THE SELF AND THE FUTURE

SUPPOSE that there were some process to which two persons, *A* and *B*, could be subjected as a result of which they might be said—question-beggingly—to have *exchanged bodies*. That is to say—less question-beggingly—there is a certain human body which is such that when previously we were confronted with it, we were confronted with person *A*, certain utterances coming from it were expressive of memories of the past experiences of *A*, certain movements of it partly constituted the actions of *A* and were taken as expressive of the character of *A*, and so forth; but now, after the process is completed, utterances coming from this body are expressive of what seem to be just those memories which previously we identified as memories of the past experiences of *B*, its movements partly constitute actions expressive of the character of *B*, and so forth; and conversely with the other body.

There are certain important philosophical limitations on how such imaginary cases are to be constructed, and how they are to be taken when constructed in various ways. I shall mention two principal limitations, not in order to pursue them further here, but precisely in order to get them out of the way.

There are certain limitations, particularly with regard to character and mannerisms, to our ability to imagine such cases even in the most restricted sense of our being disposed to take the later performances of that body which was previously *A*'s as expressive of *B*'s character; if the previous *A* and *B* were extremely unlike one another both physically and psychologically, and if, say, in addition, they were of different sex, there might be grave difficulties in reading *B*'s dispositions in any possible performances of *A*'s body. Let us forget this, and for the present purpose just take *A* and *B* as being sufficiently alike (however alike that has to be) for the difficulty not to arise; 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. Thus the feat of imagining an exchange of bodies is supposed possible in the most restricted sense. But now there is a further limitation which has to be overcome if the feat is to be not merely possible in the most restricted sense but also is to have an outcome which, on serious reflection, we are prepared to describe as *A* and *B* having changed bodies—that is, an outcome where, confronted with what was previously *A*'s body, we are prepared seriously to say that we are now confronted with *B*.

It would seem a necessary condition of so doing that the utterances coming from that body be taken as genuinely expressive of memories of *B*'s past. But memory is a causal notion; and as we actually use it, it seems a necessary condition on *x*'s present knowledge of *x*'s earlier experiences constituting memory of those experiences that the causal chain linking the experiences and the knowledge should not run outside *x*'s body. Hence if utterances coming from a given body are to be taken as expressive of memories of the experiences of *B*, there should be some suitable causal link between the appropriate state of that body and the original happening of those experiences to *B*. One radical way of securing that condition in the imagined exchange case is to suppose, with Shoemaker,[1] that the brains of *A* and of *B* are transposed. We may not need so radical a condition. Thus suppose it were possible to extract information from a man's brain and store it in a device while his brain was repaired, or even renewed, the information then being replaced: it would seem exaggerated to insist that the resultant man could not possibly have the memories he had before the operation. With regard to our knowledge of our own past, we draw distinctions between merely recalling, being reminded, and learning again, and those distinctions correspond (roughly) to distinctions between no new input, partial new input, and total new input with regard to the information in question; and it seems clear that the information-parking case just imagined would not count as new input in the sense necessary and sufficient for "learning again." Hence we can imagine the case we are concerned with in terms of information extracted into such devices from *A*'s and *B*'s brains and replaced in the other brain;

[1] *Self-Knowledge and Self-Identity* (Ithaca, N. Y., 1963), p. 23 f.

this is the sort of model which, I think not unfairly for the present argument, I shall have in mind.

We imagine the following. The process considered above exists; two persons can enter some machine, let us say, and emerge changed in the appropriate ways. If *A* and *B* are the persons who enter, let us call the persons who emerge the *A-body-person* and the *B-body-person:* the *A*-body-person is that person (whoever it is) with whom I am confronted when, after the experiment, I am confronted with that body which previously was *A*'s body—that is to say, that person who would naturally be taken for *A* by someone who just saw this person, was familiar with *A*'s appearance before the experiment, and did not know about the happening of the experiment. A non-question-begging description of the experiment will leave it open which (if either) of the persons *A* and *B* the *A*-body-person is; the description of the experiment as "persons changing bodies" of course implies that the *A*-body-person is actually *B*.

