

Survival, AI, and
cognitive
enhancement

Much of our discussion of the identity question and the survival question has focused on some farfetched examples, including teletransportation and imaginary cases of body swapping. Today we're going to talk about how the challenges posed by those examples arise in the context of technological changes which may well occur within your lifetime.

These technological changes involve 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.

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

One of the milestones in public awareness of artificial intelligence was the 1997 chess match between the world chess champion Garry Kasparov and an IBM supercomputer called "Deep Blue." Kasparov had beaten Deep Blue in 1996 — but many were shocked when Deep Blue won in 1997.

Here is a (very) simplified explanation of how Deep Blue worked. When it was its move, Deep Blue considered a range of possible moves. It then considered, for each of those moves, a range of possible response moves its opponent could make. It then considered, for each of those response moves you get the idea. For each possible configuration of pieces on the board, Deep Blue was able to evaluate how advantageous that position was for it. It then moved in such a way as to maximize the best outcome. The machine was capable of evaluating roughly 200 million configurations per second.

Chess machines have now moved well beyond Deep Blue, and it is now uncontroversial that the best of these are considerably stronger than the best human players.

In a way, this is unsurprising. We already know that machines are better than us at performing calculations quickly. If we give the machine the information about which configurations on the board are better than which other ones, and give it sufficient computing power to consider vastly more possibilities (and longer trees of moves) than we can, you might think that we should expect a machine to be able to beat us at a complex but delimited game like chess. How is this any different in principle than a machine being better than any human at multiplying large numbers?

It is instructive to think about how artificial intelligence has progressed since Deep Blue.

It is instructive to think about how artificial intelligence has progressed since Deep Blue.

In 2015 the Stockfish chess engine (which you can think of as a faster updated version of Deep Blue) played 100 games against Google's AlphaZero AI. AlphaZero won 28 and lost 0. It did this despite using less computing power — it searched 80,000 positions/second vs. Stockfish's 70 million positions/second.

How did it do this? AlphaZero was programmed in a very different way. Rather than being given as input a mass of information about various chess games and outcomes, it was (simplifying massively) simply given the rules of chess and told to play against itself, learning from its own successes and failures. According to the team who set this up, AlphaZero surpassed Stockfish after only four hours of training.

Nor is AlphaZero just a chess engine — given the rules of Go, a Chinese game which is in certain respects vastly more complex than chess, it quickly taught itself to become the best Go player in the world.

The example of AlphaZero shows that artificial intelligence is well beyond machines which simply compute human-designed algorithms very quickly. In both chess and Go, AlphaZero developed styles of play which were radically unlike anything human players had used.

Despite this, the intelligence of AlphaZero is limited. It can beat you at chess, but it cannot figure out how to make coffee, order food at a restaurant, pass a college philosophy course, or negotiate a good starting salary for a job.

It is not, that 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.

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.

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 Chalmers calls **gradual destructive uploading**.

Maria is considering whether to "go synthetic." Being a cautious person, she does this gradually. At t_1 , 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 to 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 to 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 this 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:



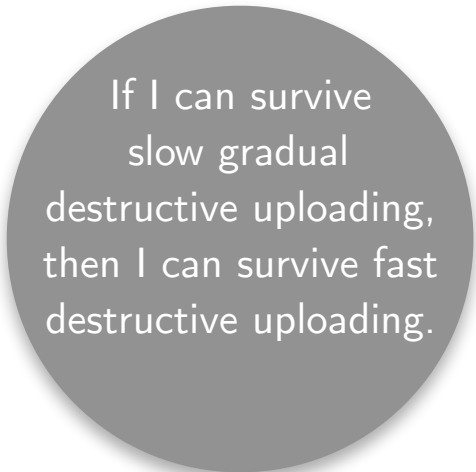
I can survive slow gradual destructive uploading.



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.

I cannot survive instant 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.

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