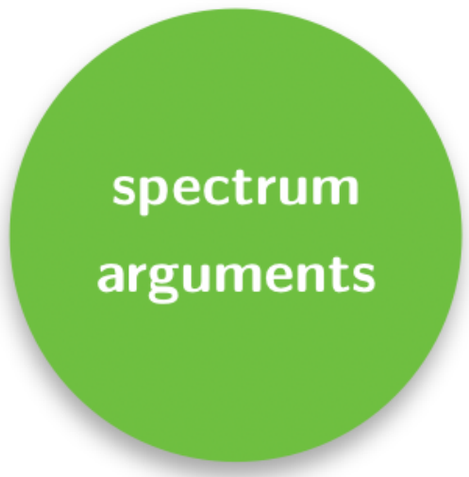


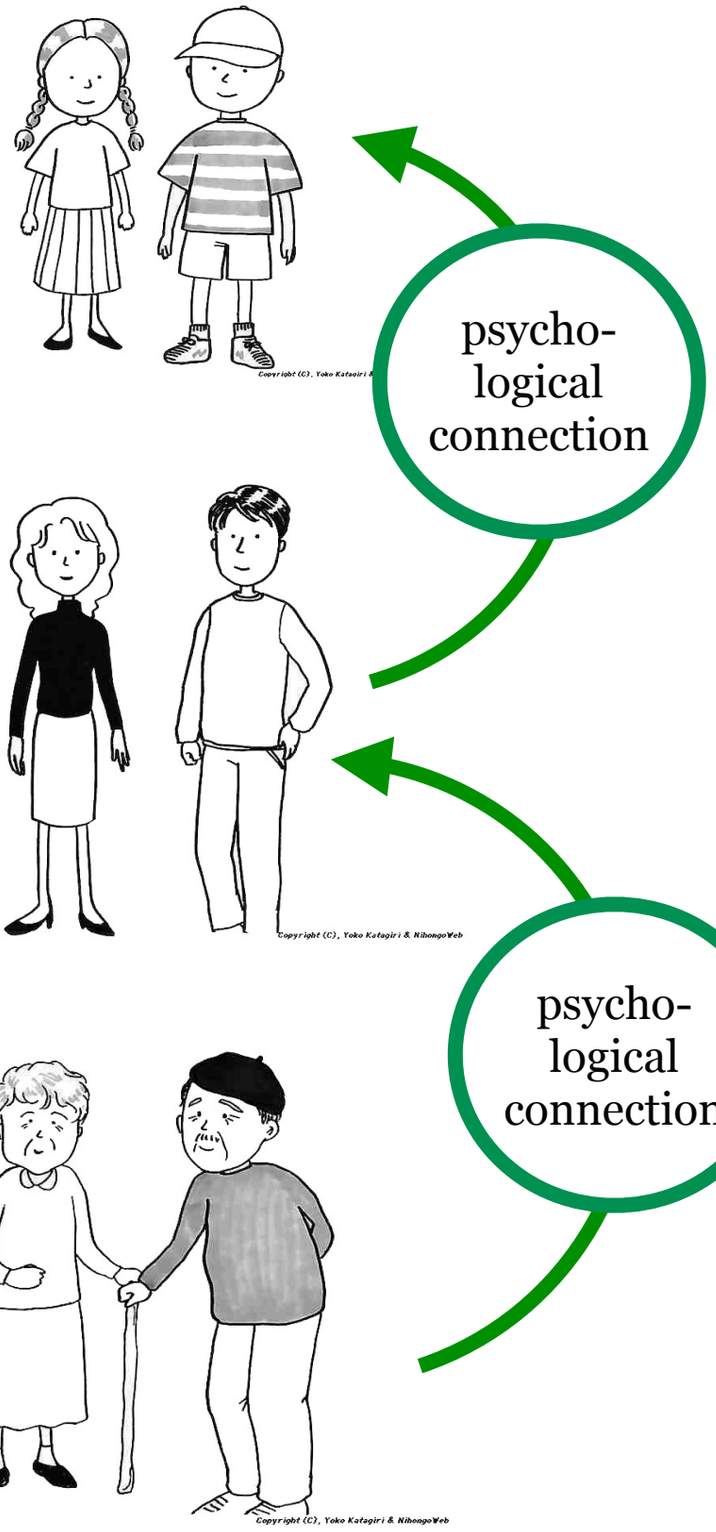
Psychological continuity, spectrum arguments, & the possibility of uploading

The survival question: What does it take for for some person at some other time to be you?

One of our answers to the survival question focuses on **psychological** relations.

Psychological survival: X is me just in case X has the right kinds of psychological connections to me

This was John Locke's theory. His view of personhood can be illustrated by considering a few different stages in the lives of some people.



This was John Locke's theory. His view of personhood can be illustrated by considering a few different stages in the lives of some people.

What makes the child, the adult, and the elderly person stages of the same person? The materialist says: because they are the same material thing. Locke thought: it is because of psychological connections between the individuals.

psychological continuity

spectrum arguments

AI & uploading

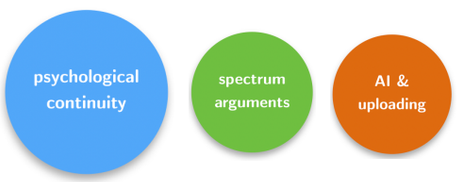
This view has some notable advantages over materialist answers to the survival question.

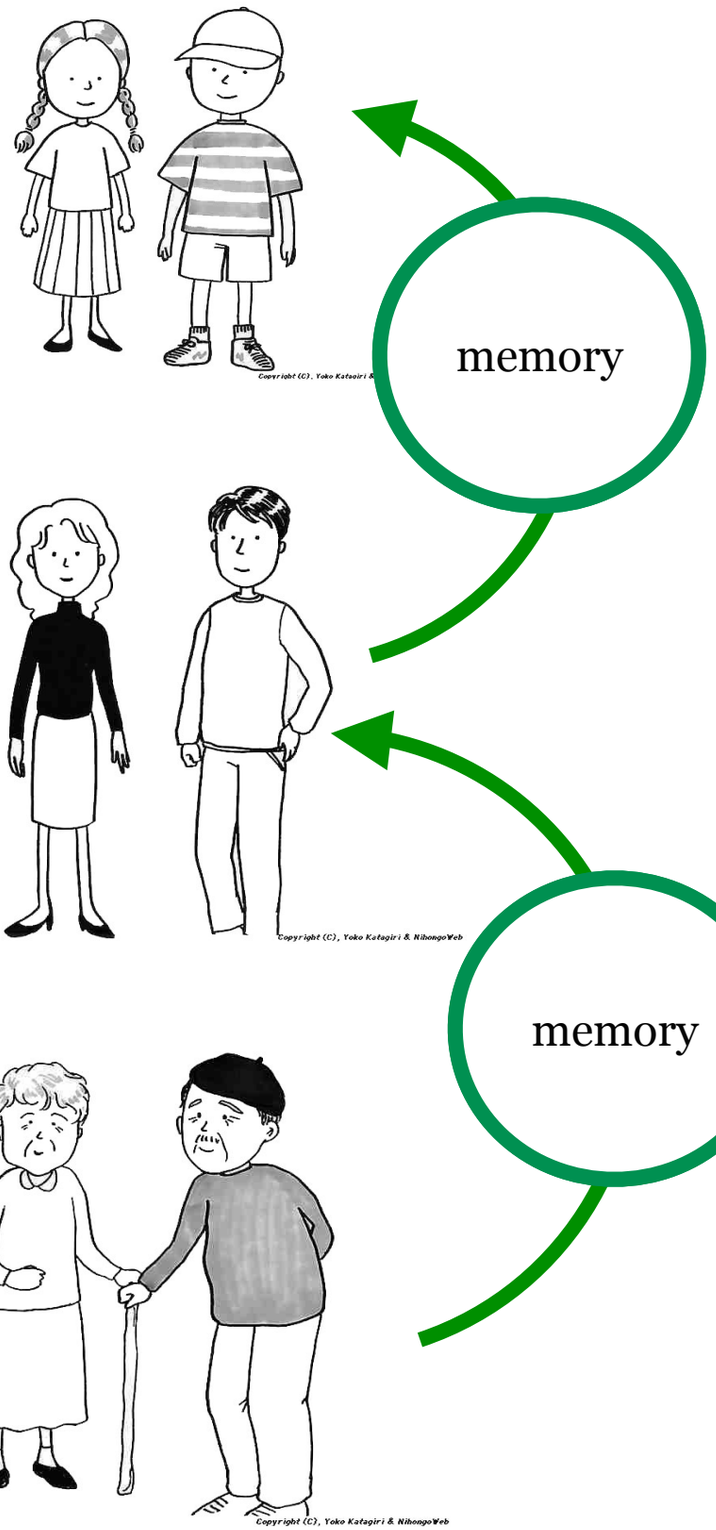
It captures the intuition that you could wake up in a different body than the one you now have.