We take two persons *A* and *B* who are going to have the process carried out on them. (We can suppose, rather hazily, that they are willing for this to happen; to investigate at all closely at this stage why they might be willing or unwilling, what they would fear, and so forth, would anticipate some later issues.) We further announce that one of the two resultant persons, the *A*-body-person and the *B*-body-person, is going after the experiment to be given $100,000, while the other is going to be tortured. We then ask each *A* and *B* to choose which treatment should be dealt out to which of the persons who will emerge from the experiment, the choice to be made (if it can be) on selfish grounds.

Suppose that *A* chooses that the *B*-body-person should get the pleasant treatment and the *A*-body-person the unpleasant treatment; and *B* chooses conversely (this might indicate that they thought that "changing bodies" was indeed a good description of the outcome). The experimenter cannot act in accordance with both these sets of preferences, those expressed by *A* and those expressed by *B*. Hence there is one clear sense in which *A* and *B* cannot both get what they want: namely, that if the experimenter, before the experiment, announces to *A* and *B* that he intends to carry out the alternative (for example), of treating the *B*-body-

person unpleasantly and the *A*-body-person pleasantly—then *A* can say rightly, "That's not the outcome I chose to happen," and *B* can say rightly, "That's just the outcome I chose to happen." So, evidently, *A* and *B* before the experiment can each come to know either that the outcome he chose will be that which will happen, or that the one he chose will not happen, and in that sense they can get or fail to get what they wanted. But is it also true that when the experimenter proceeds *after* the experiment to act in accordance with one of the preferences and not the other, then one of *A* and *B* will have got what he wanted, and the other not?

There seems very good ground for saying so. For suppose the experimenter, having elicited *A*'s and *B*'s preference, says nothing to *A* and *B* about what he will do; conducts the experiment; and then, for example, gives the unpleasant treatment to the *B*-body-person and the pleasant treatment to the *A*-body-person. Then the *B*-body-person will not only complain of the unpleasant treatment as such, but will complain (since he has *A*'s memories) that that was not the outcome he chose, since he chose that the *B*-body-person should be well treated; and since *A* made his choice in selfish spirit, he may add that he precisely chose in that way because he did not want the unpleasant things to happen to *him.* The *A*-body-person meanwhile will express satisfaction both at the receipt of the $100,000, and also at the fact that the experimenter has chosen to act in the way that he, *B*, so wisely chose. 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. For the same reasons it would seem that *A* and *B* in our example really did choose wisely, and that it was *A*'s bad luck that the choice he correctly made was not carried out, *B*'s good luck that the choice he correctly made was carried out. 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); and this in turn might be taken to show that in some sense of Descartes's obscure phrase, I and my body are "really distinct" (though, of course, nothing in these

considerations could support the idea that I could exist without a body at all).

These suggestions seem to be reinforced if we consider the cases where *A* and *B* make other choices with regard to the experiment. 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. Here again there can be no outcome which matches the expressed preferences of both of them: they cannot both get what they want. 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 acknowledgment (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, and would much prefer to be receiving what the *A*-body-person is receiving—namely, $100,000. 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. It looks, then, as though the *A*-body-person had gotten what he wanted, but not what he chose, while the *B*-body-person has gotten what he chose, but not what he wanted. So once more it looks as though they are, respectively, *B* and *A;* and that in this case the original choices of both *A* and *B* were unwise.

Suppose, lastly, that in the original choice *A* takes the line of the first case and *B* of the second: that is, *A* chooses that the *B*-body-person should get the money and the *A*-body-person the pain, and *B* chooses exactly the same thing. In this case, the experimenter would seem to be in the happy situation of giving both persons what they want—or at least, like God, what they have chosen. In this case, the *B*-body-person likes what he is receiving, recalls

choosing it, and congratulates himself on the wisdom of (as he puts it) his choice; while the *A*-body-person does not like what he is receiving, recalls choosing it, and is forced to acknowledge that (as he puts it) his choice was unwise. So once more we seem to get results to support the suggestions drawn from the first case.

