Back to the self and the future

Simon Beck

Philosophy Department, University of Natal, Private Bag X01, Scottsville 3209, South Africa

Received and accepted July 1997

The thought-experiment presented by Bernard Williams in 'The self and the future' continues to draw the attention of writers in the debate about personal identity. While few of them agree on what implications it has for the debate, almost all agree that those implications are significant ones. Some have even claimed that it has consequences not only for personal identity, but also concerning the viability of thought-experiment as a method. This paper surveys what these consequences might be at both levels – as a substantive contribution to the debate on identity, and as to what it shows about the usefulness of thought-experiments. It argues ultimately that thought-experiments like Williams's do provide a useful philosophical tool as long as we temper our expectations of them, and that it offers some support to a view of personal identity but one which is at odds with Williams's own view.

Die aandag van skrywers in die debat oor die persoonlike identiteit word voordurend gevestig op die denkeksperiment wat deur Bernard Williams in 'The self and the future' voorgestel is. Hoewel min skrywers oor die implikasies wat dit vir die debat inhou saamstem, stem amper almal ooreen dat hierdie implikasies van belang is. Party het selfs beweer dat dit gevolge inhou, nie net vir die persoonlike identiteit nie, maar ook vir die lewensvaarbaarheid van denkeksperimente as 'n metode. Hierdie werk ondersoek die gevolge op albei vlakke – as 'n substantiewe bydrae tot die debat oor identiteit, en ook wat dit vir ons wys oor die nuttigheid van denkeksperimente. Daar word per slot van rekening beweer dat deneksperimente soos die van Williams 'n nuttige instrument vir filosofie is sodra ons ons verwagtings daaroor matig, en dat dit ondersteuning vir 'n mening oor persoonlike identiteit gee wat oneens is met Williams se eie mening.

These bizarre fictions have their uses in abstract studies, as aids to a better grasp of the nature of our ideas (Leibniz 1765:314).

1. Introduction

A striking feature of the literature on problems about the identity of persons is the predominance of thought-experiments in arguments for or against the competing theories. Although thought-experiments are by no means confined to this debate, in no other area do they occur with such marked frequency. But while they are used by proponents of almost every view in the context they do have their detractors as well, and their status is thus not as secure as it once seemed.[1]

Although it is now almost 30 years old, one of the most influential thought-experiments in the literature is that which occurs in Bernard Williams's 'The self and the future'. I wish to use this case as a starting point for investigating the role of thought-experiments in the personal identity debate. I will set out Williams's thought-experiment and then discuss some of the conclusions which have been drawn from it. My hope is that something

important will emerge from this central example about what thought-experiments can do and can be expected to do. This aim has two parts to it, one methodological and one substantive: my interest is in both what the experiment shows about the method of thought-experiment and what it shows about personal identity itself.

2. Williams's thought-experiment

The thought-experiment is set out as two distinct scenarios, but its force turns on its being revealed to be one and the same scenario differently described. The first description is a fairly straightforward account of a thought-experiment of the kind used in arguments for a criterion of identity in terms of psychological continuity.

A machine has been created which is able to extract and record all of the information stored in one's brain which is relevant to and determines one's mental life. Two individuals A and B are subjected to this process, and the information from A's brain is then fed into B's brain and vice-versa. After the process, the person in B's body seems to remember having A's experiences, has A's beliefs, desires, projects, emotional attachments, and so on. Likewise, the person in A's body has the psychological features previously associated with B. The obvious intuitive response to this scenario is that a body-swap has occurred: the person in the A-body is now B, while A occupies the B-body. This intuitive response strongly supports the view that our concept of personal identity turns on psychological continuity, and that physical continuity is not necessary for identity.

Williams re-inforces this conclusion by considering the likely responses of the individuals involved. As Williams sets the case up, A and B are told about the process that is to be carried out on them and are told that one of the emerging persons will be tortured and the other rewarded. They must each make a choice beforehand as to which is to receive which treatment, presuming that this will be done on purely self-interested grounds. The support for the psychological criterion comes from the judgements the two persons would make as to how wise their earlier choices were once the operation is complete and the torture and reward handed out. For instance, had A chosen that the A-body person be tortured and the B-body person rewarded, the B-body person would seem to remember making this choice and, if this is indeed what happens, he would be satisfied that 'his' choice was the wise one (Williams 1970: 48–49).

Further reinforcement comes from considering other choices:

> Suppose that A chooses that the A-body-person should get the money, and the B-body-person get the pain, and B chooses conversely.... The experimenter announces, before the experiment, that the A-body-person will in fact get the money, and the B-body-person will get the pain. So A at this stage gets what he wants (the announced outcome matches his expressed preference). After the experiment, the distribution is carried out as announced. Both the A-body-person and the B-body-person will have to agree that what is happening is in accordance with the preference that A originally expressed. The B- body-person will naturally express this acknowledgement (since he has A's memories) by saying that this is the distribution he chose; he will recall, among other things, the experimenter announcing this outcome, his approving it as what he chose, and so forth. However, he (the B-body-person) certainly does not like what is now happening to him....The A-body-person will on the other hand recall choosing an outcome other than this one, but will reckon it good luck that the experimenter did not do what he recalls choosing (Williams 1970: 49–50).

