# The psychological theory
# &
# the dissolving self

Last time, in response to Reid's objection to Locke's version of the memory theory of persons, we ended up with a view of personal identity which included the following two claims:

**The modified memory requirement**

If there is no memory relation between A and B, whether direct or indirect, then A≠B.

**The memory guarantee**

If A does remember an experience of B, then A=B.

(Where, as last time, what it means to say that A has an indirect memory relation to B is that there is some series of persons of which A and B are both members which is such that every member of the series has memories of the preceding member of the series.)

Today we will be discussing arguments against both the modified memory requirement and the memory guarantee.

**The modified memory requirement**

If there is no memory relation between A and B, whether direct or indirect, then A≠B.

**The memory guarantee**

If A does remember an experience of B, then A=B.

Let's first discuss the example of the mad torturer, which is due to Bernard Williams.

Williams noticed that one important fact about our concept of personal identity is that there is a sense in which **one can only anticipate, or fear, experiences which are going to happen to you**. In something like the sense in which I cannot feel your toothache, it seems that there is a clear sense in which I cannot fear something painful which is going to happen to someone else - at least, the relevant sense of "fear" seems quite different.

Williams used this fact to argue against the memory theory of persons and, in particular, against the memory requirement.

**The memory guarantee**

If A does remember an experience of B, then A=B.

**The modified memory requirement**

If there is no memory relation between A and B, whether direct or indirect, then A≠B.

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.

It is hard not to agree with Williams here: fear, surely, **would** still be in the proper reaction.

**The modified memory requirement**

If there is no memory relation between A and B, whether direct or indirect, then A≠B.

**The memory guarantee**

If A does remember an experience of B, then A=B.

It is hard not to agree with Williams here: fear, surely, **would** still be in the proper reaction.

But if the memory requirement were correct, it seems that fear would **not** be the proper reaction -- at least, not fear of torture. For, if the memory requirement were correct, that would not be you being tortured, but someone else. The right reaction would be something like **pity**; we would be inclined to say, “That poor person; I hope that it will be over with quickly for them.” But this is not how we would react to the situation, it seems.

In fact, if the memory requirement were correct, the most troubling part of the story from a self-interested point of view would be the part at which the torturer announces that your memories will be erased; for that, according to that theory, would be your death.

But if we would be right to fear the torture - and if it is true that one can only have this attitude of fear towards one’s own future experiences - then this seems to show that we don’t really think of ourselves in the way that the memory theory says we should.

Can the memory theorist reasonably reply that while we would fear the torture, that this is in fact irrational, and based upon a confused idea of our own nature?