It avoids the materialist's problem with explaining how you could be the same material thing despite changes in your parts over time. No more fears that haircuts might be the end of you!

It seems to avoid the materialist's problems with making sense of the possibility of life after death. For surely God could create a being which stood in all of the right psychological connections to you, even if your body has decayed.

It also delivers these results without bringing immaterial souls into the picture. So it avoids the arguments — like the swapping arguments and the interaction argument — which make trouble for the dualist.





psychological continuity

spectrum arguments

AI & uploading

But what are the relevant psychological relations?

Locke’s answer was: relations of memory.

But, as Locke’s contemporary Thomas Reid noted, this answer leads to an immediate problem.

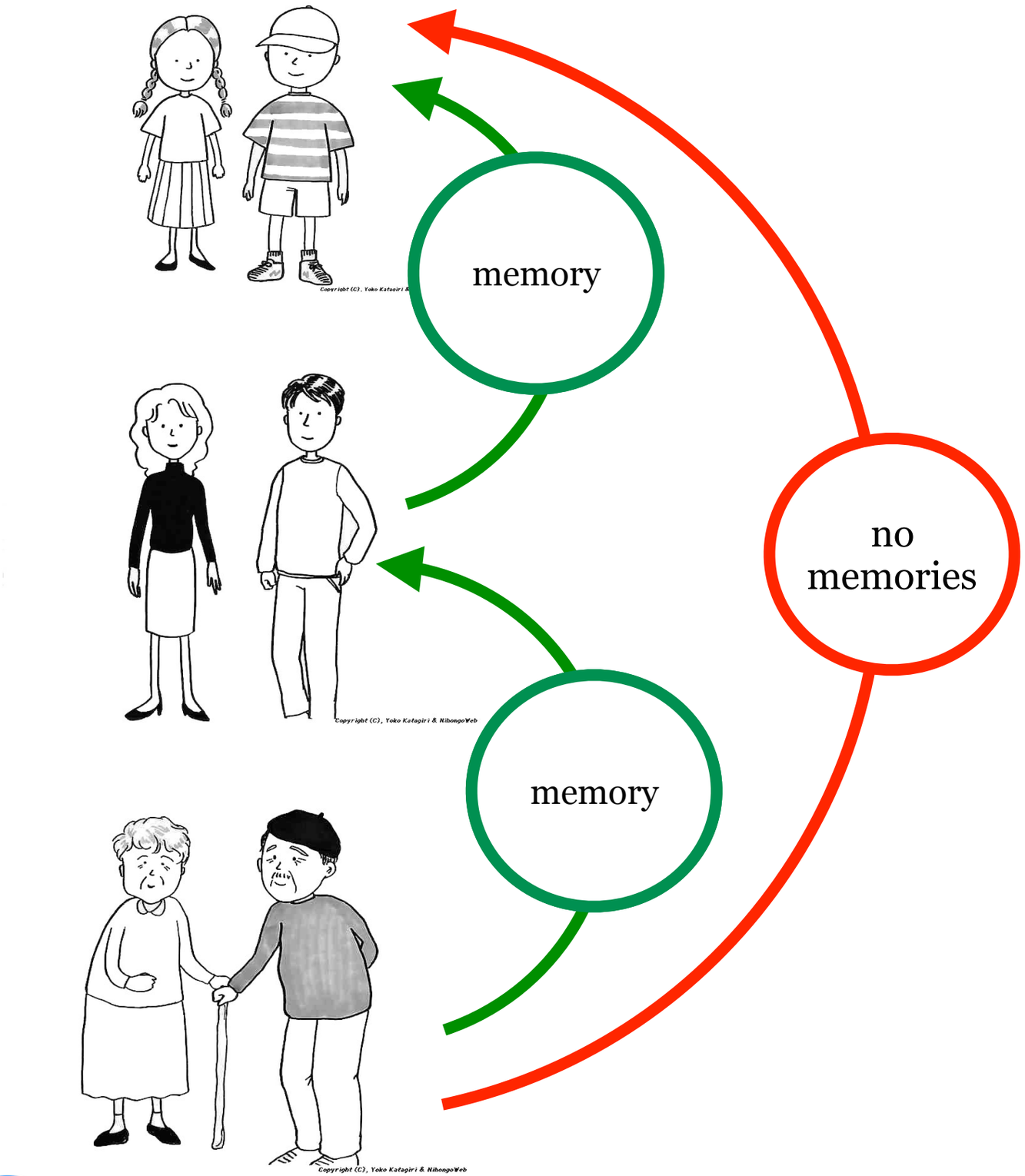
“Suppose a brave officer to have been flogged when a boy at school for robbing an orchard, to have taken a standard in his first campaign, and to have been made a general in advanced life. Suppose also, which must be admitted to be possible, that when he took the standard he was conscious of his having been flogged at school, and that when made a general he was conscious of his taking the standard, but had absolutely lost the consicousness of the flogging.

Therefore [if the psychological theory of survival is true] the general is, and at the same time is not the same person with him who was flogged at school.”