Once again, the psychological criterion is supported. The choices A and B made were those which one would expect from adherents of the physical criterion and, as Williams comments, 'in this case the original choices of both A and B were unwise' (1970: 50).

The second description is set out in a different manner. We are asked to consider a series of cases, each a development on the previous one, and challenged to state at which step some crucially relevant change – a difference which could amount to a change of identity – occurs. The first case is one in which A is operated upon in such a way that he loses all of his memories. Williams suggests that this change will not be sufficient to support a judgement that A has lost his identity. The most important consideration here, according to Williams, is that if A were told that after the operation his body will be tortured, A would still have reason to fear that torture despite the intervening memory-loss. These are the six steps which develop on this one:

(i) A is subjected to an operation which produces total amnesia;
(ii) amnesia is produced in A, and other interferences lead to certain changes in this character;
(iii) changes in his character are produced, and at the same time certain illusory memory beliefs are produced in him; these are of a quite fictitious kind and do not fit the life of any actual person;
(iv) the same as (iii), except that both the character traits and memories are designed to be appropriate to another actual person, B;
(v) the same as in (iv) except that the result is produced by putting the information into A from the brain of B, by a method which leaves B the same as he was before; and
(vi) the same happens to A as in (v) but B is not left the same, since a similar operation is conducted in the reverse direction (Willams 1970: 55–56)

At stage (i) we have no difficulty in agreeing that A survives the loss of memory; that is, the person who emerges from the operation is identical with A. Williams suggests that there is no relevant difference between stage (i) and stage (ii) which would justify a judgement that the person emerging in (ii) is not A. Furthermore, this holds for all the following stages as well. The crux comes at the final stage. Since no line can be drawn between any of the stages, we are obliged to say that the person who emerges here is A; but the problem is that the scenario set out in (vi) is precisely that described in the first experiment. In stage (vi), just as in the first experiment, the psychological features of A are transposed into B's body and vice-versa – only there our intuitions told us that the emerging person was B.

The result is that one and the same thought-experiment merely described in different terms evokes directly conflicting intuitive responses. The one response appears to support a psychological criterion of identity, the other a physical criterion, yet these two criteria are mutually exclusive.

According to Williams's exposition, then, we are faced with a conundrum. But precisely what the effects of this conundrum are is a question which needs careful investigation; certainly there have been many divergent responses expressed in the literature. One

reaction has been to argue that the conundrum is only apparent, and that in fact there is some fault with one or the other, or both, of Williams's scenarios. In this way the puzzle would be defused, and need establish nothing radically new about identity or thought-experiments. Williams's final response, although tentative, is along these lines; he suggests that one should 'take the risk' of accepting that identity goes with bodies (1970: 63).

Another response has been that Williams's thought-experiments both succeed, but that the consequent position we should adopt on personal identity is not Williams's tentative affirmation of a bodily criterion but a currently unpopular nonreductionist view.

Perhaps the most important reaction is a rather different one which sees the consequences of the conundrum as being primarily methodological, the argument being that the stalemate to which the thought-experiments bring us shows that the method of thought-experiment is fatally flawed as a way to reaching answers about our identity. I will take a look at all three arguments and formulate my own response to the problem, both as regards methodology and personal identity itself.

3. Noonan's argument that both thought-experiments are flawed

Can we just dismiss Williams's conundrum? Harold Noonan argues (Noonan 1989:Ch 10) that we can ignore any apparent consequences, methodological or other, of 'The self and the future', because both the thought-experiments presented there suffer from drastic internal problems. From the way they are described we can infer that they do not support the conclusions suggested, and certainly do not lead to the conundrum that Williams sees. This argument is not one against thought-experiments in general; it is intended specifically against those outlined above.

Noonan argues against both of Williams's experiments: that the first does not support a psychological criterion of identity, nor the second a physical one. I do not wish to enter this debate at this stage, although my final response to Williams will show that I am sympathetic with one aspect of Noonan's argument. I will argue, more immediately, that Noonan's arguments against both legs are flawed, and will then go on to look at possible consequences of the success of the experiments.

As regards the apparent body-swap outlined earlier, Noonan finds that it fails to support a psychological criterion because of the way in which Williams plays down the crucial role of certain of the participants' (A's and B's) psychological attitudes.

The problem becomes clear, according to Noonan, when one looks at the various stages Williams follows in support of his conclusion. Williams considers A's and B's reflections on the wisdom of their earlier choices made according to (i) a psychological criterion (Williams 1970: 48–49), (ii) a physical criterion (49–50), and (iii) different criteria (50) – where A chooses according to a psychological, and B a physical, criterion. In each case, as outlined above, Williams finds that the individuals' reflections would support psychological continuity as the criterion.