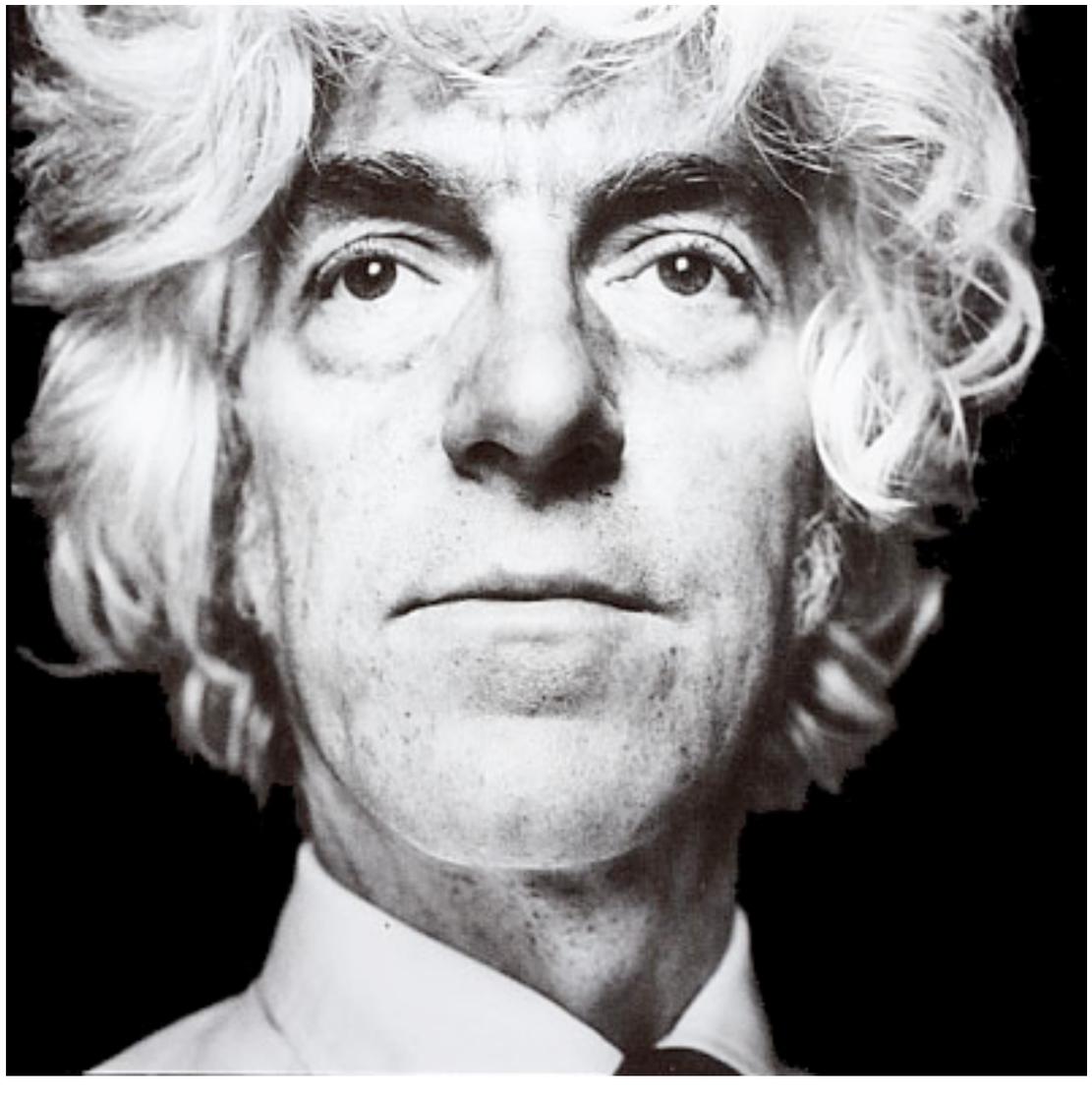
**The modified memory requirement**

If there is no memory relation between A and B, whether direct or indirect, then A≠B.

**The memory guarantee**

If A does remember an experience of B, then A=B.

A second sort of problematic case for the memory theory focuses not on the memory requirement, but the memory guarantee. This is Parfit's example of the teletransporter.



**The modified memory requirement**

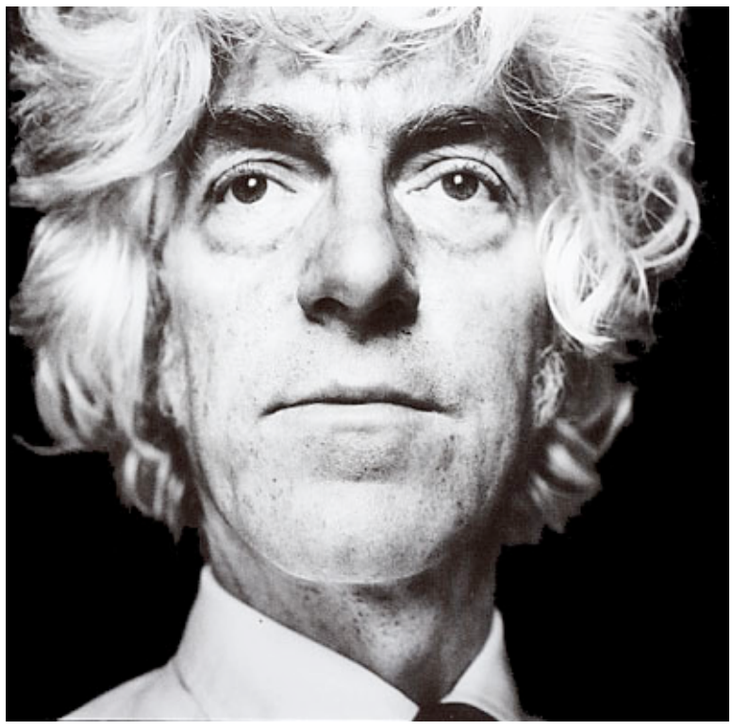
If there is no memory relation between A and B, whether direct or indirect, then A≠B.

