The psychological theory & the dissolving self

Last time we were discussing the psychological theory of persons, and closed by discussing Reid's argument against that view. Recall that the argument went like this:

Let A be a person stage of yours at the age of 5, and let B be a person stage of yours at the age of 13, and let C be a person stage of yours at the age of 17. Then then the following seems possible: C has memories of the experiences of B, and B has memories of the experiences of A, but C does not have memories of the experiences of A.

The problem is that this seems to leads to the following argument:

1. Two person stages are the same person if and only if if the later has memories of the earlier. (The Memory Theory)
2. C has memories of the experiences of B.
3. C=B (1,2)
4. B has memories of the experiences of A.
5. B=A (1,4)
6. C does not have memories of the experiences of A.
7. C≠A (1,6)
8. C=A (3,5)

C. C=A & C≠A (7,8)

We noted that this argument has a false conclusion, and hence that one of two things must be true: either the argument is invalid, or it has at least one false premise. But the argument certainly seems valid; so it seens as though it must have a false premise. So long as we think that Reid's scenario is really possible, then it seems that thefalse premise must be premise (1), which states the Memory Theory.

Because this sort of argument attempts to show the falsity of one of the premises by showing that it leads to an absurd conclusion, an argument of this sort is called a *reductio ad absurdum*, or just a *reductio.* Many times, it's easier for argue for the conclusion you want by *reductio* than by giving a direct argument for it.

Let's return to Reid's argument. That argument seemed to show that premise (1) of the argument just sketched is false. That was:

1. Two person stages are the same person if and only if if the later has memories of the earlier.

This premise can be thought of as having two parts:

1a. If two person stages are the same person, then the later has memories of the earlier.

1b. If a later person-stage has memories of the earlier one, then they are the same person.

Premise (1) is equivalent to the combination of (1a) and (1b). We mentioned earlier that if-then sentences are sometimes called *conditionals*. Sentences like premise (1) are called *biconditionals*, because they are like the combination of two conditionals, 'if p, then q' and 'if q, then p.' You can write this as 'p if and only if q' (or 'q if and only if p' -- order does not matter).

One suggestion for the memory theory is that they should give up (1a). This would be enough to solve the problem posed by the example of drunkenness and to block Reid's argument.

Does this really solve the problem posed by Reid's argument? Suppose we replace premise (1)

> 1. Two person stages are the same person if and only if if the later has memories of the earlier.

with premise (1b):

> 1b. If a later person-stage has memories of the earlier one, then they are the same person.

Then we would have the following argument:

1b. If a later person-stage has memories of the earlier one, then they are the same person. (The Modified Memory Theory)
2. C has memories of the experiences of B.
3. C=B (1b,2)
4. B has memories of the experiences of A.
5. B=A (1b,4)
6. C does not have memories of the experiences of A.
7. C≠A (1b,6)
8. C=A (3,5)

C. C=A & C≠A (7,8)

Is this argument still valid? If not, why not?

The argument is no longer valid, because the inference from premises (1b) and (6) to (7) is invalid. You can think of that inference as being of the form 'If p, then q; not-p; therefore, not-q.' But this is not a valid form of inference.

Let's put all these points together. First, in order to handle the case of false memories, the memory theory has to incorporate some distinction between memories which are caused in the right way, and ones which are not.

Second, in order to respond to Reid's argument, the memory theorist should reject the claim that

> If two person stages are the same person, then the later has memories of the earlier.

That means that they have to reject our theory as initially formulated, which was

> <u>The memory theory of persons (1st version)</u>
> Two person-stages are stages of the same person if and only if the later person-stage contains memories of the earlier person-stage.

Instead, the memory theory should be stated as follows:

> <u>The memory theory of persons (2nd version)</u>
> Two person-stages are stages of the same person if the later person-stage contains memories, caused in the right way, of the earlier person-stage.

Then we can add this claim:

If A and B are the same person, and B and C are the same person, then A and C are the same person.

Our final version of the memory theory might then intuitively be stated like this: A and B are the same person if and only if either (1) A contains memories, caused in the right way, of B, or (2) there is some series of person-stages connecting A and B which is such that each person-stage in the series has memories (caused in the right way) of the immediately preceding person-stage in the series, and B is the first person-stage in the series, and A is the last.

Does this solve all the problems with the memory theory that we have discussed so far?

In a way, Parfit thinks that the memory theory is the right view of personal identity. But he thinks that the view has some very surprising consequences.

In the article we read, he calls this view the “bundle theory.” The reason is that according to the psychological theory, a person exists over time because a series (a “bundle”) of person-stages is related in a certain way. According to dualist and materialist views, by contrast, a person’s existence over time is simply a matter of a certain thing -- whether immaterial or material -- which exists at each time. In his terms, both dualist and materialist views of this sort could be thought of as versions of the “ego theory” (even though the name suggests dualism rather than materialism).

