

Will computers ever become more intelligent than humans? What would that mean for society?

'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.

Let's use "AI" as a label for human-level general artificial intelligence.

Let's use "AI" as a label for human-level general artificial intelligence.

Some have thought that if AI is possible, then there will be an "intelligence explosion" — a process, perhaps a very rapid one, of the creation of ever more intelligent machines. This intelligence explosion is often called "the singularity."

This gives us three questions.

Will there be AI?

If there is AI, will there be a singularity?

If there is a singularity, how should we respond?

A blue circle with a thin grey border, centered at the top of the page. Inside the circle, the text "Will there be AI?" is written in white, sans-serif font.

Will there be AI?

We see the rapid growth of artificial intelligence all around us. In our phones, in our cars, and in our homes. This alone encourages the thought that AI is possible.

Estimates as to when AI will be achieved vary greatly; a recent survey of leaders in the field gave an average of the year 2100. While it is reasonable to be suspicious of future predictions of this, there is a near consensus that it will (barring catastrophes like nuclear war or extreme global warming) happen.

We see the rapid growth of artificial intelligence all around us. In our phones, in our cars, and in our homes. This alone encourages the thought that AI is possible.

Estimates as to when AI will be achieved vary greatly; a recent survey of leaders in the field gave an average of the year 2100. While it is reasonable to be suspicious of future predictions of this, there is a near consensus that it will (barring catastrophes like nuclear war or extreme global warming) happen.

Here is one way to argue for this. A computer could be designed which would emulate a human brain. We do not now have anywhere near the resources to construct such a thing; but it is hard to believe that it is in principle impossible to create a computer which would duplicate the functions of a particular brain.

It is also hard to believe that this computer would not have AI. If your brain were embedded in a system very different than your body, wouldn't your brain still have the kind of intelligence that it now has?

It is also hard to believe that this computer would not have AI. If your brain were embedded in a system very different than your body, wouldn't your brain still have the kind of intelligence that it now has?

It is true that this brain-emulating computer need not have the physical abilities which you have — it might not be able to make a cup of coffee.

But it is also hard to believe that it is in principle impossible to create a physical system in which the computer could be embedded which would duplicate your physical abilities. Of course present day robotics is nowhere near this — but it is hard to see what the in-principle stumbling blocks could be.

This makes at least a reasonably strong case that (barring catastrophes of the sort already mentioned) AI will exist.

This makes at least a reasonably strong case that (barring catastrophes of the sort already mentioned) AI will exist.

Let's turn to our second question.

If there is AI, will
there be a
singularity?

The argument for a singularity was laid out in a 1957 article by LJ Good.

The argument for a singularity was laid out in a 1957 article by LJ Good.

‘Let an ultraintelligent machine be defined as a machine that can far surpass all the intellectual activities of any man however clever. Since the design of machines is one of these intellectual activities, an ultraintelligent machine could design even better machines; there would then unquestionably be an intelligence explosion, and the intelligence of man would be left far behind. Thus the first ultraintelligent machine is the last invention that man need ever make.’

The argument suggested in this passage is simple. Among the things that intelligent things can do is create more intelligent things. The more intelligent a thing is, the more intelligent things that it can make are. And that suggests that the creation of machines more intelligent than us will lead via a series of creative acts to ever more intelligent machines.

The argument suggested in this passage is simple. Among the things that intelligent things can do is create more intelligent things. The more intelligent a thing is, the more intelligent things that it can make are. And that suggests that the creation of machines more intelligent than us will lead via a series of creative acts to ever more intelligent machines.

We can lay this out in argument form using terminology Chalmers introduces in the reading. Let AI be human level intelligence, AI+ be artificial intelligence greater than human level intelligence, and AI++ be artificial intelligence which massively exceeds human level intelligence. (Chalmers suggests that we think of AI++ as intelligence which stands to our intelligence as ours stands to mouse intelligence.)

We can then argue as follows.

We can lay this out in argument form using terminology Chalmers introduces in the reading. Let AI be human level intelligence, AI+ be artificial intelligence greater than human level intelligence, and AI++ be artificial intelligence which massively exceeds human level intelligence. (Chalmers suggests that we think of AI++ as intelligence which stands to our intelligence as ours stands to mouse intelligence.)

We can then argue as follows.