Noonan argues that what Williams must do in order to conclude thus is to ignore, or at least drastically play down, his own assumptions. For how the individuals rate earlier choices will depend crucially on their beliefs as to identity criteria. Take case (ii) for instance. If A and B choose as they do because they accept a physical criterion (i.e. A that the B-body be tortured, B choosing that fate for the A-body) and the B-body person is subsequently tortured, it is not at all clear that this person (B-body) will acknowledge that his original choice, the one he seems to remember making, was mistaken. Rather, suggests Noonan, he will dismiss this apparent memory as illusory – for both A and B believe in the physical criterion – and will complain that the torture was not meted out as he, *that is B*, chose. Williams, in suggesting what he does, apparently ignores the force that the fundamental beliefs which shape the participants' choices will have when they come to reflect on those choices. Since he does this consistently throughout the discussion of the first experiment, there is no support to be gleaned there for the psychological criterion.

It is because of this supposed misrepresentation of how those involved would react that Noonan takes Williams's first description of the thought-experiment to fail. However, it is not at all clear that it is Williams and not Noonan himself who is guilty of misrepresentation. Williams's suggestions as to how A and B will react after the experiment are based on the intuition that individuals will take themselves to be the persons they feel like, and whose lives they remember. This seems to me to be fairly uncontroversial, even when it comes to an individual who believes strongly in the physical criterion of identity. For it certainly does not follow from the fact that A overtly adopts the physical criterion that if A were to look in the mirror and see B's body, he would say, 'Oh look, I'm not the person I think I am!' Indeed, it is most implausible that A would react in this way, yet that is precisely what Noonan is suggesting. As a result, Noonan's case against the first experiment is anything but convincing.

Noonan also argues that the second experiment fails to establish its conclusion. This is because he claims that a line *can* be drawn between two of the steps Williams outlines, namely between (iv) and (v). At this point, according to Noonan, one who holds that psychological continuity is necessary for identity can insist that the person who emerges in (v) need not be identical to that in (iv) without contradicting intuitively correct claims about survival through amnesia, and so on.

The difference is that in (v) the apparent memories of the A-body person have been brought about by a 'very special causal process which ensures that the brain of A has been wiped clean of all the information it contained and that the A-body person is psychologically continuous with B' (Noonan 1989: 225), while in (iv) all that occurs is that illusory memories modelled on B's are induced in A. The difference is not just, as Williams claims, that in (v) the model for the A-body's apparent memories is also their cause. The crux of Noonan's point is that the process whereby A's psychology is altered is in no relevant way different from removing A's brain and replacing it, as in a thought-experiment like Derek Parfit's 'My Division' (Parfit 1984: 254–255), with one hemisphere of B's brain; in that case there seems to be a clear difference between what occurs and inducing illusory memories.

Williams objects to drawing the line between (iv) and (v) on the grounds that the A-body person emerging in (v) must be A, since the existence of an undisputed B prevents his being B. Noonan suggests that this is tendentious because the B-body persons's identity with B *can* be disputed by pointing out that what happens to B is better described as fission than as the production of a copy without claim to identity. The only claim that the B-body has over the A-body to B's identity is the continued occupation of B's body, in particular, continued possession of B's brain; but to use this alone is to beg the question.

Noonan's argument is a strong one against Williams's description of the second thought-experiment. Nevertheless, it is not enough to show that we do not face a conundrum. This is because the example can be altered so as to avoid Noonan's criticism, and yet still present what is at heart the same argument. The way to change the case and achieve this is to increase the number of steps in the description. Instead of the six cases which Williams outlines, one can require consideration of a *spectrum* of cases like that described by Parfit (1984: 231–233) in which each stage represents a change of a very small degree over the stage preceding it, the changes occurring a few cells at a time. Thus we would be faced by a spectrum consisting of an indefinitely large number of cases. In each case A will be operated upon to produce some psychological change, the amount of change will increase gradually as we move along the spectrum, but the difference between any two adjacent cases would be almost imperceptible. In this way, the final case in the spectrum can be that described in Williams's stage (vi), but nowhere along the spectrum does it make sense to draw a line between two stages. After all, how could a change so small represent a change in something so momentous as identity? The resulting thought-experiment presents the same sort of case which Williams intended with his second thought-experiment, but which is proof against Noonan's attack.

The sorts of consideration which Noonan presents do not show any fatally damaging flaws in the thought-experiments under discussion. Given that Williams's case still stands to be informative, the question remains as to what it does show, and especially whether it has any methodological significance.

4. Does the conundrum support the unanalysability of identity?

One influential line of argument is that which takes the thought-experiments to succeed, but to succeed in showing something other than the tentative conclusions Williams draws from them. Indeed, the strategy is taken as showing that the most popular contemporary views on personal identity are quite mistaken. The argument occurs in discussions by Richard Swinburne (1984) and Geoffrey Madell (1981), and runs along the following lines.