**The memory guarantee**

If A does remember an experience of B, then A=B.

A second sort of problematic case for the memory theory focuses not on the memory requirement, but the memory guarantee. This is Parfit's example of the teletransporter.

The initial version of the journey by teletransportation to Mars seems relatively unproblematic, even if currently technologically impossible.



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.

**The modified memory requirement**

If there is no memory relation between A and B, whether direct or indirect, then A≠B.

**The memory guarantee**

If A does remember an experience of B, then A=B.

The problems begin with the arrival of the New Scanner.

The problems posed by this case are closely related to the problems posed by Reid's example. As in the case of Reid's argument, it will be useful to introduce some terms so that we can talk about this case clearly.

Original-Parfit = Parfit before he stepped into the teletransporter.

Earth-Parfit = the person who gets out of the teletransporter on earth.

Mars-Parfit = the person who gets out of the teletransporter on Mars.

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'.

Someone politely coughs, a white-coated man who asks to speak to me in private. We go to his office, where he tells me to sit down, and pauses. Then he says: 'I'm afraid that we're having problems with the New Scanner. It records your blueprint just as accurately, as you will see when you talk to yourself on Mars. But it seems to be damaging the cardiac systems which it scans. Judging from the results so far, though you will be quite healthy on Mars, here on Earth you must expect cardiac failure within the next few days.'

The attendant later calls me to the Intercom. On the screen I see myself just as I do in the mirror every morning. But there are two differences. On the screen I am not left-right reversed. And, while I stand here speechless, I can see and hear myself, in the studio on Mars, starting to speak.

**The modified memory requirement**

If there is no memory relation between A and B, whether direct or indirect, then A≠B.

**The memory guarantee**

If A does remember an experience of B, then A=B.

Original-Parfit = Parfit before he stepped into the teletransporter.

Earth-Parfit = the person who gets out of the teletransporter on earth.

Mars-Parfit = the person who gets out of the teletransporter on Mars.

The character in the story seems to be correct when he says “If I’m here I can’t also be on Mars.” But that is just another way of saying this:

**Earth-Parfit ≠ Mars-Parfit**

The problem is that both Earth-Parfit and Mars-Parfit stand in direct memory relations to Original-Parfit. Hence, if the memory guarantee is true, we know that each of the following must be true.

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’.

Someone politely coughs, a white-coated man who asks to speak to me in private. We go to his office, where he tells me to sit down, and pauses. Then he says: ‘I’m afraid that we’re having problems with the New Scanner. It records your blueprint just as accurately, as you will see when you talk to yourself on Mars. But it seems to be damaging the cardiac systems which it scans. Judging from the results so far, though you will be quite healthy on Mars, here on Earth you must expect cardiac failure within the next few days.’

The attendant later calls me to the Intercom. On the screen I see myself just as I do in the mirror every morning. But there are two differences. On the screen I am not left-right reversed. And, while I stand here speechless, I can see and hear myself, in the studio on Mars, starting to speak.

**The modified memory requirement**

If there is no memory relation between A and B, whether direct or indirect, then A≠B.

**The memory guarantee**

If A does remember an experience of B, then A=B.

The character in the story seems to be correct when he says “If I’m here I can’t also be on Mars.” But that is just another way of saying this:

**Earth-Parfit ≠ Mars-Parfit**

The problem is that both Earth-Parfit and Mars-Parfit stand in direct memory relations to Original-Parfit. Hence, if the memory guarantee is true, we know that each of the following must be true.

**Earth-Parfit = Original-Parfit**

**Mars-Parfit = Original-Parfit**

But, for reasons which are by now familiar, the three claims in blue cannot all be true: this trio of claims is inconsistent. So, if the scenario Parfit describes is really possible, it looks as though the memory guarantee implies a contradiction. But then the memory guarantee must be false.