psychological continuity

spectrum arguments

AI & uploading



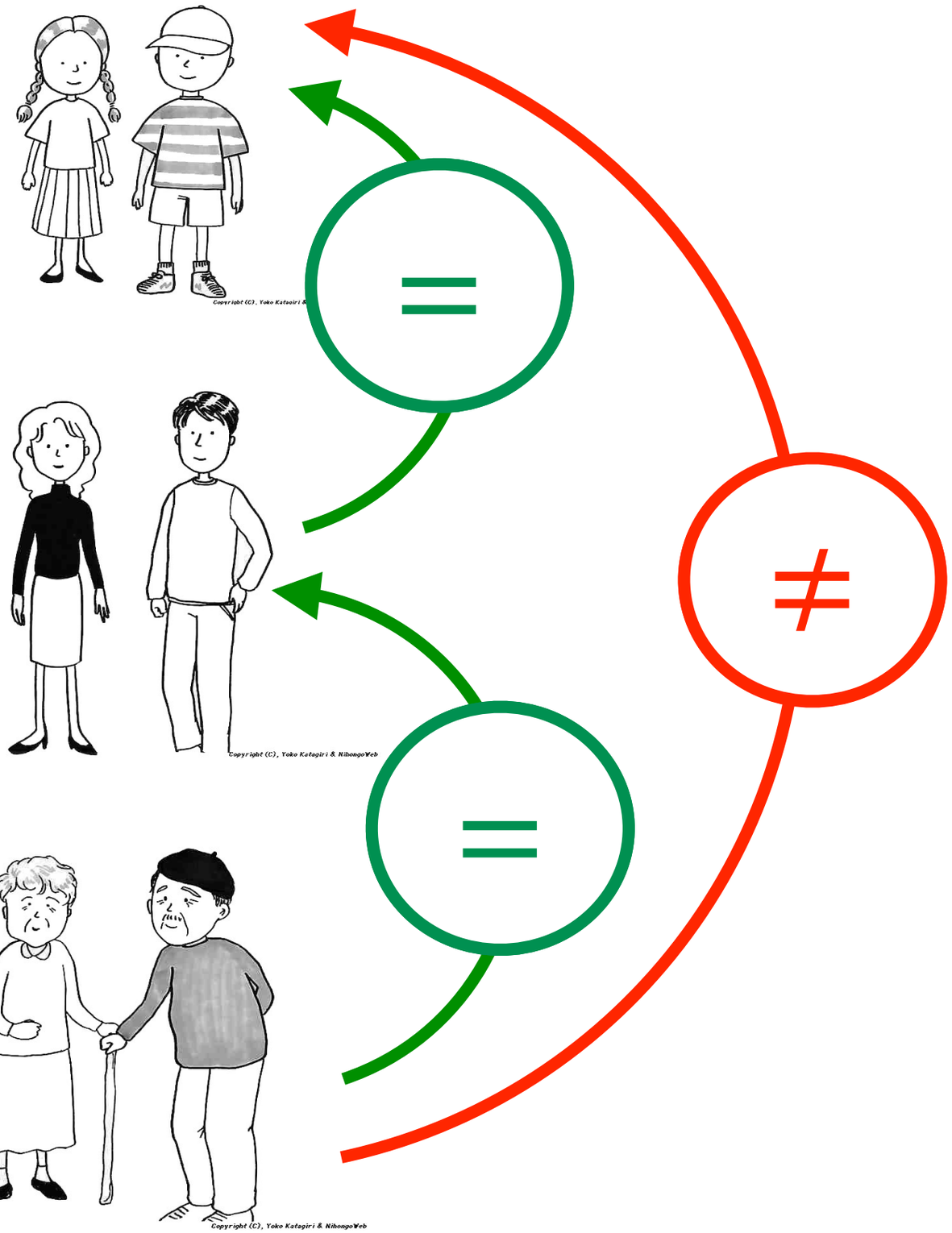
We can illustrate the kind of scenario that Reid had in mind.

This is problematic because, if identity of persons is determined by memory or its absence, Reid's example leads to an impossible constellation of identity facts.

psychological continuity

spectrum arguments

AI & uploading



How should the psychological theorist reply?

One promising reply: introduce the notion of an indirect memory relation, which is related to memory relations in the way that ancestor is related to parent.

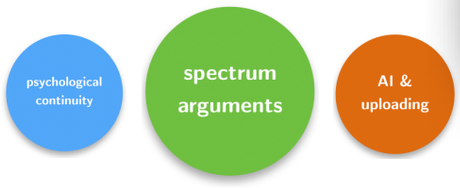
psychological continuity

spectrum arguments

AI & uploading

A more serious challenge to psychological survival is posed by spectrum arguments. We can introduce the spectrum arguments via a thought experiment Derek Parfit discusses in the reading.

> “I am the prisoner of some callous neuro-surgeon, who intends to disrupt my psychological continuity by tampering with my brain. I shall be conscious while he operates, and in pain. I therefore dread what is coming. The surgeon tells me that, while I am in pain, he will activate some neurodes that will give me amnesia. I shall suddenly lose all of my memories of my life up to the start of my pain. Does this give me less reason to dread what is coming? Surely not.
>
> The surgeon next tells me that, while I am still in pain, he will later flip another switch, that will cause me to believe that I am Napoleon, and will give me apparent memories of Napoleon’s life. I would have no reason to expect this to cause my pain to cease.
>
> The surgeon then tells me that, during my ordeal, he will later flip a third switch, that will change my character so that it becomes just like Napoleon’s. Once again, I seem to have no reason to expect the flipping of this switch to end my pain. It might at most bring some relief, if Napoleon’s character, compared with mine, involved more fortitude.”



“I am the prisoner of some callous neuro-surgeon, who intends to disrupt my psychological continuity by tampering with my brain. I shall be conscious while he operates, and in pain. I therefore dread what is coming. The surgeon tells me that, while I am in pain, he will activate some neurodes that will give me amnesia. I shall suddenly lose all of my memories of my life up to the start of my pain. Does this give me less reason to dread what is coming? Surely not.

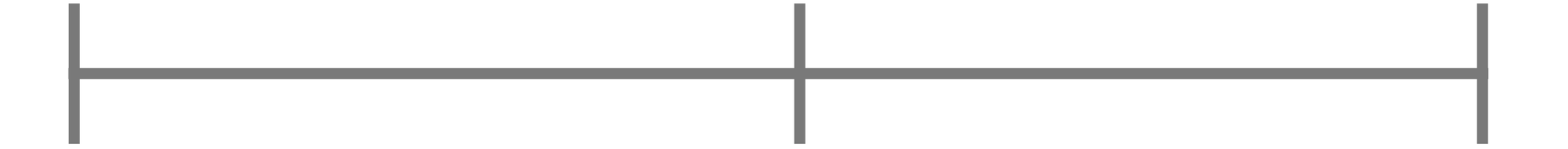
The surgeon next tells me that, while I am still in pain, he will later flip another switch, that will cause me to believe that I am Napoleon, and will give me apparent memories of Napoleon’s life. I would have no reason to expect this to cause my pain to cease.

The surgeon then tells me that, during my ordeal, he will later flip a third switch, that will change my character so that it becomes just like Napoleon’s. Once again, I seem to have no reason to expect the flipping of this switch to end my pain. It might at most bring some relief, if Napoleon’s character, compared with mine, involved more fortitude.”

Intuitively, at the end of this series of unfortunate events, you would still be in pain. But what must the psychological theorist say about this case?