Madell and Swinburne deny that personal identity can be analysed into some more familiar or better understood relation like physical or psychological continuity, and they argue that this view gains strong support from Williams's conundrum. Personal identity to Madell and Swinburne is a basic or simple relation which resists any breaking down into something more fundamental. To be the same person is to be the same person, and there is no more to be said. One is not the same person as one was *because of* the holding of some relation other than identity. One immediate implication of this view is that it would be possible for A and B to be distinct persons even though there is no difference between them whatsoever – no difference, that is, other than their being (unanalysably) non-identical. In Madell's terms, 'I might not have existed, but someone having *exactly* the life that I have had might have existed instead' (Madell 1981: 79).[2] On the other hand, you could have had a life totally different to the one you have led, and yet still be *you*.

These entailed possibilities are important for Madell and Swinburne's arguments in the context of 'The self and the future' because they show the acceptance by those who adopt a nonreductionist view like theirs of the possibility of 'bare identity' – of identities or non-identities which do not hold *in virtue of* anything. Madell and Swinburne suggest that

it is precisely an unwillingness to accept the possibility of bare identity which leads Williams and others who believe personal identity to be analysable into trouble with the examples at hand. If one throws out this prejudice, and accepts that physical and psychological continuity may be no more than *evidence for* identity, the threat posed by the two examples is removed.

Following this argument, we would have to admit that the criteria we usually use to make identity judgements let us down in Williams's cases by yielding conflicting results, but that would be just one of the epistemological problems we must be prepared to face; our criteria let us down elsewhere as well. This failure of normal criteria would not be held to mean that there is any metaphysical problem about identity at stake in 'The self and the future'. In the unlikely event that the operation should occur, then one of the emerging people would be A and one would be B; we may just not know which is which.

Madell places even greater stress than Swinburne on these particular examples:

> ... anyone who follows the two stories that Williams tells in 'The self and the future' must also reject the view that one or other story is incoherent. The outstanding fact about these stories is that *both* of them are so compelling. We are led to understand just how it is possible for there to be these two different possibilities (i.e. distinct individuals in different worlds), and for there to be no observable difference between them. Far from this being 'utterly mysterious', as Williams claims, it is precisely the conclusion that our whole argument demands (Madell 1981: 99).

It might make Madell's claim clearer to put it in terms closer to the examples under discussion than does this quoted passage. What the first version shows is how it is possible for an individual to retain her identity despite a total physical change. The second version shows that identity can be retained despite a total psychological change. That the two are the same situation differently described is neither here nor there – together they show how identity can be unaffected by a total psychological change *and* a total physical change. But this is incompatible with the view that identity can be analysed in terms of some sort of observable continuity, and is in line with Madell's prediction that any attempt at such analysis will fail. As a result, it is claimed that a 'simple' view of identity is shown to be correct. Personal identity is a simple, unanalysable relation, something over and above mental and physical continuities.

The case presented by Madell and Swinburne is a neat one, and has some appeal even if its conclusions are unpopular. As a result, their interpretation of the consequences of Williams's conundrum stands, at least conditionally. The condition is this: that Williams's two scenarios *should be* as intuitively compelling as Madell suggests they are. If they are indeed both successful and contain no internal flaws, then they do provide a strong case for the existence of bare possibility with regard to identity. In Sections 6 and 7, the question of whether this condition is satisfied will be raised.

5. Does the conundrum show thought-experiment to be a misguided method?

The potentially most devastating interpretation of the two cases and their consequences is still to be considered. This would take Williams's conundrum to show that the entire method of using thought-experiments to support theories of personal identity is mistaken.

If Williams's arguments go through, then what emerges is that we can have directly conflicting intuitions regarding the application of a concept in a given situation. Such a conflict might be taken as indicating that our commonsense conception of what it is to be the same person over time is incoherent. Alternatively, one might respond by rejecting appeals to intuitions about counterfactual situations, in this context at least. In that case the conflict which has become apparent is taken to show that our intuitions are unreliable as support for a theory of personal identity.

Whatever option is taken, the consequences are the same: we would have to stop using thought-experiments to support or reject alternative theories in this area. This is the position of Stephen White (White, 1989). Noonan suggests that this is also the conclusion which Mark Johnston draws from 'The self and the future' (Johnston, 1987), but I will point out later that Johnston's response is far less radical. White's response is an extreme one, with repercussions extending far beyond the area of philosophy concerned with personal identity. That may not be a reason for rejecting a response like White's to Williams's conundrum, but there are such reasons to be found.

A first important point is that the rejection of thought-experiments is only a consequence of Williams's conundrum if a very strong assumption is made: that what thought-experiments attempt to do is to elicit from us philosophically correct intuitions. For it is only if the intuitive response to each experiment is meant to be *the* correct response that the conflicting responses produce a contradiction. It is true that this assumption *is* made in much of the discussion in the literature which makes use of thought-experiments. When (for example) Locke presents us with his case in which the souls (and thus memories) of a prince and a cobbler are exchanged, he believes that the intuition which 'everyone sees' – that the prince and cobbler swap bodies – is the correct one (Locke 1694: 44). It is thus that he takes the experiment to show that sameness of body is not a necessary condition for identity. But I do not believe that the assumption is crucial to thought-experiments providing a useful and important method of argument.