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’.

Someone politely coughs, a white-coated man who asks to speak to me in private. We go to his office, where he tells me to sit down, and pauses. Then he says: ‘I’m afraid that we’re having problems with the New Scanner. It records your blueprint just as accurately, as you will see when you talk to yourself on Mars. But it seems to be damaging the cardiac systems which it scans. Judging from the results so far, though you will be quite healthy on Mars, here on Earth you must expect cardiac failure within the next few days.’

The attendant later calls me to the Intercom. On the screen I see myself just as I do in the mirror every morning. But there are two differences. On the screen I am not left-right reversed. And, while I stand here speechless, I can see and hear myself, in the studio on Mars, starting to speak.

**The modified memory requirement**

If there is no memory relation between A and B, whether direct or indirect, then A≠B.

**The memory guarantee**

If A does remember an experience of B, then A=B.

**Earth-Parfit ≠ Mars-Parfit**

**Earth-Parfit = Original-Parfit**

**Mars-Parfit = Original-Parfit**

But, for reasons which are by now familiar, the three claims in blue cannot all be true: this trio of claims is inconsistent. So, if the scenario Parfit describes is really possible, it looks as though the memory guarantee implies a contradiction. But then the memory guarantee must be false.

At this stage, the memory theory might seem to be in pretty bad shape. The two parts of that theory are the memory guarantee and the memory requirement, and we have now some pretty convincing arguments against each of these claims. The example of torture seems to show that the memory requirement is false, and the example of the teletransporter seems to show that the memory guarantee is false.

Parfit suggests a radical response to these problems. According to Parfit, when we talk about “personal identity” or “being the same person”, we aren’t really talking about an all-or-nothing thing. Rather, we are just talking about degrees of psychological similarity. So when I say that A and B are the same person, what I really mean is just: A and B are psychologically connected in certain interesting ways.

One useful comparison (which Parfit suggests elsewhere) is a comparison of persons to clubs, or teams. Suppose that we begin a personal identity discussion club at Notre Dame. People gradually leave and join the club, and some of the rules change, and eventually people decide that at meetings things other than personal identity may occasionally be discussed. At one of the meetings (in 2048) someone says: “Is this really the **same club** as the one formed way back in 2009?”

At this stage, the memory theory might seem to be in pretty bad shape. The two parts of that theory are the memory guarantee and the memory requirement, and we have now some pretty convincing arguments against each of these claims.

Parfit suggests a radical response to these problems. According to Parfit, when we talk about “personal identity” or “being the same person”, we aren’t really talking about an all-or-nothing thing. Rather, we are just talking about degrees of psychological similarity. So when I say that A and B are the same person, what I really mean is just: A and B are psychologically connected in certain interesting ways.

One useful comparison (which Parfit suggests elsewhere) is a comparison of persons to clubs, or teams. Suppose that we begin a personal identity discussion club at Notre Dame. People gradually leave and join the club, and some of the rules change, and eventually people decide that at meetings things other than personal identity may occasionally be discussed. At one of the meetings (in 2048) someone says: “Is this really the **same club** as the one formed way back in 2009?”

Parfit suggests, and this seems right, that this is not a very deep question. The club in 2048 is similar in some ways to our club, and different in other ways; there is no **further fact** about whether the two clubs are **really the same.** We could decide to say that they are identical or distinct, but our choice seems somewhat arbitrary.

Parfit’s radical suggestion is that people are, in this way, like clubs. When we ask, “Is Original-Parfit really the same person as Mars-Parfit, or Earth-Parfit?” we are not asking a very deep question. Each is similar in certain important ways to Original-Parfit, and that is pretty much the end of the story. There is simply no further, fundamental fact about which one is identical to Original-Parfit.

This view has some surprising consequences. One is that questions about death and survival also do not have all-or-nothing answers. Think about Earth-Parfit after he comes out of the New Scanner. One naturally thinks that he should be very upset about the fact that he is going to die soon. But, if Parfit is right, he should be much consoled by the fact that Mars-Parfit, who is psychologically extremely similar to him, will continue to live - after all, ordinary survival just is a matter of there being someone psychologically quite similar to me who continues to exist. (Compare the survival of a club.)