As Parfit says, we can think of these kinds of psychological changes as falling on a spectrum.

the psychological spectrum

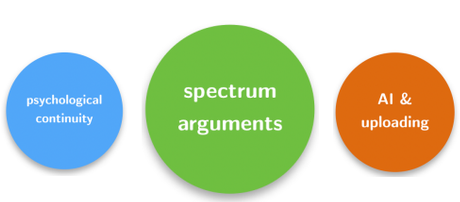


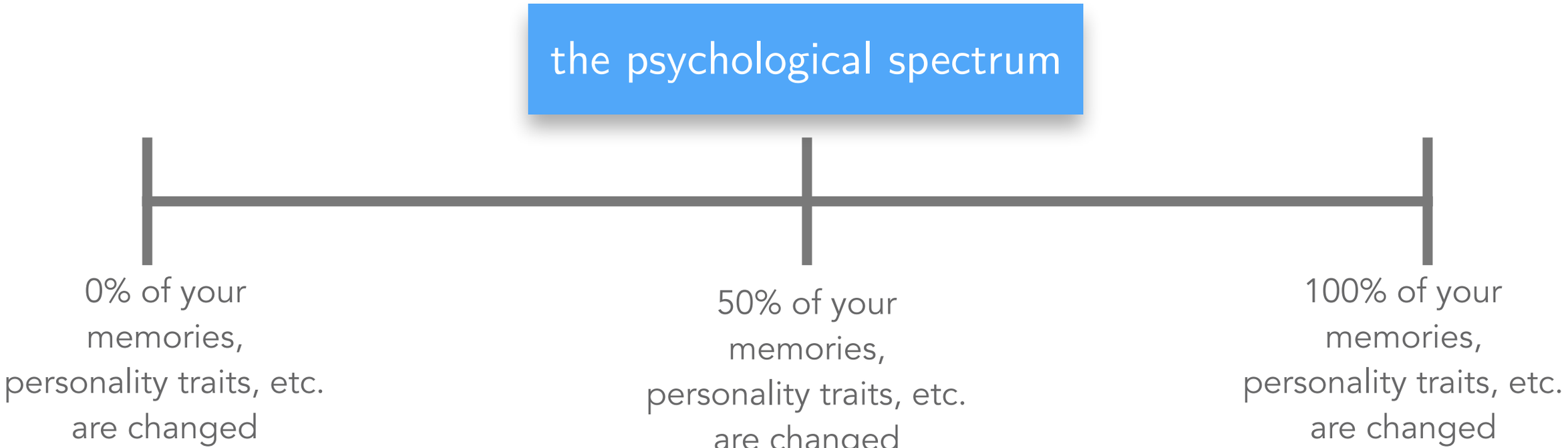
0% of your memories, personality traits, etc. are changed

50% of your memories, personality traits, etc. are changed

100% of your memories, personality traits, etc. are changed

The intuitive response to the torture example implies that you survive even on the far right edge of the spectrum. This seems to show that the psychological theory of survival is false.





The intuitive response to the torture example implies that you survive even on the far right edge of the spectrum. This seems to show that the psychological theory of survival is false.

Let's think about how someone who endorses psychological survival might respond. To do that, it will help to look at a different kind of example.

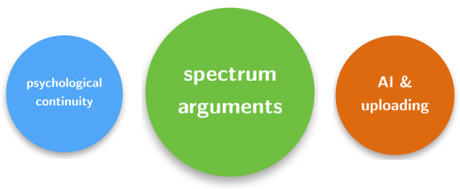
Let's think about how someone who endorses psychological survival might respond. To do that, it will help to look at a different kind of example.

Suppose that I am an impoverished philosophy professor, and definitely not rich.

Now suppose that a wealthy benefactor who loves philosophy decides to give me some money. But he does this in an eccentric way: by adding 1 cent to my bank account every second.

At the end of 10 years, I will have $3.1 million in my bank account, and will be rich.

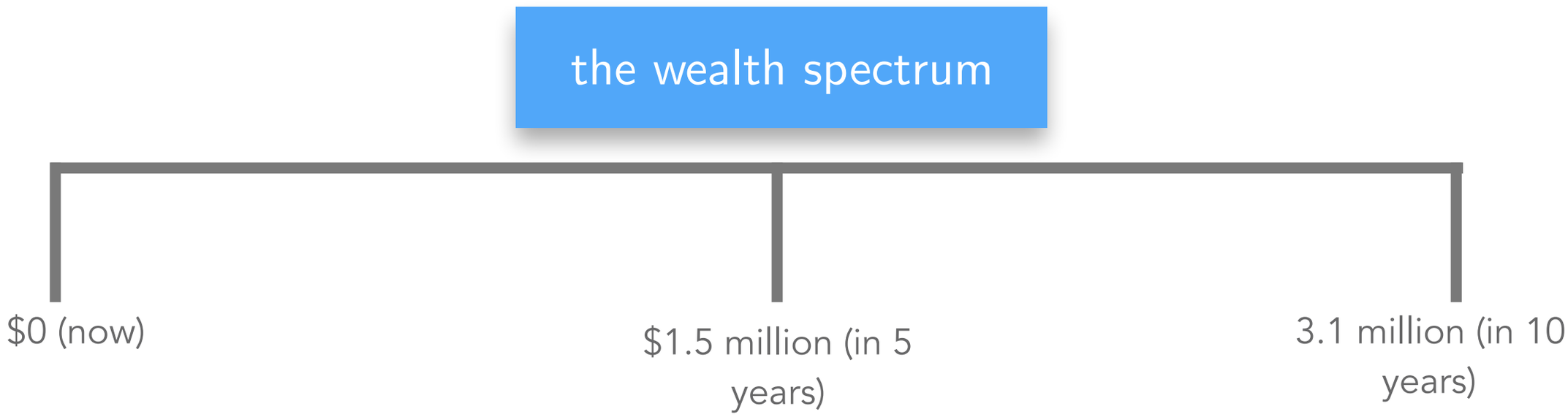
We can chart my progress using the wealth spectrum.



Now suppose that a wealthy benefactor who loves philosophy decides to give me some money. But he does this in an eccentric way: by adding 1 cent to my bank account every second.

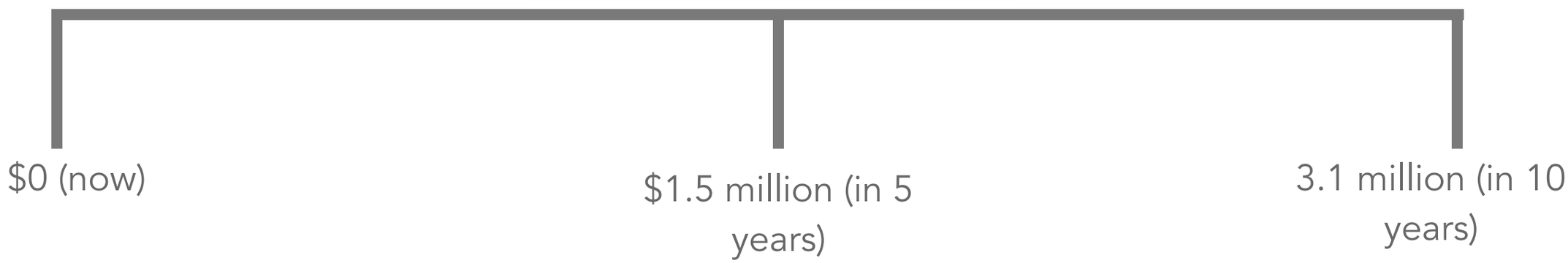
At the end of 10 years, I will have $3.1 million in my bank account, and will be rich.

We can chart my progress using the wealth spectrum.



When did I become rich?

the wealth spectrum



When did I become rich?

It appears that there are exactly three ways to answer this question.

Sharp Cut Off

There is a precise point in the spectrum at which I switched from being non-rich to being rich.

Indeterminacy

At the beginning I was non-rich; at the end I am rich; but there is no sharp cut off. Instead, there is a range of cases in which it is not determinately true either that I am rich or that I am non-rich.

Never rich

Even at the end of the spectrum, I am still not rich.

the wealth spectrum

$0 (now)

$1.5 million (in 5 years)

3.1 million (in 10 years)

Sharp Cut Off

There is a precise point in the spectrum at which I switched from being non-rich to being rich.