Let us now consider the question, not of *A* and *B* choosing certain outcomes to take place after the experiment, but of their willingness to engage in the experiment at all. If they were initially inclined to accept the description of the experiment as "changing bodies" then one thing that would interest them would be the character of the other person's body. In this respect also what would happen after the experiment would seem to suggest that "changing bodies" was a good description of the experiment. If *A* and *B* agreed to the experiment, being each not displeased with the appearance, physique, and so forth of the other person's body; after the experiment the *B*-body-person might well be found saying such things as: "When I agreed to this experiment, I thought that *B*'s face was quite attractive, but now I look at it in the mirror, I am not so sure"; or the *A*-body-person might say "When I agreed to this experiment I did not know that *A* had a wooden leg; but now, after it is over, I find that I have this wooden leg, and I want the experiment reversed." It is possible that he might say further that he finds the leg very uncomfortable, and that the *B*-body-person should say, for instance, that he recalls that he found it very uncomfortable at first, but one gets used to it: but perhaps one would need to know more than at least I do about the physiology of habituation to artificial limbs to know whether the *A*-body-person would find the leg uncomfortable: that body, after all, has had the leg on it for some time. But apart from this sort of detail, the general line of the outcome regarded from this point of view seems to confirm our previous conclusions about the experiment.

Now let us suppose that when the experiment is proposed (in non-question-begging terms) *A* and *B* think rather of their psychological advantages and disadvantages. *A*'s thoughts turn primarily to certain sorts of anxiety to which he is very prone, while *B* is concerned with the frightful memories he has of past experiences which still distress him. They each hope that the

experiment will in some way result in their being able to get away from these things. They may even have been impressed by philosophical arguments to the effect that bodily continuity is at least a necessary condition of personal identity: *A*, for example, reasons that, granted the experiment comes off, then the person who is bodily continuous with him will not have this anxiety, while the other person will no doubt have some anxiety—perhaps in some sense his anxiety—and at least that person will not be he. The experiment is performed and the experimenter (to whom *A* and *B* previously revealed privately their several difficulties and hopes) asks the *A*-body-person whether he has gotten rid of his anxiety. This person presumably replies that he does not know what the man is talking about; he never had such anxiety, but he did have some very disagreeable memories, and recalls engaging in the experiment to get rid of them, and is disappointed to discover that he still has them. The *B*-body-person will react in a similar way to questions about his painful memories, pointing out that he still has his anxiety. These results seem to confirm still further the description of the experiment as "changing bodies." And all the results suggest that the only rational thing to do, confronted with such an experiment, would be to identify oneself with one's memories, and so forth, and not with one's body. The philosophical arguments designed to show that bodily continuity was at least a necessary condition of personal identity would seem to be just mistaken.

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.

If this is right, the whole question seems now to be totally mysterious. For what we have just been through is of course merely one side, differently represented, of the transaction which we considered before; and it represents it as a perfectly hateful prospect, while the previous considerations represented it as something one should rationally, perhaps even cheerfully, choose out of the options there presented. It is differently presented, of course, and in two notable respects; but when we look at these two differences of presentation, can we really convince ourselves that the second presentation is wrong or misleading, thus

leaving the road open to the first version which at the time seemed so convincing? Surely not.