One useful comparison (which Parfit suggests elsewhere) is a comparison of persons to clubs, or teams. Suppose that we begin a personal identity discussion club at Notre Dame. People gradually leave and join the club, and some of the rules change, and eventually people decide that at meetings things other than personal identity may occasionally be discussed. At one of the meetings (in 2048) someone says: “Is this really the **same club** as the one formed way back in 2009?”

Parfit suggests, and this seems right, that this is not a very deep question. The club in 2048 is similar in some ways to our club, and different in other ways; there is no **further fact** about whether the two clubs are **really the same.** We could decide to say that they are identical or distinct, but our choice seems somewhat arbitrary.

Parfit’s radical suggestion is that people are, in this way, like clubs. When we ask, “Is Original-Parfit really the same person as Mars-Parfit, or Earth-Parfit?” we are not asking a very deep question. Each is similar in certain important ways to Original-Parfit, and that is pretty much the end of the story. There is simply no further, fundamental fact about which one is identical to Original-Parfit.

This view has some surprising consequences. One is that questions about death and survival also do not have all-or-nothing answers. Think about Earth-Parfit after he comes out of the New Scanner. One naturally thinks that he should be very upset about the fact that he is going to die soon. But, if Parfit is right, he should be much consoled by the fact that Mars-Parfit, who is psychologically extremely similar to him, will continue to live -- after all, ordinary survival just is a matter of there being someone psychologically quite similar to me who continues to exist. (Compare the survival of a club.)

One might think that Earth-Parfit could protest:

> **“But Mars-Parfit isn’t me! Why should I feel better about dying because this other guy will live!”**

But if Parfit is right, this is just confused. Mars-Parfit sort of is Earth-Parfit -- they are psychologically similar in important ways and, if Parfit is right, that is all there is to personal identity. **If Parfit is right, Earth-Parfit has nothing to complain about.**

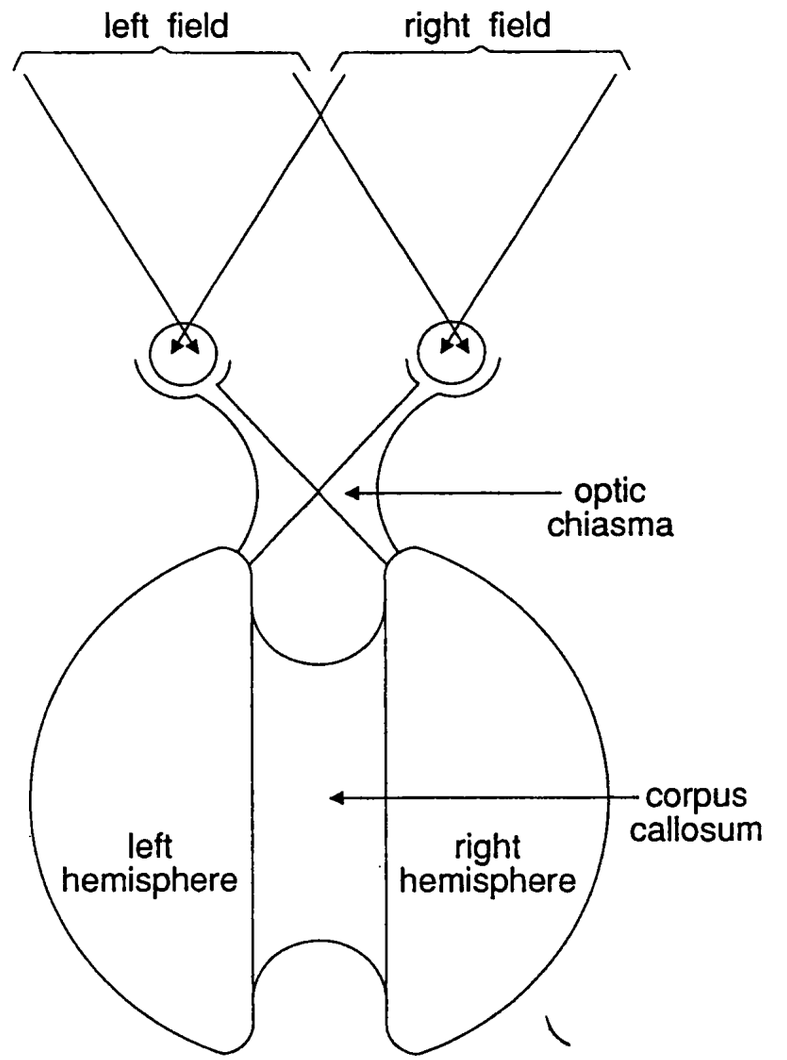
Many people find this particular consequence, and Parfit’s denial of the reality of facts about personal identity in general, extremely difficult to believe. But it does find some support in some psychological research done over the last few decades.