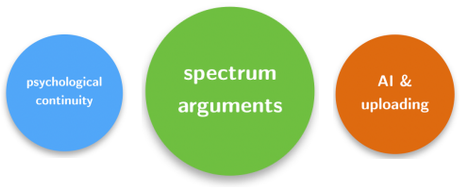
Indeterminacy

At the beginning I was non-rich; at the end I am rich; but there is no sharp cut off. Instead, there is a range of cases in which it is not determinately true either that I am rich or that I am non-rich.

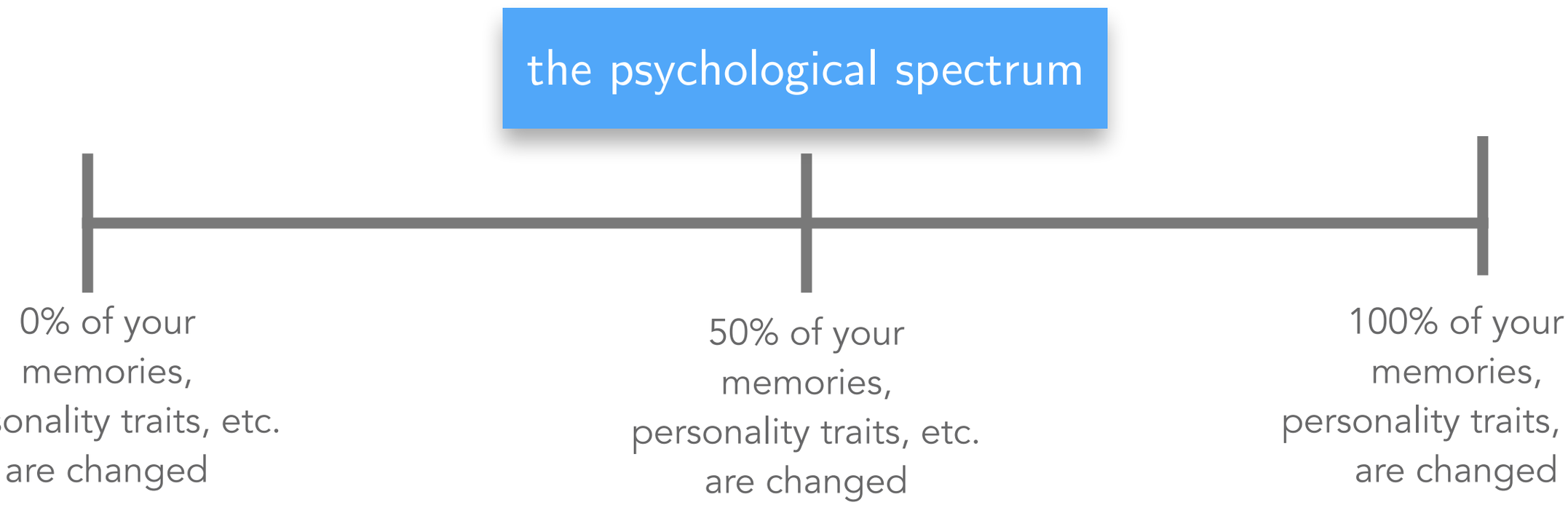
Never rich

Even at the end of the spectrum, I am still not rich.

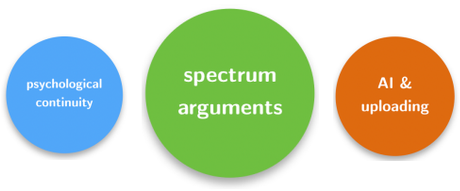
Which answer is most plausible in the case of the wealth spectrum?



Now let's return to the psychological spectrum.



Here we have basically the same three choices.



the psychological spectrum

Sharp Cut Off

There is a precise point in the spectrum at which, for the first time, I would not survive the surgery. Perhaps it is when 43.13% of the person's memories and character are changed.

Indeterminacy

In the first cases I survive; in the last cases I do not survive; but there is no sharp cut off. Instead, there is a range of cases in which it is not determinately true either that the person is me or that the person is not me.

Survive All

Even in the cases at the right edge of the spectrum, I survive.

Which of these three is the most plausible?

Sharp Cut Off

There is a precise point in the spectrum at which, for the first time, I would not survive the surgery. Perhaps it is when 43.13% of the person's memories and character are changed.

Indeterminacy

In the first cases I survive; in the last cases I do not survive; but there is no sharp cut off. Instead, there is a range of cases in which it is not determinately true either that the person is me or that the person is not me.

Survive All

Even in the cases at the right edge of the spectrum, I survive.

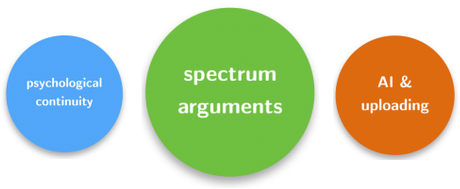
Which of these three is the most plausible?

The psychological theorist seems forced into either the Sharp Cut Off view or Indeterminacy. But both of those can seem hard to believe.

The materialist can endorse Survive All — which looks at first glance more attractive.

What should the dualist say?

The dualist might endorse Survive All — that would make this case like the case of psychology-swapping that we discussed last time.



Sharp Cut Off

There is a precise point in the spectrum at which, for the first time, I would not survive the surgery. Perhaps it is when 43.13% of the person's memories and character are changed.

Indeterminacy

In the first cases I survive; in the last cases I do not survive; but there is no sharp cut off. Instead, there is a range of cases in which it is not determinately true either that the person is me or that the person is not me.

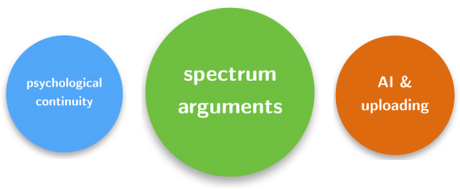
Survive All

Even in the cases at the right edge of the spectrum, I survive.

What should the dualist say?

The dualist might endorse Survive All — that would make this case like the case of psychology-swapping that we discussed last time.

Alternatively, the dualist might say that a certain amount of psychological change forces the body's connection to the soul to be severed, so that the soul attached to the body changes. Then it looks like the dualist's choices are the same as the psychological theorist's.



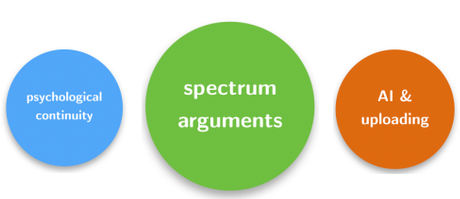
What should the dualist say?

The dualist might endorse Survive All — that would make this case like the case of psychology-swapping that we discussed last time.

Alternatively, the dualist might say that a certain amount of psychological change forces the body's connection to the soul to be severed, so that the soul attached to the body changes. Then it looks like the dualist's choices are the same as the psychological theorist's.

So far, the psychological spectrum might seem to give our materialist theory of survival the advantage over psychological survival.

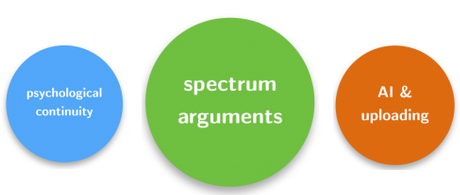
But, as Parfit points out, we can come up with a parallel example which makes trouble for the materialist theory of survival.