The first difference is that in the second version the torture is throughout represented as going to happen to *me:* "you," the man in charge persistently says. Thus he is not very neutral. But should he have been neutral? Or, to put it another way, does his use of the second person have a merely emotional and rhetorical effect on me, making me afraid when further reflection would have shown that I had no reason to be? It is certainly not obviously so. The problem just is that through every step of his predictions I seem to be able to follow him successfully. And if I reflect on whether what he has said gives me grounds for fearing that I shall be tortured, I could consider that behind my fears lies some principle such as this: that my undergoing physical pain in the future is not excluded by any psychological state I may be in at the time, with the platitudinous exception of those psychological states which in themselves exclude experiencing pain, notably (if it is a psychological state) unconsciousness. In particular, what impressions I have about the past will not have any effect on whether I undergo the pain or not. This principle seems sound enough.

It is an important fact that not everything I would, as things are, regard as an evil would be something that I should rationally fear as an evil if it were predicted that it would happen to me in the future and also predicted that I should undergo significant psychological changes in the meantime. For the fact that I regard that happening, things being as they are, as an evil can be dependent on factors of belief or character which might themselves be modified by the psychological changes in question. Thus if I am appallingly subject to acrophobia, and am told that I shall find myself on top of a steep mountain in the near future, I shall to that extent be afraid; but if I am told that I shall be psychologically changed in the meantime in such a way as to rid me of my acrophobia (and as with the other prediction, I believe it), then I have no reason to be afraid of the predicted happening, or at least not the same reason. Again, I might look forward to meeting a certain person again with either alarm or excitement because of my memories of our past relations. In some part, these memories

operate in connection with my emotion, not only on the present time, but projectively forward: for it is to a meeting itself affected by the presence of those memories that I look forward. If I am convinced that when the time comes I shall not have those memories, then I shall not have just the same reasons as before for looking forward to that meeting with the one emotion or the other. (Spiritualism, incidentally, appears to involve the belief that I have just the same reasons for a given attitude toward encountering people again after I am dead, as I did before: with the one modification that I can be sure it will all be very nice.)

Physical pain, however, the example which for simplicity (and not for any obsessional reason) I have taken, is absolutely minimally dependent on character or belief. No amount of change in my character or my beliefs would seem to affect substantially the nastiness of tortures applied to me; correspondingly, no degree of predicted change in my character and beliefs can unseat the fear of torture which, together with those changes, is predicted for me.

I am not at all suggesting that the *only* basis, or indeed the only rational basis, for fear in the face of these various predictions is how things will be relative to my psychological state in the eventual outcome. I am merely pointing out that this is one component; it is not the only one. For certainly one will fear and otherwise reject the changes themselves, or in very many cases one would. Thus one of the old paradoxes of hedonistic utilitarianism; if one had assurances that undergoing certain operations and being attached to a machine would provide one for the rest of one's existence with an unending sequence of delicious and varied experiences, one might very well reject the option, and react with fear if someone proposed to apply it compulsorily; and that fear and horror would seem appropriate reactions in the second case may help to discredit the interpretation (if anyone has the nerve to propose it) that one's reason for rejecting the option voluntarily would be a consciousness of duties to others which one in one's hedonic state would leave undone. The prospect of contented madness or vegetableness is found by many (not perhaps by all) appalling in ways which are obviously not a function of how things would then be for them, for things would then be for them not

appalling. In the case we are at present discussing, these sorts of considerations seem merely to make it clearer that the predictions of the man in charge provide a double ground of horror: at the prospect of torture, and at the prospect of the change in character and in impressions of the past that will precede it. And certainly, to repeat what has already been said, the prospect of the second certainly seems to provide no ground for rejecting or not fearing the prospect of the first.

I said that there were two notable differences between the second presentation of our situation and the first. The first difference, which we have just said something about, was that the man predicted the torture for *me*, a psychologically very changed "me." We have yet to find a reason for saying that he should not have done this, or that I really should be unable to follow him if he does; I seem to be able to follow him only too well. The second difference is that in this presentation he does not mention the other man, except in the somewhat incidental role of being the provenance of the impressions of the past I end up with. He does not mention him at all as someone who will end up with impressions of the past derived from me (and, incidentally, with $100,000 as well—a consideration which, in the frame of mind appropriate to this version, will merely make me jealous).