Many people find this particular consequence, and Parfit's denial of the reality of facts about personal identity in general, extremely difficult to believe. But it does find some support in some psychological research done over the last few decades.

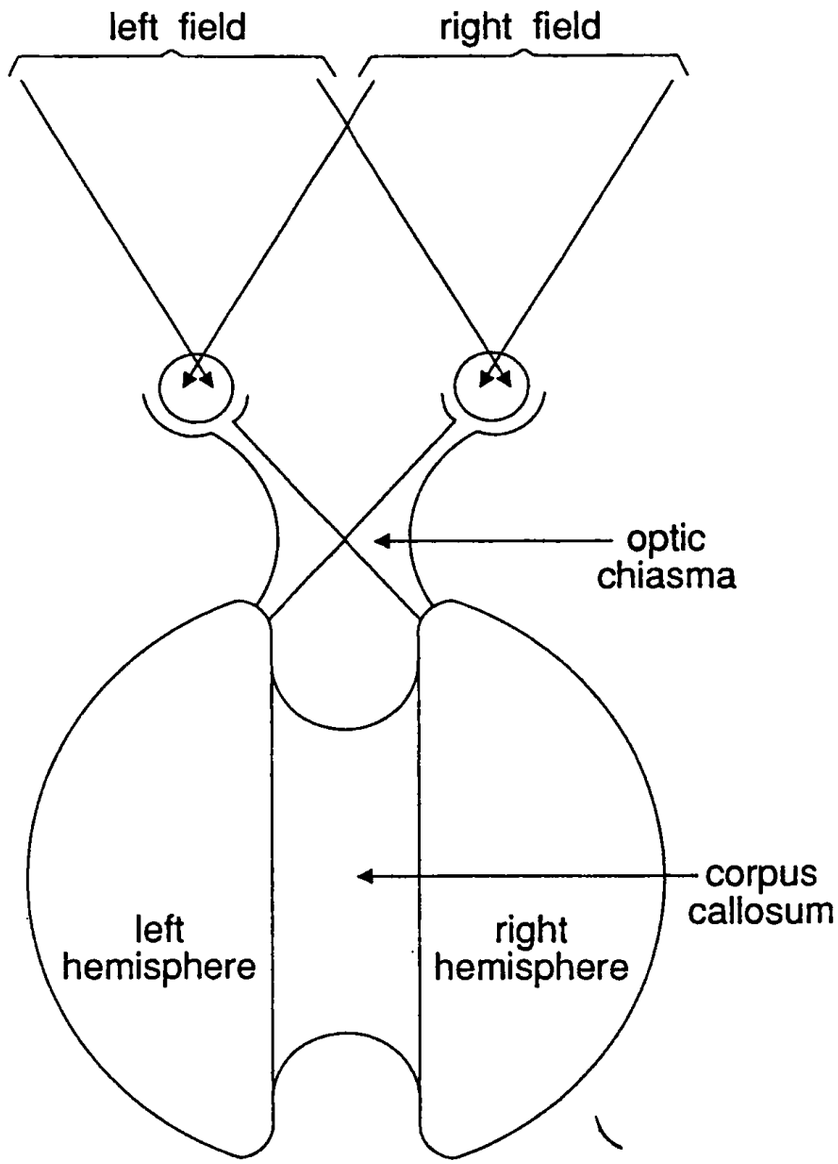
These are studies of patients whose corpus callosum has been severed. The corpus callosum is a pathway which connects the left and right hemispheres of the human brain and, in normal subjects, allows the two hemispheres of the brain to exchange information.

