Williams on expectations and the self

PHIL 20208
Jeff Speaks

October 10, 2006

1 Exchanging bodies

Williams begins by describing a case which we are, intuitively, inclined to think of as a case of two persons exchanging bodies. This is similar to Locke's example of prince and the cobbler:

> "For should the soul of a Prince, carrying with it the consciousness of the Princes past life, enter and inform the body of a Cobler as soon as deserted by his own soul, every one sees, he would be the same Person with the Prince, accountable only for the Prince's actions . . . "

Williams' example does not involve the transfer of an immaterial soul from one body to another; rather, he imagines that

> "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, in the information then being replaced: it would seem exaggerated to insist that the resultant man could not possibly have the memories he had before he had the operation."

So we are to imagine two people, A and B, who have had their personalities swapped by a method somewhat like this; each have had the information from their brains uploaded to a device, and then downloaded into the body of the other. Suppose further that "after the experiment, persons familiar with A and B are just *overwhelmingly struck* by the B-ish character of the doings associated with what was previously A's body, and conversely." We can then ask: is the person with the A-body after the experiment A, or B?

As Williams notes, it can seem very plausible that in this sort of case, we genuinely have two persons exchanging bodies. He dramatizes this by imagining that an announcement is made before the experiment that one of the two persons after the experiment will be given $100,000, and one of them will be tortured. Suppose that before the experiment A says that the person in B's body should get the reward, and that B says that the person in A's body should get the reward. Now suppose after the experiment that the experimenter gives money to the A-body-person. In this case the A-body person will say that he's glad that the experimenter went with his wishes, etc., and the B-body person will complain that the experimenter disregarded his wishes. As Williams says,

> "These facts make a strong case for saying that the experimenter has brought it about that B did in the outcome get what he wanted and A did not. It is therefore a strong case for saying that the B-body-person really is A, and the A-body-person really is B; and therefore for saying that the process of the experiment really is that of changing bodies. ...
>
> This seems to how 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) ..."

In other words, these cases strongly suggest that, if you were to undergo this experiment, you would be rational to choose that the person in the body other than yours post-experiment will receive the cash reward — after all, that person will be you.

Why this result is required by, and fits nicely with, psychological theories of personal identity. Why it seems to refute 'body' theories and puts some pressure on 'immaterial soul' theories.

2 Expectations and the self

Williams runs through several different versions of the above thought experiment, all of which lead to the same conclusion: the experiment is genuinely a case of two persons switching bodies. But on p. 167 he turns to a different sort of case:

> "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 ... that will still be a torture that, so long as I do know about the prediction, I look forward to in fear."

Williams then imagines the following, apparently slight, modifications to the case:

1. I will lose not only my memory of the prediction of the torture, but also all of my other memories. As Williams says, "This does not cheer me up, either, since I can readily conceive of being involved in an accident ... 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 ..."

2. I will not only lose my current memories, but also have a set of other memories, different than the ones I now have: "I can at least conceive the possibility ... 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 authoritatively told that I was going to be tortured, but would merely compound the horror ..."

3. The new set of memories I will have will exactly fit the memories of some other person.

4. I will acquire this new set of memories by having the information in that person's brain copied into mine. As Williams says, "Fear, surely, would still be the proper reaction, and not because one did not know what was going to happen, but because ... one did know what was going to happen: torture ... to be preceded by certain mental derangements as well."

Williams' description of this case, and the proper expectations to have if one were in it, seems perfectly correct: none of steps (1)-(4) should, as he puts it, "cheer me up." But this seems to conflict with our intuitions above about prince/cobbler cases of exchanging bodies.

One difference between the two cases: the second case seems to materially involve just one person, rather than two. But why should this matter?

Williams suggests that this case should make us doubt our verdict about the first class of cases; if presented with the experiment described at the outset of the article,

> "we should perhaps decide that if we were the person A then, if we were to decide selfishly, we should pass the pain to the B-body-person."

Why this verdict poses a problem for psychological theories of personal identity.

Can we solve the problem in a way reminiscent of Hume by saying that the issue about which one of the people will be me is a merely verbal, or conventional issue? What does Williams think of this?