1. There will be AI.
 2. If there will be AI, there will be AI+.
 3. If there will be AI+, there will be AI++.
-
- C. There will be AI++. (1,2,3)

We have already talked about the justification for (1). How about (2) and (3)?

1. There will be AI.
 2. If there will be AI, there will be AI+.
 3. If there will be AI+, there will be AI++.
-
- C. There will be AI++. (1,2,3)

We have already talked about the justification for (1). How about (2) and (3)?

In defense of (2): suppose that we construct a computer which emulates a human brain, as discussed above. Presumably it would then be possible to give that computer massively more computing speed and memory than a brain. (After all, we already know how to make computers which exceed brains in these respects.) That computer would then have AI+.

It may be tempting to reply that nothing can make a machine which is more intelligent than itself. But that neglects the fact that we can already make machines which are better than us in many ways, and neglects the kind of process by which AlphaZero was created.

1. There will be AI.
 2. If there will be AI, there will be AI+.
 3. If there will be AI+, there will be AI++.
-
- C. There will be AI++. (1,2,3)

So suppose that there will be AI+. Why think that premise (3) is true?

The argument here is the one implicit in the quote from Good. If we can think of ways to extend a machine which has AI to one which has AI+, presumably a machine which has AI+ will be able to think of ways of making a yet smarter machine. And that one will be able to think of ways to make a yet smarter one. And so on, apparently without end. And that leads to the singularity.

One response would be to say: "OK, this is possible, but it will take an enormously long time. It took us many millennia to make a computer — it might well take millennia for a computer to make an even marginally smarter computer."

One response would be to say: "OK, this is possible, but it will take an enormously long time. It took us many millennia to make a computer — it might well take millennia for a computer to make an even marginally smarter computer."

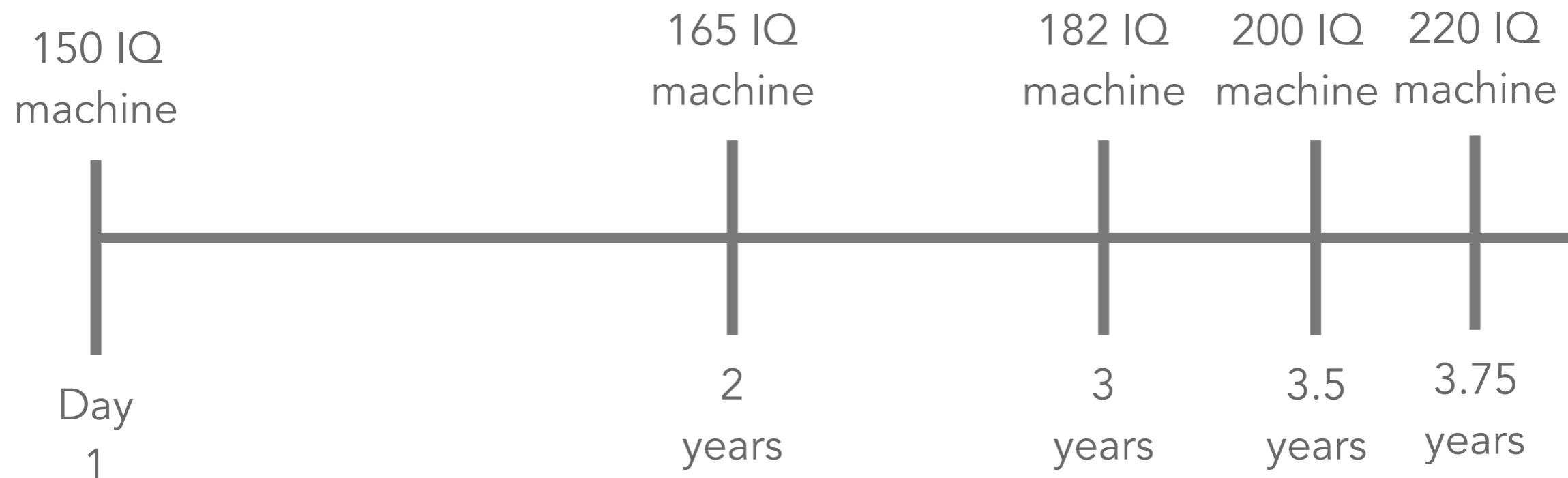
But on some not-crazy assumptions, this could actually proceed with frightening speed.

Suppose (to use Chalmers' example) that a machine with AI+ can produce another machine which is twice as fast and 10% more intelligent. (The increase in speed is not out of line with current improvements in computer processing speed.)