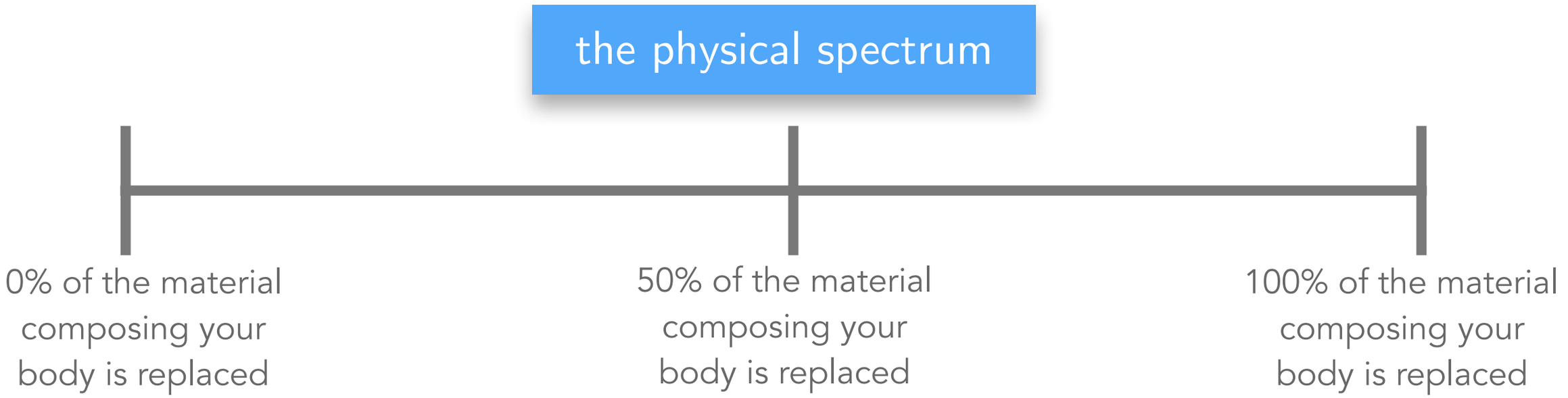
But, as Parfit points out, we can come up with a parallel example which makes trouble for the materialist theory of survival.

> ‘‘Consider the physical spectrum. In a case close to the near end, scientists would replace 1% of the cells in my brain and body with exact duplicates. In the case in the middle of the spectrum, they would replace 50%. In a case near the far end, they would replace 99%, leaving only 1% of my original brain and body. At the far end, the ‘replacement’ would involve the complete destruction of my brain and body, and the creation out of new organic matter of a Replica of me.’’

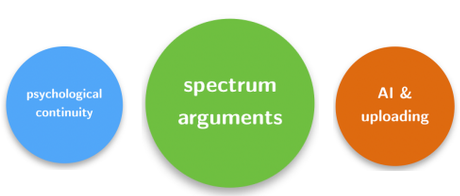
We can represent this case much as we represented the psychological spectrum.



We can represent this case much as we represented the psychological spectrum.



And it looks like we have the same choices about how to respond.



the physical spectrum

And it looks like we have the same choices about how to respond.

Sharp Cut Off

There is a precise point in the spectrum at which, for the first time, I would not survive the surgery. Perhaps it is when 43.13% of the matter composing the organism is replaced.

Indeterminacy

In the first cases I survive; in the last cases I do not survive; but there is no sharp cut off. Instead, there is a range of cases in which it is not determinately true either that the person is me or that the person is not me.

Survive All

Even in the cases at the right edge of the spectrum, I survive.

Sharp Cut Off

There is a precise point in the spectrum at which, for the first time, I would not survive the surgery. Perhaps it is when 43.13% of the matter composing the organism is replaced.

Indeterminacy

In the first cases I survive; in the last cases I do not survive; but there is no sharp cut off. Instead, there is a range of cases in which it is not determinately true either that the person is me or that the person is not me.

Survive All

Even in the cases at the right edge of the spectrum, I survive.

Which of these is the most plausible treatment of the physical spectrum?

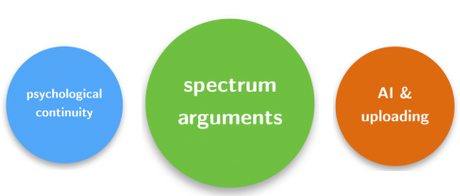
Here the situation is just the opposite as with the psychological spectrum. The materialist is forced into endorsing Sharp Cut Off or Indeterminacy; the psychological theorist can endorse Survive All.

What should the dualist say?

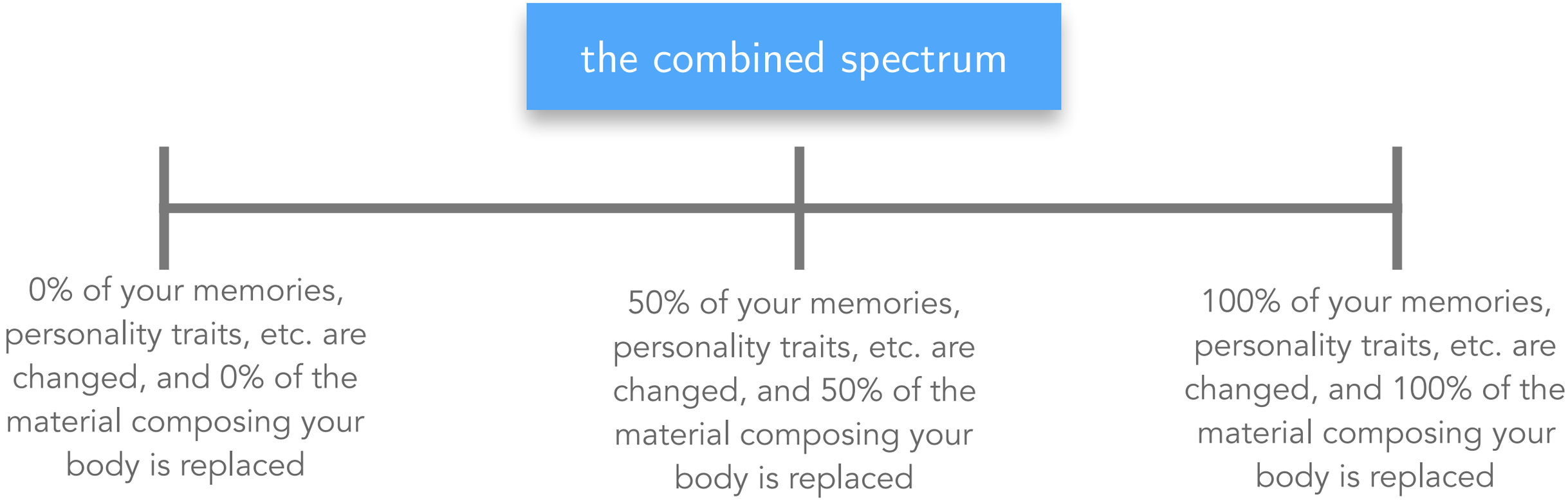
The last and most puzzling spectrum argument is the combined spectrum:

> ‘‘At the near end of this spectrum is the normal case in which a future person would be fully continuous with me as I am now, both physically and psychologically. This person would be me in just the way that, in my actual life, it will be me who wakes up tomorrow. At the far end of this spectrum the resulting person would have no continuity with me as I am now, either physically or psychologically. In this case the scientists would destroy my brain and body, and then create, out of new organic matter, a perfect Replica of someone else. Let us suppose this person to be Greta Garbo. We can suppose that, when Garbo was 30, a group of scientists recorded the states of all the cells in her brain and body.’’

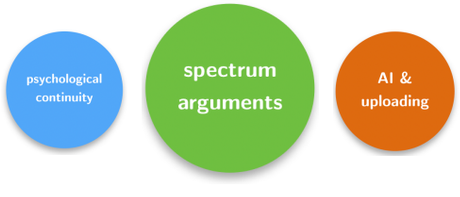
In the intermediate stages, the person is to some degree physically like you and to some degree physically like Garbo, and to some degree psychologically like you and to some degree psychologically like Garbo.