But why *should* he mention this man and what is going to happen to him? My selfish concern is to be told what is going to happen to me, and now I know: torture, preceded by changes of character, brain operations, changes in impressions of the past. The knowledge that one other person, or none, or many will be similarly mistreated may affect me in other ways, of sympathy, greater horror at the power of this tyrant, and so forth; but surely it cannot affect my expectations of torture? But—someone will say—this is to leave out exactly the feature which, as the first presentation of the case showed, makes all the difference: for it is to leave out the person who, as the first presentation showed, will be you. It is to leave out not merely a feature which should fundamentally affect your fears, it is to leave out the very person for whom you are fearful. So of course, the objector will say, this makes all the difference.

But can it? Consider the following series of cases. In each case

we are to suppose that after what is described, *A* is, as before, to be tortured; we are also to suppose the person *A* is informed beforehand that just these things followed by the torture will happen to him:

(*i*) *A* is subjected to an operation which produces total amnesia;

(*ii*) amnesia is produced in *A*, and other interference leads to certain changes in his character;

(*iii*) changes in his character are produced, and at the same time certain illusory "memory" beliefs are induced 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 the "memory" impressions are designed to be appropriate to another actual person, *B;*

(*v*) the same as (*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;

(*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.

I take it that no one is going to dispute that *A* has reasons, and fairly straightforward reasons, for fear of pain when the prospect is that of situation (*i*); there seems no conceivable reason why this should not extend to situation (*ii*), and the situation (*iii*) can surely introduce no difference of principle—it just seems a situation which for more than one reason we should have grounds for fearing, as suggested above. Situation (*iv*) at least introduces the person *B*, who was the focus of the objection we are now discussing. But it does not seem to introduce him in any way which makes a material difference; if I can expect pain through a transformation which involves new "memory"-impressions, it would seem a purely external fact, relative to that, that the "memory"-impressions had a model. Nor, in (*iv*), do we satisfy a causal condition which I mentioned at the beginning for the "memories" actually being memories; though notice that if the

job were done thoroughly, I might well be able to elicit from the *A*-body-person the kinds of remarks about his previous expectations of the experiment—remarks appropriate to the original *B*—which so impressed us in the first version of the story. I shall have a similar assurance of this being so in situation (*v*), where, moreover, a plausible application of the causal condition is available.

But two things are to be noticed about this situation. First, if we concentrate on *A* and the *A*-body-person, we do not seem to have added anything which from the point of view of his fears makes any material difference; just as, in the move from (*iii*) to (*iv*), it made no relevant difference that the new "memory"-impressions which precede the pain had, as it happened, a model, so in the move from (*iv*) to (*v*) all we have added is that they have a model which is also their cause: and it is still difficult to see why that, to him looking forward, could possibly make the difference between expecting pain and not expecting pain. To illustrate that point from the case of character: if *A* is capable of expecting pain, he is capable of expecting pain preceded by a change in his dispositions—and to that expectation it can make no difference, whether that change in his dispositions is modeled on, or indeed indirectly caused by, the dispositions of some other person. If his fears can, as it were, reach through the change, it seems a mere trimming how the change is in fact induced. The second point about situation (*v*) is that if the crucial question for *A*'s fears with regard to what befalls the *A*-body-person is whether the *A*-body-person is or is not the person *B*,[2] then that condition has not yet been satisfied in situation (*v*): for there we have an undisputed *B* in addition to the *A*-body-person, and certainly those two are not the same person.