There is then a methodological moral that I wish to draw from 'The self and the future': that we must cease to view thought-experiments as revealing *the* correct intuition about personal identity and its necessary and sufficient conditions. It still remains to be shown how they can serve any useful purpose once this view is abandoned.

Let us take another look at the first experiment (or a purged version thereof) and note closely what goes on there, especially the cognitive responses it invokes in us. Another thought-experiment, used in a different context, may be a helpful guide here. The experiment is one of Hilary Putnam's and concerns, among other things, the meaning of 'cat'. Putnam sets out the following scenario:

> Suppose ... that there never have been cats, i.e. non-fake cats. Suppose evolution has produced many things that come close to the cat but that it never actually produced the cat, and that the cat as we know it is and always was an artifact. Every movement of a cat, every twitch of a muscle, every meow, every flicker of an eyelid is thought out by a man in a control center on Mars and is then executed by the cat's body as the result of signals that emanate not from the cat's 'brain' but from a highly minituarized radio receiver located, let us say, in the cat's pineal gland (Putnam 1963: 53–54).

He then asks whether, should this be discovered to be the case, there are or ever were any cats. Does one respond that there never were any cats since being an animal is essential to being a cat, or does one respond that there are cats – and thus that cats do not have to be animals? His response is that there are and were cats, despite it now being the case that cats are not animals. However, while Putnam is clear about his own answer which is what seems to be the general response, he suggests that things are not quite as clear cut as this makes them seem (1963: 54). Peter Unger brings out more clearly the complexity of our cognitive response in his discussion of the example:

> Notice that we do *not* make *just one* response to the question asked. Even while our *dominant* response is to believe that the correct answer is 'Yes', we make a *dominated* response, of believing oppositely, that the correct answer is 'No'. Or, at the very least, we have a felt tendency to believe in that negative direction (Unger 1982: 119).

Here Unger seems to be correct. We respond that these familiar objects *are* cats, and yet we also want to say that cats are animals; in that way, we experience the 'felt tendency' against our dominant response of which Unger speaks. In the same way we have a felt tendency to deny our dominant response to talk of 'demon possession' by the witch-doctor of old. Our dominant response is to say that 'demon possession' occurs when a demon enters a human body, but since that never happens there is no such thing as demon possession. Our tendency to deny this (i.e. our 'dominated response') shows in our wish to say something like, 'demon possession is just epilepsy', which implies that there is such a thing.

In the first of Williams's thought-experiments something very similar can be observed. Our dominant response to the question, 'who emerges in the A-body after the operation?' is 'B'. But, as in the case of Putnam's experiment and the case of demon possession, we have a dominated response to the question which is 'A'. In the case of the supposed body-swap the first response is far stronger, that is, our attitude is less equivocal than towards Putnam's experiment, but it is no less misleading to talk as if we have one single (and supposedly correct) intuitive response.

Unger explains responses to Putnam's case in terms of certain of our beliefs and their relative strengths: our 'existence belief' that there are cats is stronger than our 'property belief' that cats are animals. In the demon case, we might say that our property belief is stronger than our existence belief. A similar sort of explanation may be appropriate in the personal identity case, this time in terms of the relative strength of our belief that psychological continuity is crucial to our identity over the belief that bodily continuity is. However, talk of this sort of belief is a bit uncomfortable in the context of commonsense responses, for it is far from obvious that people other than philosophers have such beliefs.

Perhaps the best way of avoiding this worry is to describe the position of the kind of thought-experiment at stake as follows. In mastering a language, we adopt certain classificatory habits – we become disposed to classify certain individuals as 'cats', 'persons', 'the same person as ...', and so on. These habits rest on implicit criteria according to which the classification is done; what the thought-experiments do is to make these implicit criteria explicit. They serve to bring out the principles at work, even if these principles cannot properly be viewed as fully-fledged beliefs. In some cases, like Williams's,

implicit criteria which are usually co-satisfied can be made to come apart, with the possible result that we discover to which underlying principles we are more strongly attached.

To describe the results of thought-experiments as the discovery of the relative strength of commitment to underlying principles which they reveal serves to avoid unfounded faith in the existence of 'the correct intuition'. But it does not mean that thought-experiments have all their promise removed and become pointless. After all, these principles are constitutive of our conceptual system and we can still devise and use experiments which will decide which of two crucial and conflicting criteria or principles is the more fundamental, in virtue of being the one we are least prepared to give up or deny.[3] Williams's example, then, can be used to show that the relation of psychological continuity is more fundamental in our conceptual system than its counterpart concerning bodily continuity. Of course, Williams's other experiment still then poses a problem, even if the problem is now slightly changed.

Even if we no longer assume that the thought-experiment will lead us directly to metaphysical truth – that is, to one clear, coherent and correct intuition, or to some implicit principle which is a strict and universal rule, 'The self and the future' still poses a problem. For if, as suggested, the first thought-experiment reveals that our adherence to psychological features is more fundamental than to bodily ones, the second experiment contradicts this by apparently reversing the direction of dominance. As I remarked earlier, the fact that both experiments seem so compelling needs to be explained.