In the intermediate stages, the person is to some degree physically like you and to some degree physically like Garbo, and to some degree psychologically like you and to some degree psychologically like Garbo.



We again have just three choices.



the combined spectrum

0% of your memories, personality traits, etc. are changed, and 0% of the material composing your body is replaced

50% of your memories, personality traits, etc. are changed, and 50% of the material composing your body is replaced

100% of your memories, personality traits, etc. are changed, and 100% of the material composing your body is replaced

Sharp Cut Off

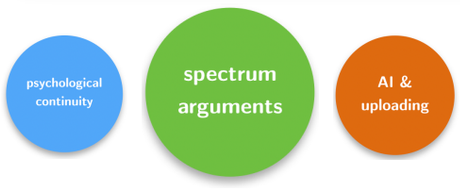
There is a precise point in the spectrum at which, for the first time, I would not survive the surgery. Perhaps it is when 43.13% of your psychological traits have changed and the same percentage of the matter composing the organism is replaced.

Indeterminacy

In the first cases I survive; in the last cases I do not survive; but there is no sharp cut off. Instead, there is a range of cases in which it is not determinately true either that the person is me or that the person is not me.

Survive All

Even in the cases at the right edge of the spectrum, I survive.



Unlike our other spectrum cases, we can all agree that Survive All looks pretty implausible.

Sharp Cut Off

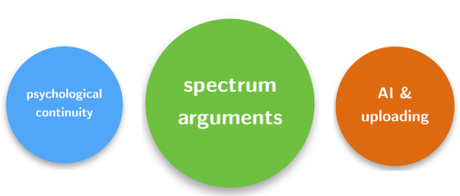
There is a precise point in the spectrum at which, for the first time, I would not survive the surgery. Perhaps it is when 43.13% of your psychological traits have changed and the same percentage of the matter composing the organism is replaced.

Indeterminacy

In the first cases I survive; in the last cases I do not survive; but there is no sharp cut off. Instead, there is a range of cases in which it is not determinately true either that the person is me or that the person is not me.

Unlike our other spectrum cases, we can all agree that Survive All looks pretty implausible.

Here is an argument against Sharp Cut Off. If Sharp Cut Off were true, then there are two adjacent procedures on the combined spectrum which are such that I should care an enormous amount which procedure happens to me. (After all, I would survive one but not the other.) But in reality it would never be rational to care which of two such similar procedures I should undergo.



Sharp Cut Off

There is a precise point in the spectrum at which, for the first time, I would not survive the surgery. Perhaps it is when 43.13% of your psychological traits have changed and the same percentage of the matter composing the organism is replaced.

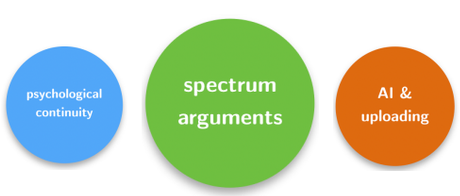
Indeterminacy

In the first cases I survive; in the last cases I do not survive; but there is no sharp cut off. Instead, there is a range of cases in which it is not determinately true either that the person is me or that the person is not me.

Does Sharp Cut Off look more plausible if one is a dualist? Couldn't one then say that there is a point in the combined spectrum at which the soul would lose its connection to the body, and that this would explain the existence of a cut off point?

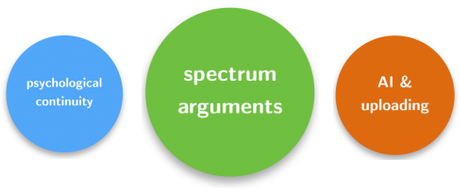
But even here there are puzzles. Suppose that you underwent one of the procedures in the middle of the combined spectrum? Could you tell afterwards whether you had survived?

And what should the dualist say about cases to the right of the cut off point (wherever that is) -- is a new soul created, or joined to the body for the first time, by the procedure?



So Parfit thinks that the moral of the spectrum arguments is not that the psychological theory is false, but that we should change a fundamental part of our view about what our own continued existence amounts to.

> “[One] assumes that, in each of these cases, the resulting person either would or would not be me. This is not so. The resulting person would be me in the first few cases. In the last case he would not be me. In many of the intervening cases, neither answer would be true. I can always ask, ‘Am I about to die? Will there be some person living who will be me?’ But, in the cases in the middle of this Spectrum, there is no answer to this question.”

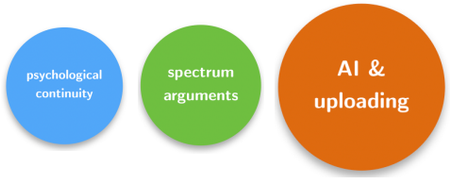


The spectrum cases are challenges for any view of survival. They are also related to practical questions which may become pressing in the course of your lifetime with improvements in artificial intelligence.

'Artificial intelligence' is a term for the ability of machines to perform tasks intelligently: for example, to strategize and to solve problems.

So defined, artificial intelligence is now all around us. There are plenty of examples of AI systems which are vastly better than humans at performing various tasks.

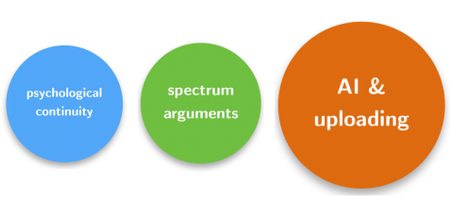
What does not yet exist is a **general** artificial intelligence: an artificial intelligence capable of doing all or almost all of the things that an ordinary adult human being can do. No machine in existence (that we know of) has general artificial intelligence.



What does not yet exist is a **general** artificial intelligence: an artificial intelligence capable of doing all or almost all of the things that an ordinary adult human being can do. No machine in existence (that we know of) has general artificial intelligence.

One very interesting question is whether, and when, we will develop human level artificial intelligence. A recent survey of researchers in the field gave an average guess of the year 2100 — but opinions vary widely.

We will focus on one way in which human level (and greater than human level) AI might be achieved, and some of the philosophical challenges it poses.



Our topic today is one way in which human level (and greater than human level) AI might be achieved, and some of the philosophical challenges it poses.

Suppose that it is the year 2045. We have now developed silicon devices which replicate but improve upon the functioning of neurons or clusters of neurons. The silicon devices do just the same things as the neurons they replace, but more quickly and more efficiently.

You have the opportunity to have part of your brain replaced with silicon devices of this kind. Lots of your friends have done this, and they can process information much more quickly than they used to be able to. You find yourself consistently underperforming relative to your peers who have had the synthetic replacement done — and you suspect that your newly super-smart friends are beginning to find it kind of boring to talk to you.

If given the opportunity to go in for partial synthetic replacement, would you do it?

You have the opportunity to have part of your brain replaced with silicon devices of this kind. Lots of your friends have done this, and they can process information much more quickly than they used to be able to. You find yourself consistently underperforming relative to your peers who have had the synthetic replacement done — and you suspect that your newly super-smart friends are beginning to find it kind of boring to talk to you.

If given the opportunity to go in for partial synthetic replacement, would you do it?

Once you have part of your brain replaced in this way, it seems to be irresistible to gradually have all of your brain replaced in this way (assuming that the surgery is affordable). Why would you want to keep part of your underperforming biological brain around?

