of personal identity can ignore them. For Methuselah, however, the fading-out of personal identity looms large as a fact of life. It is incumbent on us to make it literally true that he will be a different person after one and one-half centuries or so." [11]

Leaving aside for the moment that appeals to psychological continuity seem to be missing something important about identity across time, it is worth noting that the revised Lockean account Parfit offers can handle Reid's objection but not the backward causation problem. Partfit understands psychological continuity to consist of "the holding of overlapping chains of strong connectedness."[12] And strong psychological connectedness between any two days, involves at least half the number of direct connections that hold, over every day, in the lives of nearly every actual person.[13] So if a blow to the head today leaves you with slightly less than half the normal psychological connections between today and yesterday, then there is not enough connections to establish psychological continuity between you today and any

person yesterday, thus your origins have changed and you did not exist in the last century or even the previous week. Your origins were much more recent.

One can avoid this problem by claiming that *any* degree of memory connections is sufficient for continuity but only if one is willing to accept that one doesn't survive a stroke that is of sufficient severity that one is left a permanent amnesiac regarding any pre-stroke aspects of one's biography. If one does survive such a stroke, then one has new origins given that psychological continuity doesn't extend back to a time before the stroke. However, one might try to claim that the psychological continuity account allows that we survive such stroke-induced amnesia in a way that preserves our earlier origins because psychological continuity persists through other psychological states than autobiographical memory. For instance, one may have the memory of how to speak English, do long division and read a map. There also may be continuity of impersonal beliefs (the world is round) and generic desires (for food and shelter). But these seem to have little to do with your identity, i.e., what distinguishes you from any other adult. And notice that adopting such an approach means not only that we would have moved away from the original Lockean memory criterion, but we would be working with even a more watered down version of psychological continuity than before.

If a psychological identity criterion must involve some appeal to psychological connectedness as Locke, Lewis, Parfit and Schectman imply, the threat of backward causation can be avoided only by adopting a modal rigidity and a rather embarrassing overpopulation. If one believes there are a lot of overlapping persons as does Lewis, then a blow to the head that eliminates one person whose earliest memory was of 1937 doesn't introduce a new person.[14] The second person already existed connected from 2004 to T_2 (1938). But not only does this mean accepting that counterintuitive explosion of embedded people but it will necessitate it

being impossible for someone, rather than their counterpart, to have lived any differently.[15] If the actual person hit on the head could have avoided the blow, then the normal forgetfulness of aging or some later trauma to the head would likely mean a change in the earliest psychological connection and thus the return of the problems of backward causation and someone having two origins.

Notes

[1] John Locke, *An Essay Concerning Human Understanding.* Ed. Peter Nidditch (Oxford: Oxford Univ. Press, 1975), p. 335. Similar ideas are expressed at pp. 340-341, 345.

[2] Derek Parfit, *Reasons and Persons*. (Oxford: Oxford Univ. Press, 1984), pp. 205-06.

[3] Ibid., p. 206. For an account of psychological continuity without insisting on strong connectedness, see Jeff McMahan's account of "broad psychological continuity" in his *The Ethics of Killing: Problems at the Margins of Life* (Oxford: Oxford Univ. Press, 2002), p. 50.

[4] Locke, p. 345.

[5] Or if you don't like the sound of a proposition that was true no longer being so, we could say that it is still true that earlier in 2004 you originated in 1937, but it is also true that that sometime later in 2004 you originated in 1938.

[6] Locke, p. 328. Italics added to the quote.

[7] This denial of coming into existence twice is not meant to exclude the possibility of intermittent existence, only that someone can have two first moments of existence.

[8] Reid famously imagined an old general who remembers his first battle as a young soldier but doesn't recall being flogged as a small child after stealing an apple from an orchard. Yet at the time of his first military campaign, he could recollect being flogged. This would mean that the

general was identical to the soldier of the early campaign, that soldier was identical to the boy, but the general was not the same person as the boy. This violates the transitivity of the identity relation.

[9] Marya Schectman, *The Constitution of Selves* (Ithaca: Cornell Univ. Press, 1996), pp. 28-29.

[10] Parfit, p. 206.

[11] David Lewis, "Survival and Identity," in *Philosophical Papers. Vol. 1* (Oxford: Oxford Univ. Press, 1983), p. 66.

[12] Parfit, p. 206.

[13] Ibid.

[14] David Lewis, "Many, but Almost One," in *Papers in Metaphysics and Epistemology*. (Cambrdige: Cambridge Univ. Press, 1999), pp. 164-82.

[15] Some four dimensionalists, like Ted Sider in his *Four Dimensionalism: An Ontology of Persistence and Time* (Oxford: Oxford Univ. 2002), have tried to argue that four dimentionsalism need not commit its adherents to counterpart theory, pace the argument in Peter van Inwagen's "Four-dimensional Objects," *Nous* 24 (1990), pp. 245-255.