This problem is important, not only for Williams's thought-experiments, but for thought-experiments in general. Should we find that thought-experiments reveal some minor conflict or tension in the implicit principles which underlie our concepts, this would not necessarily be of any fatal consequence to the method. But the existence of a case which shows a generally felt, direct inversion of intuitions like that under discussion would be a serious threat to the usefulness and reliability of the method, since it raises the suspicion that should any given thought-experiment be differently described we might feel an intuition totally opposed to the one we currently feel. In the next few sections I will be concerned to avert this threat to the method in general by averting the threat Williams's examples pose to the coherence of the concept of personal identity in particular.

6. What Williams's experiments *do* show

A degree of relativity has already been introduced by giving up the search for the correct intuition, and at the same time it should be acknowledged that our responses to a thought-experiment may be affected by the way and the context in which an imagined situation is described or presented. For instance, Parfit's description of the outcome of his thought-experiment involving a teletransporter makes it much easier for us to agree that we would survive the experience: he describes it as 'I' who enters the teletransporter, and 'I' who wakes up on Mars in the body made of new matter (Parfit 1984: 199).[4] And should Putnam's experiment be presented in the context of a zoology seminar, we may well react differently to Putnam. But these differences in reaction will usually not be radical ones: they may affect what we are disposed to say, yet it is extremely unlikely that they will reverse responses generally.

Williams's two experiments seem to be a special case. Even so, they do not justify the radical response outlined at the start of the previous section – that they be taken to undermine the method itself. They do not do so as long as some feature, or features, of their presentation can be isolated the presence of which explains the response evoked, and as long as this isolation and explanation produces a change of response. These claims are not very different from those made by Mark Johnston, where he does not reject thought-experiment as a method but suggests that our intuitive reactions to them are 'defeasible judgements' (Johnston 1987: 64, 80–83).[5]

In the case of the two experiments in question, such features can be seen in the second experiment. Perhaps the primary reason for singling out the second one as requiring explanation is that it doesn't take the straightforward form of the first. It is not as if we are simply presented with two scenarios, to the first of which we respond 'A emerges in the B-body' and to the other, 'A emerges in the A-body'. In that way although the two thought-experiments do at some stage involve descriptions of one and the same situation, they are nevertheless very different thought-experiments. It is not as if the second experiment is simply the first differently described. Ever since Locke's example we have found scenarios of apparent body-swaps compelling, and once the second thought-experiment is closely scrutinized, it becomes clear that we have no counter-example to that here.

How we are led to deny that A emerges in the B-body can be seen from the way in which the second case is presented. In the first place, Williams describes things as if they merely happen within the life of *one* person – amnesia is produced *in A*, illusory memories are induced in *A* – implying that it is still A who emerges after the change; but this is precisely what we are asked to judge, and our responses, I suggest, are being prejudiced by the description given.

A second misleading point is the way that the huge jump between steps (ii) and (iii) is disguised in the description. In step (ii) we are told that amnesia and 'certain character changes' are produced in A. In step (iii) there are 'changes in his character' and 'certain illusory memories are induced'. This does not sound like a very important development, certainly not enough to herald a change of identity – but it only comes across like that because the whole story is not being made clear. In step (iv) we learn that the psychological changes wrought are such that the character traits and memories shown and experienced match up to those of some other actual person, but that what goes on is essentially '*the same as (iii)*' (my italics). This means that the character and memory changes occurring in (iii) are sufficiently different from A's to be those of a *totally different*, even if non-actual, person. But then what happens in (iii) is nothing like the few minor changes that Williams's description suggests, but a massive – a total – invasion of A's psychology. Now, we may have no objection to the idea of a person surviving memory loss and some character change, but a total and irreversible change like this is another matter altogether. Were the extent of the change in (iii) made clear in the description, then it is much less likely that anyone would find the move from (ii) to (iii) intuitively acceptable.

These two points are enough to show why the general response to the second experiment has been as envisaged by Williams, and are sufficient to show that that response would not be straightforwardly elicited by a *clearer* description of the changes that are supposedly occurring over the various stages, or at least to show that that should remain a dominated response. But the modification I suggested in Section 3 as a way of getting

around Noonan's criticism of the experiment – that is, by introducing a spectrum of cases in which each case represents only a very slight change from the previous one – avoids this sort of objection. The objections just expressed are only to the way in which Williams describes the problem; a similar problem can be described in less misleading terms.

Even though the second thought-experiment can be set up in a way which avoids the problem of too large a change between its stages, I wish to argue that there remain reasons for not taking it and its apparent consequences too seriously. The original problem is set out as a series of cases, each one representing an apparently trivial development on the previous step. This step-by-step development along with the claim that to draw a line between any two adjacent steps would be irrational is a familiar pattern of argument. The amendment proposed in Section 3 not only avoids some objections, but offers a useful insight when it comes to the way the thought-experiment shapes our responses by drawing attention to this familiar pattern operating here. For once it is recognised that we are dealing with something of the same form as a Sorites problem, we should no longer be surprised that we are led into the invidious position that step (vi) represents, nor should we see the reformulated problem as presenting any fatal threat.