But in situation (*vi*), we seemed to think, that is finally what he is. But if *A*'s original fears could reach through the expected changes in (*v*), as they did in (*iv*) and (*iii*), then certainly they can reach through in (*vi*). Indeed, from the point of view of *A*'s expectations and fears, there is less difference between (*vi*) and (*v*) than there is between (*v*) and (*iv*) or between (*iv*) and (*iii*). In

[2] This of course does not have to be the crucial question, but it seems one fair way of taking up the present objection.

those transitions, there were at least differences—though we could not see that they were really relevant differences—in the content and cause of what happened to him; in the present case there is absolutely no difference at all in what happens to him, the only difference being in what happens to someone else. If he can fear pain when (*v*) is predicted, why should he cease to when (*vi*) is?

I can see only one way of relevantly laying great weight on the transition from (*v*) to (*vi*); and this involves a considerable difficulty. This is to deny that, as I put it, the transition from (*v*) to (*vi*) involves merely the addition of something happening to *somebody else;* what rather it does, it will be said, is to involve the reintroduction of *A* himself, as the *B*-body-person; since he has reappeared in this form, it is for this person, and not for the unfortunate *A*-body-person, that *A* will have his expectations. This is to reassert, in effect, the viewpoint emphasized in our first presentation of the experiment. But this surely has the consequence that *A* should not have fears for the *A*-body-person who appeared in situation (*v*). For by the present argument, the *A*-body-person in (*vi*) is not *A;* the *B*-body-person is. But the *A*-body-person in (*v*) is, in character, history, everything, exactly the same as the *A*-body-person in (*vi*); so if the latter is not *A*, then neither is the former. (It is this point, no doubt, that encourages one to speak of the difference that goes with [*vi*] as being, on the present view, the *reintroduction* of *A*.) But no one else in (*v*) has any better claim to be *A*. So in (*v*), it seems, *A* just does not exist. This would certainly explain why *A* should have no fears for the state of things in (*v*)—though he might well have fears for the path to it. But it rather looked earlier as though he could well have fears for the state of things in (*v*). Let us grant, however, that that was an illusion, and that *A* really does not exist in (*v*); then does he exist in (*iv*), (*iii*), (*ii*), or (*i*)? It seems very difficult to deny it for (*i*) and (*ii*); are we perhaps to draw the line between (*iii*) and (*iv*)?

Here someone will say: you must not insist on drawing a line—borderline cases are borderline cases, and you must not push our concepts beyond their limits. But this well-known piece of advice, sensible as it is in many cases, seems in the present case to involve an extraordinary difficulty. It may intellectually comfort observers of *A*'s situation; but what is *A* supposed to make of it? To be told

that a future situation is a borderline one for its being myself that is hurt, that it is conceptually undecidable whether it will be me or not, is something which, it seems, I can do nothing with; because, in particular, it seems to have no comprehensible representation in my expectations and the emotions that go with them.

If I expect that a certain situation, *S*, will come about in the future, there is of course a wide range of emotions and concerns, directed on *S*, which I may experience now in relation to my expectation. Unless I am exceptionally egoistic, it is not a condition on my being concerned in relation to this expectation, that I myself will be involved in *S*—where my being "involved" in *S* means that I figure in *S* as someone doing something at that time or having something done to me, or, again, that *S* will have consequences affecting me at that or some subsequent time. There are some emotions, however, which I will feel only if I will be involved in *S*, and fear is an obvious example.

Now the description of *S* under which it figures in my expectations will necessarily be, in various ways, indeterminate; and one way in which it may be indeterminate is that it leave open whether I shall be involved in *S* or not. Thus I may have good reason to expect that one out of us five is going to get hurt, but no reason to expect it to be me rather than one of the others. My present emotions will be correspondingly affected by this indeterminacy. Thus, sticking to the egoistic concern involved in fear, I shall presumably be somewhat more cheerful than if I knew it was going to be me, somewhat less cheerful than if I had been left out altogether. Fear will be mixed with, and qualified by, apprehension; and so forth. These emotions revolve around the thought of the eventual determination of the indeterminacy; moments of straight fear focus on its really turning out to be me, of hope on its turning out not to be me. All the emotions are related to the coming about of what I expect: and what I expect in such a case just cannot come about save by coming about in one of the ways or another.