Suppose further that the machine with AI+ has an IQ of 150. (IQ would presumably break down for superintelligence, but it will be useful to fix ideas.) Then the four years after the creation of AI+ will look like this:

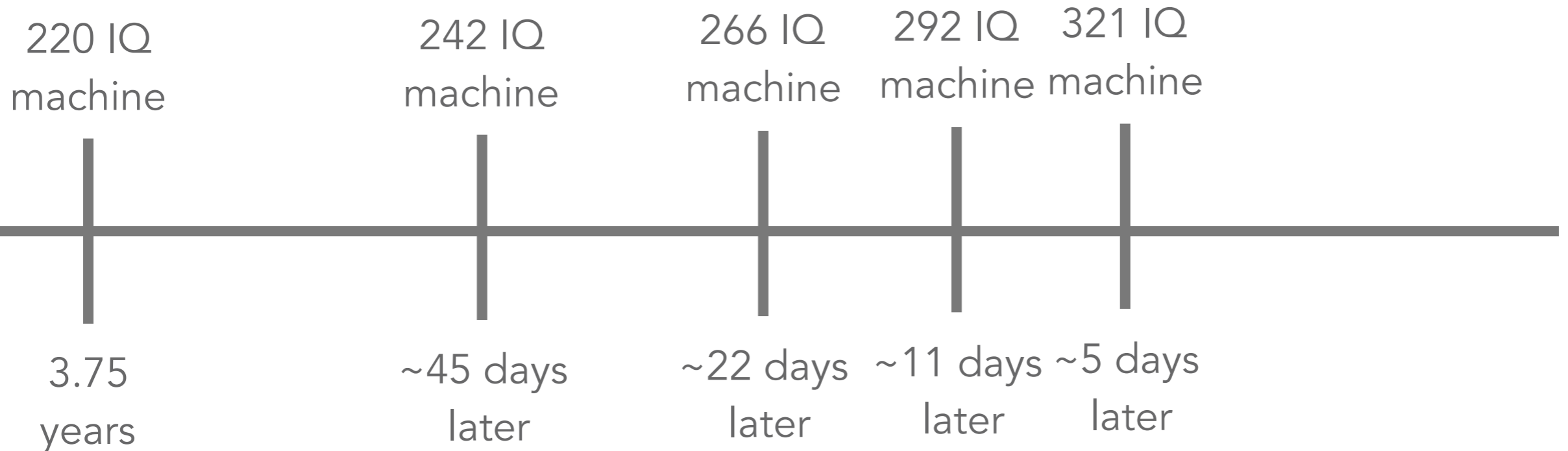
Suppose (to use Chalmers' example) that a machine with AI+ can produce another machine which is twice as fast and 10% more intelligent. (The increase in speed is not out of line with current improvements in computer processing speed.)

Suppose further that the machine with AI+ has an IQ of 150. (IQ would presumably break down for superintelligence, but it will be useful to fix ideas.) Then the four years after the creation of AI+ will look like this:



Suppose (to use Chalmers' example) that a machine with AI+ can produce another machine which is twice as fast and 10% more intelligent in two years. (The increase in speed is not out of line with current improvements in computer processing speed.)

Suppose further that the machine with AI+ has an IQ of 150. (IQ would presumably break down for superintelligence, but it will be useful to fix ideas.) Then the four years after the creation of AI+ will look like this:



There is no finite limit to the IQ of the machine created in four years time.

There is thus a strong argument, not just for the eventual existence of a singularity, but for the rapid occurrence of the singularity. The question is thus less why this would happen than why it wouldn't.

One possible reason has already been discussed — a global catastrophe of some kind. Are there others?

One would be a slowdown in the growth of processing speed. This would slow the process — but it could still be quite rapid.

Another would be some kind of upper bound in the possible levels of intelligence — but there is no obvious reason to think that, even if there is such an upper bound, it will occur anywhere near human level intelligence.

Another would be some kind of upper bound in the possible levels of intelligence — but there is no obvious reason to think that, even if there is such an upper bound, it will occur anywhere near human level intelligence.

Perhaps the strongest obstacle to a rapid singularity is us: we could decide that we do not want the singularity to occur. How likely is this?

One might think that it is not very likely. After all, all that is needed for the singularity is one group creating an AI+ which is designed to create ever more intelligent machines.