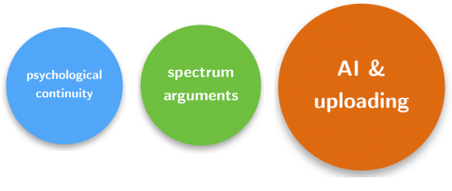
Suppose that you were now given the opportunity to have your synthetic brain supplemented with improved memory, so that more of your memories could be reliably stored and retrieved. Would you opt for that as well?

Suppose that you were now given the opportunity to have your synthetic brain supplemented with improved memory, so that more of your memories could be reliably stored and retrieved. Would you opt for that as well?

You would now have become, at least in part, an AI system with greater than human level intelligence. Your intelligence would be in many ways like human intelligence — but you would have much faster processing speed and much better memory.

Having traded in your brain for an artificial system, you might become annoyed with the limitations of your other biological parts.

For example, we could presumably replace all of your organs and body parts with synthetic systems which were not subject to decay, and which worked much better than your current biological parts. Perhaps you would no longer have to sleep or eat (though you might have the option to do so).



Having traded in your brain for an artificial system, you might become annoyed with the limitations of your other biological parts.

For example, we could presumably replace all of your organs and body parts with synthetic systems which were not subject to decay, and which worked much better than your current biological parts. Perhaps you would no longer have to sleep or eat (though you might have the option to do so).

This might make you effectively immortal (barring some disaster). After all, replacement of any of your failed parts would now be a straightforward matter.

Would you trade in the rest of your biological parts for synthetic replacements? (Again, it may help to imagine that your friends have all done this, and are now annoyed with your “biological” limitations.)

This might make you effectively immortal (barring some disaster). After all, replacement of any of your failed parts would now be a straightforward matter.

Would you trade in the rest of your biological parts for synthetic replacements? (Again, it may help to imagine that your friends have all done this, and are now annoyed with your “biological” limitations.)

At this point it seems that you would have become an artificial intelligence. You would no longer be a biological organism.

The scenarios just laid out show that it is not wildly implausible to think that you will be faced with choices like this in your lifetime, and that it is not wildly implausible to think that decisions which lead to this outcome would be very tempting.

But at this stage it is natural to pose the following question: would the synthetic being which results from these changes be you? Would you survive?

Let’s look at three examples.

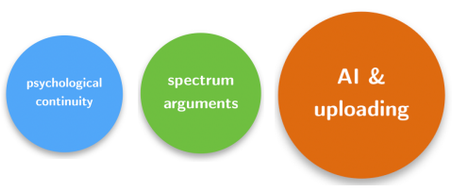
First, let's consider a process of what David Chalmers calls gradual destructive uploading.

Maria is considering whether to "go synthetic." Being a cautious person, she does this gradually. At t1, she has one neuron replaced by a silicon device which replicates the functioning of that neuron.

Would she notice a change? It seems that she would not.

So now suppose that she has a second neuron replaced. Would she notice a change? Again, it seems that she would not.

This process might continue until all of Maria's neurons have been replaced. Gradually, this synthetic system inside her head could then be supplemented in ways which gave it more memory and greater processing speed. Here Maria would notice a difference — she would be able gradually so solve problems faster, and remember much more. But it does not seem as though changes of this kind could make it "no longer Maria."

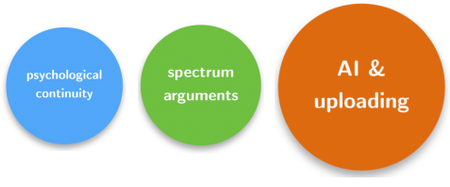


This process might continue until all of Maria's neurons have been replaced. Gradually, this synthetic system inside her head could then be supplemented in ways which gave it more memory and greater processing speed. Here Maria would notice a difference — she would be able gradually so solve problems faster, and remember much more. But it does not seem as though changes of this kind could make it "no longer Maria."

Once we have gone his far, it seems pretty clear that we could provide synthetic replacements of all of Maria's body parts without her ceasing to exist. Surely replacing Maria's index finger with a synthetic replacement need not involve a change in identity!

Now imagine the same process, but that it occurs much faster; perhaps each replacement occurs in a fraction of a second. Surely this would not matter; the time it takes to perform a replacement seems irrelevant.

This argument seems to show that one can survive gradual destructive uploading.



Let's, following Chalmers, call the outcome of these procedures "DigiMaria." Our argument suggests that DigiMaria = Maria.

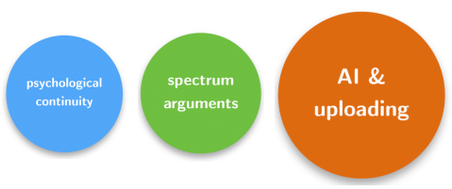
Let’s look at a second example.

Caleb is considering whether to go synthetic. But he does not have Maria’s patience, and is nervous about having parts of his body destroyed.

He is therefore given the option of going for instant nondestructive uploading. A synthetic version of Caleb — DigiCaleb — is created while Caleb watches. DigiCaleb is like Caleb in certain ways (just as DigiMaria is like Maria in certain ways) — but of course DigiCaleb is much smarter than Caleb, and less prone to bodily damage of various kinds.

Is Caleb identical to DigiCaleb? Surely not. Caleb could not take cyanide and expect to survive as DigiCaleb; the presence of an improved twin in the room won’t change the fact that cyanide will kill Caleb.

Our argument suggests that nondestructive uploading does not preserve identity; the synthetic thing created may resemble you in various ways, but it is not you.



Mindful of Caleb's fate, Emily decides to take a different path. Like Caleb, she lacks the patience for gradual uploading. But she wants to become a synthetic thing, and knows that Caleb failed to achieve this.

So Emily decides to go for instant destructive uploading. In this process, Emily's body is destroyed, and right away a synthetic version — DigiEmily — is created.

Did Emily survive the procedure?

A strong case can be made that she did not, because Emily seems relevantly just like Caleb — the only difference is that Emily was destroyed whereas Caleb was not. But why should that matter?

It seems that things came to an end for Emily when her body was destroyed; the fact that DigiEmily was later created seems irrelevant to her survival. But if she did not survive, then she is not DigiEmily (she isn't anyone any more).

If you agree with this, then it seems that one cannot survive instant destructive uploading.

So far, you might think, so good. One can survive gradual destructive uploading but not instant destructive uploading, so I will just opt for the gradual version of the procedure! So let's suppose:

I can survive slow gradual destructive uploading.

I cannot survive instant destructive uploading.

Maybe that is correct. But there is at least a tension here. First, it looks like the speed of the gradual destructive uploading should not matter. So it looks like:

If I can survive slow gradual destructive uploading, then I can survive fast destructive uploading.

I can survive slow gradual destructive uploading.

If I can survive slow gradual destructive uploading, then I can survive fast destructive uploading.

I cannot survive instant destructive uploading.

But now consider a super-super-fast version of gradual uploading; perhaps the entire process is complete in a small fraction of a second. Could that really be importantly different from instant uploading? There is at least some tendency to think that the difference between a super-super-fast sequence of changes and a simultaneous change could not matter. That suggests:

If I can survive fast destructive uploading., then I can survive instant destructive uploading.

I can survive slow gradual destructive uploading.

If I can survive slow gradual destructive uploading, then I can survive fast destructive uploading.

If I can survive fast destructive uploading., then I can survive instant destructive uploading.

I can survive instant destructive uploading.

I cannot survive instant destructive uploading.

But this is a contradiction. So one of our assumptions must be false. Which one is it?