There are other ways in which indeterminate expectations can be related to fear. Thus I may expect (perhaps neurotically) that something nasty is going to happen to me, indeed expect that when it happens it will take some determinate form, but have no

range, or no closed range, of candidates for the determinate form to rehearse in my present thought. Different from this would be the fear of something radically indeterminate—the fear (one might say) of a nameless horror. If somebody had such a fear, one could even say that he had, in a sense, a perfectly determinate expectation: if what he expects indeed comes about, there will be nothing more determinate to be said about it after the event than was said in the expectation. Both these cases of course are cases of *fear* because one thing that is fixed amid the indeterminacy is the belief that it is to me to which the things will happen.

Central to the expectation of *S* is the thought of what it will be like when it happens—thought which may be indeterminate, range over alternatives, and so forth. When *S* involves me, there can be the possibility of a special form of such thought: the thought of how it will be for me, the imaginative projection of myself as participant in *S*.[3]

I do not have to think about *S* in this way, when it involves me; but I may be able to. (It might be suggested that this possibility was even mirrored in the language, in the distinction between "expecting to be hurt" and "expecting that I shall be hurt"; but I am very doubtful about this point, which is in any case of no importance.)

Suppose now that there is an *S* with regard to which it is for conceptual reasons undecidable whether it involves me or not, as is proposed for the experimental situation by the line we are discussing. It is important that the expectation of *S* is not *indeterminate* in any of the ways we have just been considering. It is not like the nameless horror, since the fixed point of that case was that it was going to happen to the subject, and that made his state unequivocally fear. Nor is it like the expectation of the man who expects one of the five to be hurt; his fear was indeed equivocal, but its focus, and that of the expectation, was that when *S* came about, it would certainly come about in one way or the other. In the present case, fear (of the torture, that is to say, not of the initial experiment) seems neither appropriate, nor inappropriate, nor

[3] For a more detailed treatment of issues related to this, see *Imagination and the Self*, British Academy (London, 1966); reprinted in P. F. Strawson (ed.), *Studies in Thought and Action* (Oxford, 1968).

appropriately equivocal. Relatedly, the subject has an incurable difficulty about how he may think about *S*. If he engages in projective imaginative thinking (about how it will be for him), he implicitly answers the necessarily unanswerable question; if he thinks that he cannot engage in such thinking, it looks very much as if he also answers it, though in the opposite direction. Perhaps he must just refrain from such thinking; but is he just refraining from it, if it is incurably undecidable whether he can or cannot engage in it?

It may be said that all that these considerations can show is that fear, at any rate, does not get its proper footing in this case; but that there could be some other, more ambivalent, form of concern which would indeed be appropriate to this particular expectation, the expectation of the conceptually undecidable situation. There are, perhaps, analogous feelings that actually occur in actual situations. Thus material objects do occasionally undergo puzzling transformations which leave a conceptual shadow over their identity. Suppose I were sentimentally attached to an object to which this sort of thing then happened; then it might be that I could neither feel about it quite as I did originally, nor be totally indifferent to it, but would have some other and rather ambivalent feeling toward it. Similarly, it may be said, toward the prospective sufferer of pain, my identity relations with whom are conceptually shadowed, I can feel neither as I would if he were certainly me, nor as I would if he were certainly not, but rather some such ambivalent concern.

But this analogy does little to remove the most baffling aspect of the present case—an aspect which has already turned up in what was said about the subject's difficulty in thinking either projectively or non-projectively about the situation. For to regard the prospective pain-sufferer *just* like the transmogrified object of sentiment, and to conceive of my ambivalent distress about his future pain as just like ambivalent distress about some future damage to such an object, is of course to leave him and me clearly distinct from one another, and thus to displace the conceptual shadow from its proper place. I have to get nearer to him than that. But is there any nearer that I can get to him without expecting his pain? If there is, the analogy has not shown us it. We can

certainly not get nearer by expecting, as it were, *ambivalent* pain; there is no place at all for that. There seems to be an obstinate bafflement to mirroring in my expectations a situation in which it is conceptually undecidable whether I occur.