With the standard 'paradox of the heap', one is led by a series of small steps to a strongly counter-intuitive judgement; namely, that the two remaining stones in front of one still constitute a heap. But one does not just accept this judgement, nor does one accept that this shows a hopeless incoherence in one's belief-system. Likewise with Williams's example, one should realise that the conclusion we are led to is *strongly counter-intuitive* in that it conflicts with some of our fundamental principles; that is what the first experiment shows, and what we have realized ever since Locke. But then, as with the case of the concept of a heap, we can accept that the most the second experiment shows is that there is some vagueness in our concept of personal identity. That is hardly a startling admission and has nothing like the consequences of the devastating incoherence with which we seemed to be threatened.[6] Actual phenomena like the gradual onset of conditions like senility or Huntington's corea already force us to accept that there can be cases in which it is indeterminate whether or not the concepts of person or same person apply.

In discussing some of the dangers involved in the use of thought-experiments, Unger issues the following warning which is of import here:

> Even in favourable, revealing contexts, not all examples will elicit responses indicative of philosophically interesting attitudes in the area at which the case may be aimed. Dominant responses to these potentially dangerous examples may indicate, instead, certain rather general psychological tendencies that we have. Although sometimes useful to us in other ways, for purposes of philosophical inquiry these tendencies will be distorting (Unger 1990:12).

In the case of the revised version of Williams's second thought-experiment, and indeed of the original version, it is, I suggest, precisely one of these tendencies that is to be detected. Faced by the use of a Sorites-like development of the argument by small stages, we tend to ignore the larger picture. If we did not have this tendency we would never react in the sympathetic way we do to Williams's second experiment, but would see it for the body-swap that it is.

7. Williams's own response and why it is unconvincing

This account of how the description of the second experiment misleads us solves the problem posed by his conundrum. However, there is another argument which must be dealt with before we can move on to other thought-experiments. This argument is the final response which Williams makes to his own experiments. As I have done, he argues that one side of the experiment misleads by the way in which it is described; but unlike my case, he holds that it is the *first* description which is at fault.[7] This is Williams's argument:

> The apparently decisive arguments of the first presentation, which suggested that A should identify himself with the B-body-person, turned on the extreme neatness of the situation in satisfying, if any could, the description of 'changing bodies'. But this neatness is basically artificial; it is the product of the will of the experimenter to produce a situation which would naturally elicit, with minimum hesitation, that description. By the sorts of methods he employed, he could easily have left off earlier or gone on further. He could have stopped at [the point equivalent to] situation (v), leaving B as he was; or he could have gone on and produced two persons each with A-like character and memories.... If he had done either of those, we should have been in yet greater difficulty about what to say; he just chose to make it as easy as possible for us to find something to say.... The experimenter has...produced the one situation out of a range of equally possible situations which we should be most disposed to call a change of bodies. As against this, the principle that one's fears can extend to future pain whatever psychological changes precede it seems positively straightforward. Perhaps, indeed, it is not; but we need to be shown what is wrong with it. Until we are shown what is wrong with it, 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 (Williams 1970: 62–63, my parenthesis.

Does this leave us back in the position suggested by Noonan, able to ignore 'The self and the future' because both descriptions are flawed? I don't think that it does, because Williams's case here is not convincing. It is not the tentativeness of his proposal which fails to convince, but rather the fact that the considerations he raises are weak ones, especially in the light of the strong case presented against the second experiment in Section 6.

In the first place, we *can* show what is wrong with the principle that one's fears can extend to future pain despite radical intervening psychological changes. The principle is only plausible if the situation is described in a question-begging way (i.e. that *A* will be tortured), and the description of the degree of psychological change is fudged, as the discussion of Section 6 makes clear.

In the second place, it is by no means clear that the will of the experimenter to produce an easy response is as guilty as Williams suggests. Certainly, the experimenter could have stopped his description at (v), leaving two people with the character and memory of B, or he could have gone on to make two people like A as well. These situations raise interesting issues, but they do not stop one from reacting to the original experiment as one originally did. In *Reasons and persons*, Derek Parfit pays close attention to thought-experiments that raise exactly these issues, and argues convincingly that those experiments have no consequences which conflict with the view that personal identity can be

analysed in terms of psychological continuity.[8] Williams admits that his final answer is a risky one (1970: 63), and I contend he plumps for the wrong option.

8. Conclusion

In terms of the personal identity debate, the thought-experiments described in 'The self and the future' do not provide sufficient grounds for embracing the view that personal identity is unanalysable. Nor do they, on a more careful reading, support a bodily criterion of personal identity. Both of these positions are let down by the faults of the second experiment. Nevertheless, we are not justified (as Noonan has suggested we are) in ignoring the experiments altogether. This means that the thought-experiments ultimately suggest some support for the view that if personal identity is to be analysed, it is to be analysed in terms of psychological continuity.

