Drew McDermott
Response to “The Singularity: A Philosophical Analysis”

I agree with David Chalmers about one thing: it is useful to see the arguments for the singularity written down using the philosophers’ signature deductive framework, in which controversial premises are made explicit. If all concur that the form of the argument has been captured, then they can get down to the brass tacks of refuting or rebutting the numbered premises.

To give away my inclinations up front, I tend to disagree with Chalmers about the prospects for the singularity, agree about uploading, and agree with some of his conclusions about controlling the singularity, and disagree with others.

I have not much to say about the distinctions among the various kinds of singularity discussed at the beginning of the paper. They are well described by Chalmers, but most of the time they are not relevant, or if they are it is hard to judge what the consequences of the distinctions might be.

Prospects

Should we be discussing the singularity? Sure, it’s a free country. I personally think that even though it’s true that the “the singularity [would] be one of the most important events in the history of the planet,” other more likely scenarios may rule it out. In those scenarios the problems we face *right now* may, if not solved, escalate to the point of disaster, making the possibility of a singularity moot. I’ll lay out the scenario I most fear at the end of this review.

But putting those nightmares aside, what dreams or nightmares might the singularity bring about, if any?

For ultraintelligent machines to exist, intelligent machines must exist, and, as Chalmers says, “every path to AI has proved surprisingly difficult to date.” He lays out the arguments that AI is inevitable in spite of its difficulty, and I accept the conclusion, if not all of his arguments.

The next step is to argue that if there is AI then there will be AI+ “soon after,” where AI+ is artificial intelligence surpassing human intelligence. I’ll accept this conclusion as well, because I agree with Chalmers that it is unlikely human beings’ level of intelligence is unsurpassable. (My argument is different: as soon as bipedal primates reached a minimum level of intelligence sufficient for civilization to develop, it did, and here we are; it would be a big coincidence if that level was also a maximum.)

My skepticism is mainly focused on the third premise: If there is AI+, there will be AI++ (soon, perhaps very soon, after), where AI++ is artificial intelligence far surpassing humans'. The case that there *exists* a level of intelligence far surpassing humans' is much weaker than the case that there exists a level surpassing it at all. No argument for this presupposition is given in the paper.

The argument for the third premise takes the form of mathematical induction. As stated, the argument is unsound, because a series of increases from AI_n to AI_{n+1}, each exponentially smaller than the previous one, will reach a limit. But the text clarifies that one might need to assume that AI_{n+1} is at least δ times larger than AI_n, where δ is a constant ratio.

I find this claim unsatisfying. For one thing, it is unnecessary. The mathematical induction should, I think, be thought of as shorthand for an "empirical induction" of a fixed number of steps, at most $\lceil log_\delta P \rceil$, where P is the ratio between AI++ and AI (=AI_0). (There may be weaker ways to patch the argument, not introducing this minimum ratio δ, but I don't know what they would be.)

But the claim is also, unfortunately, unsupported by any argument, except a set of proposals to shore up the notion of an "extendible method" of producing intelligence, a method that can "easily be improved, yielding more intelligent systems."

Three extendible methods are put forward: direct programming, machine learning, and artificial evolution. Direct programming is extendible because of the well-known fact that "almost every program that has yet been written [is] improvable in multiple respects." What respects? I can think of some: bug elimination, inner-loop optimization, and porting to faster hardware. None of these are extendible, except possibily the last, which works only if Moore's Law stays true indefinitely. But Moore's Law is a good example of why extendible methods are a mirage. What seems like a smooth, tidy curve in the dreams of application programmers is a series of hairy technological innovations from the point of view of hardware designers and systems programmers. The methods used to make circuits smaller are quite different from the methods used to automate the adaptation of programs to run on multi-core processors. But even the phrase "methods used to make circuits smaller" is misleading because there are no such methods; at each stage completely novel ways of shrinking the components of the semiconductor fab line had to be developed at ever-escalating costs.

As far as machine learning is concerned, I am quite puzzled by the idea that a good learning algorithm can be improved indefinitely. For instance, why do so few neural networks have more than one hidden layer? Because, it turns out, backpropagation to more than one such layer provides too weak a signal to learn at a useful rate. And yet many concepts are difficult to capture in shallow networks. There has been progress in this area, but it required new ideas. Waiting for a new idea to come along is not an extendible method, indeed not a method at all.