The bafflement seems, moreover, to turn to plain absurdity if we move from conceptual undecidability to its close friend and neighbor, conventionalist decision. This comes out if we consider another description, overtly conventionalist, of the series of cases which occasioned the present discussion. This description would reject a point I relied on in an earlier argument—namely, that if we deny that the *A*-body-person in (*vi*) is *A* (because the *B*-body-person is), then we must deny that the *A*-body-person in (*v*) is *A*, since they are exactly the same. "No," it may be said, "this is just to assume that we say the same in different sorts of situation. No doubt when we have the very good candidate for being *A*—namely, the *B*-body-person—we call him *A;* but this does not mean that we should not call the *A*-body-person *A* in that other situation when we have no better candidate around. Different situations call for different descriptions." This line of talk is the sort of thing indeed appropriate to lawyers deciding the ownership of some property which has undergone some bewildering set of transformations; they just have to decide, and in each situation, let us suppose, it has got to go to somebody, on as reasonable grounds as the facts and the law admit. But as a line to deal with a person's fears or expectations about his own future, it seems to have no sense at all. If *A*'s fears can extend to what will happen to the *A*-body-person in (*v*), I do not see how they can be rationally diverted from the fate of the exactly similar person in (*vi*) by his being told that someone would have a reason in the latter situation which he would not have in the former for deciding to call another person *A*.

Thus, to sum up, it looks as though there are two presentations of the imagined experiment and the choice associated with it, each of which carries conviction, and which lead to contrary conclusions. The idea, moreover, that the situation after the experiment is conceptually undecidable in the relevant respect seems not to assist, but rather to increase, the puzzlement; while the idea (so often appealed to in these matters) that it is conven-

tionally decidable is even worse. Following from all that, I am not in the least clear which option it would be wise to take if one were presented with them before the experiment. I find that rather disturbing.

Whatever the puzzlement, there is one feature of the arguments which have led to it which is worth picking out, since it runs counter to something which is, I think, often rather vaguely supposed. It is often recognized that there are "first-personal" and "third-personal" aspects of questions about persons, and that there are difficulties about the relations between them. It is also recognized that "mentalistic" considerations (as we may vaguely call them) and considerations of bodily continuity are involved in questions of personal identity (which is not to say that there are mentalistic and bodily criteria of personal identity). It is tempting to think that the two distinctions run in parallel: roughly, that a first-personal approach concentrates attention on mentalistic considerations, while a third-personal approach emphasizes considerations of bodily continuity. The present discussion is an illustration of exactly the opposite. The first argument, which led to the "mentalistic" conclusion that *A* and *B* would change bodies and that each person should identify himself with the destination of his memories and character, was an argument entirely conducted in third-personal terms. The second argument, which suggested the bodily continuity identification, concerned itself with the first-personal issue of what *A* could expect. That this is so seems to me (though I will not discuss it further here) of some significance.

I will end by suggesting one rather shaky way in which one might approach a resolution of the problem, using only the limited materials already available.

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 situation (*v*), leaving *B* as he was; or he could have gone on and produced two persons each with *A*-like character and memories, as well as one or two with *B*-like characteristics. 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. Now if we had some model of ghostly persons in bodies, which were in some sense actually moved around by certain procedures, we could regard the neat experiment just as the *effective* experiment: the one method that really did result in the ghostly persons' changing places without being destroyed, dispersed, or whatever. But we cannot seriously use such a model. The experimenter has not in the sense of that model *induced* a change of bodies; he has rather 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. It would be risky: that there is room for the notion of a *risk* here is itself a major feature of the problem.

BERNARD WILLIAMS

King's College, Cambridge