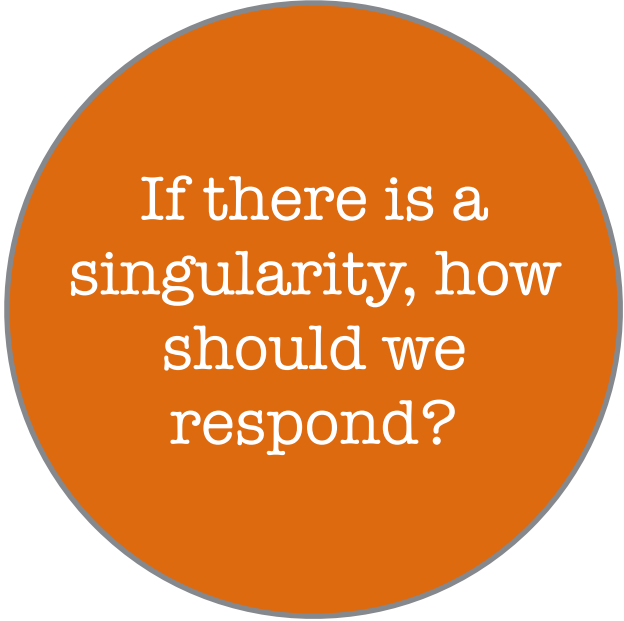
Could this be blocked by some sort of international agreement? Perhaps; but given the obvious military uses of artificial intelligence and the levels of distrust between countries, there will be very strong incentives to not fall behind in the race to create greater than human intelligence.

One might think that it is not very likely. After all, all that is needed for the singularity is one group creating an AI+ which is designed to create ever more intelligent machines.

Could this be blocked by some sort of international agreement? Perhaps; but given the obvious military uses of artificial intelligence and the levels of distrust between countries, there will be very strong incentives to not fall behind in the race to create greater than human intelligence.

We have seen that there is a plausible argument that the singularity will occur, perhaps even in your lifetime. That leads to our last question:

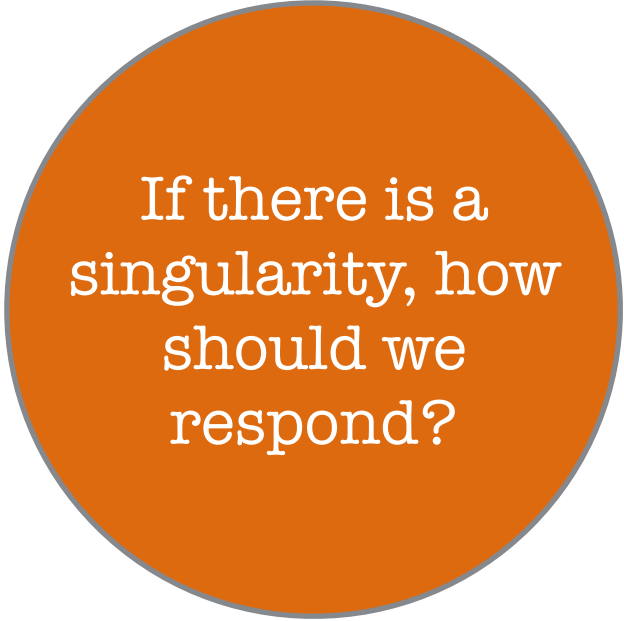
If there is a singularity, how should we respond?



If there is a
singularity, how
should we
respond?

This is an important choice, as Chalmers says:

‘If there is a singularity, it will be one of the most important events in the history of the planet. An intelligence explosion has enormous potential benefits: a cure for all known diseases, an end to poverty, extraordinary scientific advances, and much more. It also has enormous potential dangers: an end to the human race, an arms race of warring machines, the power to destroy the planet. So if there is even a small chance that there will be a singularity, we would do well to think about what forms it might take and whether there is anything we can do to influence the outcomes in a positive direction.’



If there is a
singularity, how
should we
respond?

Chalmers thinks that there are four options: extinction, isolation, inferiority, and integration.

Let's set aside extinction, which seems less than ideal.

To opt for isolation would be to opt for a world in which, while artificial super-intelligence exists, we remove ourselves from it, and live apart from it.

This may be practically impossible; there would be enormous temptations to interact with the AI++ systems, and to use them to help solve problems like disease and climate change, as well as for less well-meaning ends.

Let's set aside extinction, which seems less than ideal.