As far as our methodological interests are concerned, I conclude that one aspect of the popular response to Williams's thought-experiments has been correct in that they *do* have something to teach us. It is not that the method of thought-experiment must be rejected. Nevertheless, the lesson is that we must temper our expectations of thought-experiments like those presented; they can indeed be informative, but they inform by revealing which principles underlying the application of a concept are most important to us given our conceptual scheme, not by some direct route to metaphysical reality.

Notes

1. The following list of works expressing opposition gives some idea of how widespread the censure of thought-experiments regarding personal identity has become: Baillie 1990, Johnston 1987, Kitcher 1979, Lowe 1990, White 1989, Wiggins 1980, Wilkes 1988. I do not plan to address the considerations raised by these writers here; I am simply interested in the case for opposition to the method raised by Williams's thought-experiments. I address some of the other arguments in 'Should we tolerate people who split?' (Beck 1992).
2. Madell uses a thought-experiment along these lines as an *argument for* his view on personal identity.
3. This is not to say that this is the only role thought experiments can play. They can also serve to bring to light that a particular view has internal inconsistencies or that it conflicts with some deeply entrenched principle. A good example of such an experiment is Parfit's 'My division' (1994: 253–265). See note 8.
4. Another fairly clear example of this sort of description influencing response can be seen in Chisholm's description and response to a case not unlike Williams's second experiment (Chisholm 1969:104–105).
5. There is experimental evidence supporting my view that our intuitive responses are defeasible judgements, which is particularly relevant to Williams's experiments and the terms in which they are described. Kahneman and Tversky set out case studies which show how people react in contradictory ways to choices they are offered under conditions of uncertainty, the subjects' reactions depending on whether the consequences of the choices are described in terms of gains or losses (Kahneman & Tversky 1984).
6. Even the so-called paradoxes which beset vague concepts may well not be insurmountable problems. Mark Sainsbury reviews a number of possible responses in his *Paradoxes* (Sainsbury, 1989).
7. Peter Unger concurs with Williams in taking this option (Unger 1990: 159). He also appeals to Williams as offering support to his case, in the process badly misrepresenting what Williams actually claims. I argue that Unger's case fails to add anything to Williams's in my 'Let's exist again (like we did last summer)' (forthcoming).
8. Interestingly, what emerges there is that cases of fission represent a different kind of thought-experiment to cases of body-swapping, for in the cases of fission (as Parfit points out) we end up in difficulty *no matter how* we respond intuitively. the thought-experiment thus works as a form of *reductio ad absurdum* rather than a test of our deepest commitments.

Bibliography

Baillie, J. 1990. Identity, survival and sortal concepts. *Philosophical Quarterly*, **40**: 183–194.
Beck, S. 1992. Should we tolerate people who split? *Southern Journal of Philosophy*, **30**: 1–17.
Beck, S. Forthcoming. Let's exist again (like we did last summer).
Chisholm, R. 1969. The loose and popular and the strict and philosophical senses of identity, in N. Care & R. Grimm (Eds), *Perception and personal identity*. Cleveland: Case Western Reserve Univ. Pr.
Johnston, M. 1987. Human beings. *Journal of Philosophy* 84: 59–83.
Kahneman, D. & Tversky, A. 1984. Choices, values and frames. *American Psychologist*, **39**: 341–350.
Kitcher, P. 1979. Natural kinds and unnatural persons. *Philosophy*, **54**: 541–547.
Leibniz, G.W. 1765. *New essays on human understanding*. P. Remnant & J. Bennett (Eds and Trans.). Cambridge: Cambridge Univ. Pr. 1981.
Locke, J. 1694. Of identity and diversity, in J. Perry (Ed.), *Personal identity*. Berkeley: University of California Press, 1975.
Lowe, E.J. 1990. Review of H. Noonan. Personal identity. *Mind*, **99**: 477–479.
Madell, G. 1981. *The identity of the self*. Edinburgh: Univ. Pr.
Noonan, H. 1989. *Personal identity*. London: Routledge and Kegan Paul.
Parfit, D. 1984. *Reasons and persons*. Oxford: Clarendon Press.
Putnam, H. 1963. It ain't necessarily so, in J. Rosenberg & B. Travis, *Readings in the philosophy of language*. Englewood Cliffs: Prentice-Hall, 1970.
Sainsbury, R.M. 1989. *Paradoxes*. Cambridge: Cambridge Univ. Pr.
Swinburne, R. 1984. Personal identity: the dualist theory, in S. Shoemaker & R. Swinburne, *Personal identity*. Oxford: Blackwell.
Unger, P. 1982. Towards a psychology of common sense. *American Philosophical Quarterly*, **19**: 117–129.
Unger, P. 1990. *Identity, consciousness and value*. Oxford: Oxford Univ. Pr.
White, S. 1989. Metapsychological relativism and the self. *Journal of Philosophy*, **86**: 298–323.
Wiggins, D. 1980. *Sameness and substance*. Oxford: Blackwell.
Wilkes, K. 1988. *Real people*. Oxford: Clarendon Press.
Williams, B. 1970. The self and the future, in *Problems of the self*. Cambridge: Cambridge Univ. Pr. 1973.