One way to bring out some of the surprising aspects of the psychological theory is by considering the example of *teletransportation*. Here's how Parfit describes the case in a longer work of his, entitled *Reasons and Persons*:

I enter the Teletransporter. I have been to Mars before, but only by the old method, a space-ship journey taking several weeks. This machine will send me at the speed of light. I merely have to press the green button. Like others, I am nervous. Will it work? I remind myself what I have been told to expect. When I press the button, I shall lose consciousness, and then wake up at what seems a moment later. In fact I shall have been unconscious for about an hour. The Scanner here on Earth will destroy my brain and body, while recording the exact states of all of my cells. It will then transmit this information by radio. Travelling at the speed of light, the message will take three minutes to reach the Replicator on Mars. This will then create, out of new matter, a brain and body exactly like mine. It will be in this body that I shall wake up.

Though I believe that this is what will happen, I still hesitate. But then I remember seeing my wife grin when, at breakfast today, I revealed my nervousness. As she reminded me, she has been often teletransported, and there is nothing wrong with *her*. I press the button. As predicted, I lose and seem at once to regain consciousness, but in a different cubicle. Examining my new body, I find no change at all. Even the cut on my upper lip, from this morning's shave, is still there.

If you believe the psychological theory, it seems natural to think of teletransportation as an unproblematic (and especially convenient) mode of transportation. After all, the person who emerges from the cubicle has exactly the same memories as the person who stepped in; and the person has these memories *because* of the experiences of the person who stepped into the cubicle on earth. Indeed, by comparison with the sorts of examples of body-switching discussed last time, teletransportation seems rather unproblematic.

But it is not unproblematic, as Parfit's continuation of his story shows:

> Several years pass, during which I am often Teletransported. I am now back in the cubicle, ready for another trip to Mars. But this time, when I press the green button, I do not lose consciousness. There is a whirring sound, then silence. I leave the cubicle, and say to the attendant: 'It's not working. What did I do wrong?'
>
> 'It's working', he replies, handing me a printed card. This reads: 'The New Scanner records your blueprint without destroying your brain and body. We hope that you will welcome the opportunities which this technical advance offers.'
>
> The attendant tells me that I am one of the first people to use the New Scanner. He adds that, if I stay for an hour, I can use the Intercom to see and talk to myself on Mars.
>
> 'Wait a minute', I reply, 'If I'm here I can't *also* be on Mars'.

Why might this sort of example seem to pose a problem for the psychological view?

The problem here is analogous to the problem posed by Reid's argument.

Let Original-Parfit = Parfit before he stepped into the teletransporter.
Let Earth-Parfit = the person who gets out of the teletransporter on earth.
Let Mars-Parfit = the person who gets out of the teletransporter on Mars.

Then it seems as though all of the following are true, if the psychological theory is true:

Original-Parfit = Mars-Parfit
Original-Parfit = Earth-Parfit
Mars-Parfit ≠ Earth-Parfit

Why does it seem as though, if the psychological theory is true, each of these claims must be true?

The problem is that all three of these **cannot** be true. So something has to give.

There seem to be two ways that the psychological theorist can respond to this sort of example:

1. Try to set up your theory so that while Original-Parfit = Earth-Parfit, Original-Parfit ≠ Mars-Parfit. According to this option, as Parfit puts it

> **if you chose to be teletransported, believing this to be the fastest way of travelling, you would be making a terrible mistake. This would not be a way of travelling, but a way of dying.**

2. Say that, strictly speaking, Original-Parfit ≠ Earth-Parfit and Original-Parfit ≠ Mars-Parfit. Strictly speaking, no person at one time is identical to a person at any other time. Personal identity (and hence survival over time) is just a matter of degree of similarity.

Option 1 seems more sane, at first glance. So let's see how we might argue that Original-Parfit = Earth-Parfit, even though Original-Parfit ≠ Mars-Parfit.

A natural line of response would be to say that personal identity requires not just psychological continuity, but also physical continuity. And, clearly, Earth-Parfit is physically continuous with Original-Parfit in a way that Mars-Parfit is not.

Why might this line of response be worrying for the psychological theorist, given the sorts of examples used to motivate her theory?

Moreover, as Parfit observes, it seems that the proponent of any criterion of physical continuity must say that while replacing 100% of one's cells is inconsistent with identity, replacing 1% of one's cells is not. (After all, we don't think successful organ replacements always result in the creation of a new person.) But this leads to a kind of dilemma:

> **If these beliefs were correct, there must be some critical percentage, somewhere in this range of cases, up to which the resulting person would be you, and beyond which he would merely be your Replica. Perhaps, for example, it would be you who would wake up if the proportion of cells replaced were 49 per cent, but if just a few more cells were also replaced, this would make all the difference, causing it to be someone else who would wake up.**
>
> **That there must be some such critical percentage follows from our natural beliefs. But this conclusion is most implausible. How could a few cells make such a difference? Moreover, if there is such a critical percentage, no one could ever discover where it came. Since in all these cases the resulting person would believe that he was you, there could never be any evidence about where, in this range of cases, he would suddenly cease to be you.**

So Parfit endorses the second response to cases of teletransportation: all that there is to say about such cases is that both people are, to some extent, the same person as Original-Parfit.