If the corpus callosum is severed, the two hemispheres of the brain cannot exchange information. So any sensory data about the environment available to, for example, the left hemisphere, will not be available to guide the movements of the left hand, which is controlled by the right hemisphere. Information available only to the right hemisphere will not be reportable in speech, since speech is controlled by the left hemisphere.

The results of giving sensory data to just one of the hemispheres of the brain of such a patient are striking.

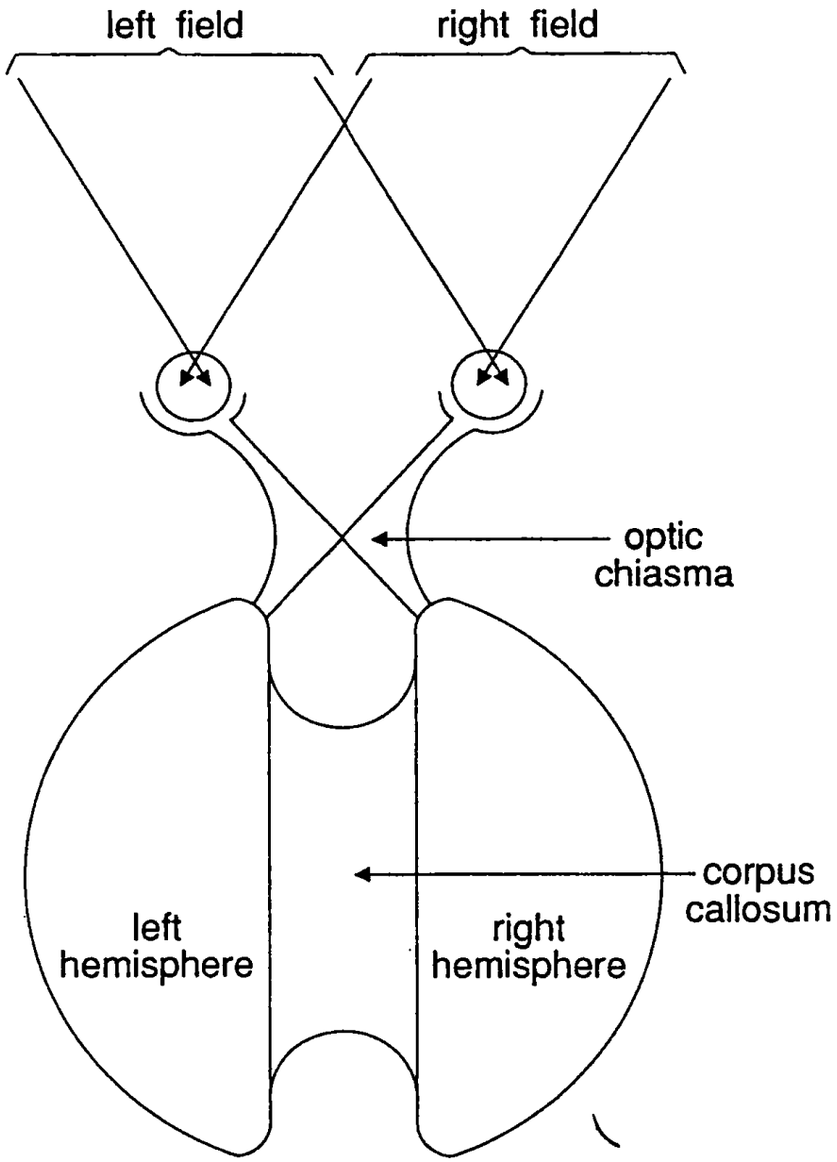


The results of giving sensory data to just one of the hemispheres of the brain of such a patient are striking.