To opt for isolation would be to opt for a world in which, while artificial super-intelligence exists, we remove ourselves from it, and live apart from it.

This may be practically impossible; there would be enormous temptations to interact with the AI++ systems, and to use them to help solve problems like disease and climate change, as well as for less well-meaning ends.

To opt for inferiority would be to live along with the AI++ systems roughly as we are now. Chalmers thinks that this option "threatens to diminish the significance of our lives."

Why might one think this?

Here is one reason: AI++ systems would be able to do anything vastly better than a human can do it. This would include not just solving mathematical and scientific problems, but also solving political and ethical problems, creating works of art, and cooking food.

To opt for inferiority would be to live along with the AI++ systems roughly as we are now. Chalmers thinks that this option “threatens to diminish the significance of our lives.”

Why might one think this?

Here is one reason: AI++ systems would be able to do anything vastly better than a human can do it. This would include not just solving mathematical and scientific problems, but also solving political and ethical problems, creating works of art, and cooking food.

There would therefore be literally nothing left for us to do. Whatever you might think of doing, a machine could do it vastly better; whatever job you might think you want to have, a machine could do it vastly better. So in a sense there would be no such thing as real achievement. Would this diminish the significance of your life?

There would therefore be literally nothing left for us to do. Whatever you might think of doing, a machine could do it vastly better; whatever job you might think you want to have, a machine could do it vastly better. So in a sense there would be no such thing as real achievement. Would this diminish the significance of your life?

Here's an analogy which might suggest that it would not. We consider winning the 100 meter dash in the Olympics an achievement. But of course cars can go 100 meters faster than people can. Could we think of achievement in this way? Yes, one could never be the best at anything, or even pretty good at anything, relative to an AI++ machine. But you could still try to be the best human!

Whether even this kind of achievement could remain is connected to the question of whether some would opt for Chalmers' fourth option — integration.

Whether even this kind of achievement could remain is connected to the question of whether some would opt for Chalmers' fourth option — integration.

On this option, we become superintelligent ourselves. How might this work?

This would presumably involve becoming at least partly non-biological things. Suppose, for example, that a computer could be made, part of which emulated your current brain. It could then be supplemented in the ways we have already discussed to give it vastly greater memory and vastly greater processing speed than your brain now has.

If given the opportunity to have surgery in which your brain was replaced with such a system, would you do it?

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

For example, AI++ 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?

Here's a last thought experiment. If everyone around you was trading in their brains and body parts for synthetic replacements, and thus becoming vastly more intelligent and physically able, would this convince you to do so as well? Remember that they are not just a bit smarter than you — they would become so much smarter than you that talking to you would become for them something like what you talking to a dog is like now.

It is natural to think that the answers to these questions about what one would or should do are connected to questions about whether the synthetic individual which resulted from these replacements would really be you.

This leads directly to the kinds of questions we asked when we were thinking about Parfit's views of personal identity.

Let's look first at an argument that the resulting synthetic thing would be you, and then an argument that it would not. (Both are adapted from Chalmers' 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 is Maria.

Let's look at another 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.

This is a case of **nondestructive uploading**. 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. 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.

Maybe that is correct. But there is at least a tension here.

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.

So there is some tendency to accept all of the following claims:

One can survive slow gradual uploading.

If one can survive slow gradual uploading, one can survive fast gradual uploading.

If one can survive fast gradual uploading, one can survive instant uploading.

One cannot survive instant uploading.

So there is some tendency to accept all of the following claims:

One can survive slow gradual uploading.

If one can survive slow gradual uploading, one can survive fast gradual uploading.

If one can survive fast gradual uploading, one can survive instant uploading.

One cannot survive instant uploading.

But of course not all of these could be true.

There is a residual question here, which may be a practical question for you.

Suppose that it is 2098, and you have been diagnosed with a disease which will give you one year to live. AI has just been achieved, and it is very likely that the possibility of undergoing synthetic replacement is just a few years away. You have the chance to have your brain and body scanned, knowing that after you die these can be used to create a synthetic replica of yourself.

On the one hand, you are convinced by the arguments that you cannot survive instant destructive uploading — and this case seems just like that one, except with a bigger time gap between the destruction and the creation of the new synthetic entity. So it seems unlikely that that thing will be you. And it seems a little weird to create a synthetic copy of yourself.

On the other hand ... philosophy is hard, and maybe you would survive!

What would you do?