But this view can seem crazy; if it is true, then almost all of our normal beliefs about ourselves and our identity over time are wrong. It is hard to believe that the only thing to say about the example of teletransportation is that the the person on Mars is to some extent identical to your prior self, and that the person on Earth is to some extent identical to your prior self; both are, to some extent, you. We are strongly inclined to believe that there must be some fact of the matter about which, if either, is you.

We might sum up this “commonsense” view of persons by saying that personal identity is always an all-or-nothing matter; any person either is you, or is not you --- there’s no middle ground.

Parfit thinks that the example of teletransportation (plus the discussion of incrementally larger physical changes) provides one sort of argument against this commonsense view about persons. But in the article we read, he also provides one other argument.

This argument is based on the sorts of split-brain cases he describes. Here's an initial, simplified description of such a case:

> IT WAS THE split-brain cases which drew me into philosophy. Our knowledge of these cases depends on the results of various psychological tests, as described by Donald MacKay.[1] These tests made use of two facts. We control each of our arms, and see what is in each half of our visual fields, with only one of our hemispheres. When someone's hemispheres have been disconnected, psychologists can thus present to this person two different written questions in the two halves of his visual field, and can receive two different answers written by this person's two hands.
>
> Here is a simplified imaginary version of the kind of evidence that such tests provide. One of these people looks fixedly at the centre of a wide screen, whose left half is red and right half is blue. On each half in a darker shade are the words, "How many colours can you see?" With both hands the person writes, "Only one." The words are now changed to read, "Which is the only colour that you can see?" With one of his hands the person writes "Red," with the other he writes "Blue."

(This is in important respects a simplification of the experimental data; for those who want a less simplified discussion, see the optional reading on the course web site by Nagel, "Brain bisection and the unity of consciousness.")

Even more dramatic examples result from consideration of speech, which is controlled by the brain's left hemisphere. Here are some examples. (The descriptions are from the Nagel article.)

> If a concealed object is placed in the left hand and the person is asked to guess what it is, wrong guesses will elicit an annoyed frown, since the right hemisphere, which receives the tactile information, also hears the answers. If the speaking hemisphere should guess correctly, the result is a smile. A smell fed to the right nostril (which stimulates the right hemisphere) will elicit a verbal denial that the subject smells anything, but if asked to point with the left hand at a corresponding object he will succeed in picking out e.g. a clove of garlic, protesting all the while that he smells absolutely nothing, so how can he possibly point to what he smells.

Even more dramatic examples result from consideration of speech, which is controlled by the brain's left hemisphere. Here are some examples. (The descriptions are from the Nagel article.)

> One particularly poignant example of conflict between the hemispheres is as follows. A pipe is placed out of sight in the patient's left hand, and he is then asked to write with his left hand what he was holding. Very laboriously and heavily, the left hand writes the letters P and I. Then suddenly the writing speeds up and becomes lighter, the I is converted to an E, and the word is completed as PENCIL. Evidently the left hemisphere has made a guess based on the appearance of the first two letters, and has interfered, with ipsilateral control. But then the right hemisphere takes over control of the hand again, heavily crosses out the letters ENCIL, and draws a crude picture of a pipe.[6]

How do these split brain cases challenge the commonsense view of persons?

In such cases -- think of the simple blue/red case -- it seems that there are two separate streams of consciousness. But it seems that one person can't have two separate streams of consciousness, at least of this sort. So it seems that, in the case of split brain patients, there are (at least) two persons inhabiting a single body.

But now imagine a surgery to repair the corpus callosum. Surely such a surgery needn't involve ending the life of a person. But then if there were (at least) two persons before the surgery, there must be (at least) two persons afterward. (Similarly, it is odd to think that severing someone's corpus callosum involves the creation of a person.)

But you and I are just like a split-brain patient after such surgery (or before having their corpus callosum severed). So if there are (at least) two persons inhabiting the body of the split brain patient, the same is true of us.

But this is absurd. What these cases show, according to Parfit, is that our concept of a person is inherently unstable; any way of "counting persons" in these cases leads to crazy results. We should conclude that all there really are are experiences with certain connections between them. Our talk about persons is just a convenient way of grouping these experiences together, but doesn't really correspond to anything in reality. (Compare this to Parfit's example of clubs.)

How should someone who does not want to take this view of persons respond?