The results are as follows. What is flashed to the right half of the visual field, or felt unseen by the right hand, can be reported verbally. What is flashed to the left half field or felt by the left hand cannot be reported, though if the word 'hat' is flashed on the left, the left hand will retrieve a hat from a group of concealed objects if the person is told to pick out what he has seen. At the same time he will insist verbally that he saw nothing. Or, if two different words are flashed to the two half fields (e.g. 'pencil' and 'toothbrush') and the individual is told to retrieve the corresponding object from beneath a screen, with both hands, then the hands will search the collection of objects independently, the right hand picking up the pencil and discarding it while the left hand searches for it, and the left hand similarly rejecting the toothbrush which the right had lights upon with satisfaction.

The results of giving sensory data to just one of the hemispheres of the brain of such a patient are striking.




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?

The following two principles seem quite plausible (especially if, like memory theorists, we think that the nature of persons is tied closely to consciousness):

**Ownership**

Every conscious experience must be an experience of someone.

**Awareness**

If someone has a conscious experience, it must be at least in principle possible for them to be aware of that experience.

Now think about a case in which a split-brain patient has a red stimulus presented to the right half of their visual field, and a blue stimulus presented to the left half of their visual field. If you ask the subject what color they see, they will say “Red”, since this was the color presented to the part of the eye which feeds input to the left hemisphere of the brain, which controls speech.

So it is clear that there is a conscious experience of red; so, by **Ownership**, there must be someone who is having this experience. Let’s call this person “Mr. Red.”

If you put a pen in the left hand of the left hand of the subject, and ask what color was just seen, that hand will write “Blue.” So it seems that there must have been a conscious experience of blue - otherwise, how would the hand know what color to write?

But if there is a conscious experience of blue, by **Ownership** someone must have had this experience. Let us call the person who has this experience “Mr. Blue.”

**Ownership**

Every conscious experience must be an experience of someone.

**Awareness**

If someone has a conscious experience, it must be at least in principle possible for them to be aware of that experience.

So it is clear that there is a conscious experience of red; so, by **Ownership**, there must be someone who is having this experience. Let's call this person "Mr. Red."

But if there is a conscious experience of blue, by **Ownership** someone must have had this experience. Let us call the person who has this experience "Mr. Blue."

Now the crucial question is: Is Mr. Red the same person as Mr. Blue? It seems to follow from **Awareness** that they are not the same person. After all, if you ask Mr. Red whether he has had any experience of blue, he will say "No." And no amount of introspection on his part will allow him to remember having a conscious experience of this sort; and of course this is not because he forgot having the experience, but because he was never aware of having it. But then, by **Awareness**, he *didn't* have it.

Hence it seems that Mr. Red ≠ Mr. Blue. So there are two persons in the body of the split brain patient.

This is a bit weird on its own. But further oddities result from consideration of what this conclusion says about non-split-brain patients, like us.

There seem to be three things we can say:

1. While the split brain patients are in experiments of this sort, there are two persons inhabiting their body; but, at other times, there is just one person inhabiting their body.

2. Split brain patients always have two persons inhabiting their body, but non-split brain subjects do not.

3. All of us, split-brain and non-split-brain subjects alike, have two (or more) persons inhabiting their body.

There seem to be three things we can say:

But each of these options seems, for various reasons, absurd.

1. While the split brain patients are in experiments of this sort, there are two persons inhabiting their body; but, at other times, there is just one person inhabiting their body.

If this were true, then simply flashing some red and blue lights at someone would bring a new person into existence; and turning off the lights would kill that person.

2. Split brain patients always have two persons inhabiting their body, but non-split brain subjects do not.

If this were true, then severing the corpus callosum of an epileptic patient would bring a new person into existence; and reversing the surgery would kill that person.

3. All of us, split-brain and non-split-brain subjects alike, have two (or more) persons inhabiting their body.

Non-split brain patients never have conscious experiences of which they are not aware; but then it would follow that there is a person inhabiting my body which never has any conscious experiences at all. But then in what sense does that person even exist?

There remains a fourth option: our talk about persons, or subjects of experience, is just a convenient fiction for talking about conscious experiences. The split-brain cases illustrate that there are cases in which this convenient fiction breaks down; in cases like the one described above, there is a red experience and a blue experience, and that is all that we can say; there is no further fact about whether these experiences are experiences of the same person, or not. But this is, of course, the radical view we ascribed to Parfit.