The third of the three extendible methods is artificial evolution. So far evolutionary algorithms have proven to do fairly well in solving some fairly interesting problems. I don't quite see what's extendible about them, or why they're of special interest, except that *we* were produced by an evolutionary algorithm, in a sense. The current picture of that process is not encouraging, because to get to an intelligent primate a series of lucky breaks had to occur. At one point in fairly recent biological history (within the last 200,000 years) the population of our ancestors was down to a few thousand individuals. It could so easily have gone all the way to zero, and the planet would still be waiting for its first intelligence to appear. What makes us think the path through the high-dimensional space of environments and agents will require fewer lucky twists and turns in the future?

Finally, let me point out a general argument against the existence of extendible methods.

> *Theorem:* There are no extendible methods unless $P = NP$.
>
> I have a wonderful proof of this result which is unfortunately too big to fit in this review. The whole thing is at `http://cs-www.cs.yale.edu/homes/dvm/papers/no-extendible-methods.pdf`.

If you doubt that $P = NP$, as I do, then it's hard to believe that any extendible method, including waiting for Moore, exists.

But sweep away all those objections, there is still the "motivational defeater" Chalmers describes as "AI+ systems [being] disinclined to create their successors, perhaps because we design them to be so disinclined, or perhaps because they will be intelligent enough to realize that creating successors in not in their interests." If it's possible to start the sequence AI$+_1$, AI$+_2$, $\ldots$, I don't think we can say in advance anything at all about the motivations of what will amount to *artificial persons.* In particular, I seriously doubt we can design them to be inclined or disinclined one way or another. Even if the first in the series has motivations under our control, the later elements will be designed with the help of members of the earlier cohorts, and so will be difficult for us to understand. If we knew all there was to know about a fellow human's brain circuitry and memories, could we predict what they would do by any method other than simulating them (and their surroundings) on a faster computer than they "have," or "are"? I think not. But simulating AI$+_2$ before bringing it into existence is logically impossible, because the simulation would essentially *be* AI+$_2$, running in a virtual world, with all the problems of keeping it contained that Chalmers discusses. So the only way to control the motivations of the next generation of AI+'s is to create several different versions, fiddling with whatever parameters we have that might have an impact on its motivations as we go from one version to the next. This achieves control of the next generation's motivations in the same sense that control of the trajectory of the next ballistic missile you launch may be achieved by launching several different ballistic missiles in succession, correcting the trajectory slightly each time. Which is to say, it doesn't achieve control of anything, but might create a series of misplaced explosions.

Consequences

Early in the paper Chalmers says that "An intelligence explosion has enormous potential benefits: a cure for all known diseases, an end to poverty, extraordinary scientific advances, and much more" (sect. 1). Later we are offered a choice among being helped to die out, being consigned to a virtual world, being allowed to live as inferiors, or being succeeded by a mechanical superrace that may or may not remember being us; or, in his concise formula, "extinction, isolation, inferiority, or integration"(sect. 8).[1]

The difference between these two lists illustrates the fact that mere intelligence is not enough to guarantee an end to poverty (to pick one of the most intractable problems); it might bring about an end to the human race. In between these two possibilities is the far more likely scenario in which AI+ or AI++ finds a way to end poverty, and human greed and fear (and the greed and fear of the AIs involved in the political process) prevent it from succeeding. Or: the AI working for the United Nations finds cures for diseases almost as fast as the AIs working for contending powers think up new ones to use in biological warfare.

Chalmers investigates in greatest detail the possibility that as AI++ is developed, we gradually become integrated with it. But would it really still be "we"? One gets into a lot of tangles by assuming that there is a definite answer to this question. It is easy to construct cases in which one's intuitions are driven one way or the other, no matter where they start, as Chalmers explains so capably.

I tend to believe that personal identity is mostly a matter of social convention; it's not settled purely by facts about the contenders. (I think this

[1] We can upgrade this list to the cooler slogan "the five I's": interment, isolation, inferiority, impersonation, or improvement, the last two items replacing "integration," which covers both bogus and genuine continuity of identity.

is a subspecies of what Chalmers calls the “deflationary” view.) If at some point it becomes a normal convention that people survive uploading, and the uploaded include many prominent citizens, who are indignant at the idea that they’re not conscious, or that the DigiX they are now differs from BioX, the biological entity they started as, then at that point going virtual will have become a normal stage of life. Almost everyone will expect to survive it, and this mass belief will make it so.

The biological incarnation of human beings, if it is still necessary, would be viewed the way intelligent amphibians might view tadpoles, as a “larval” phase of human existence. A refusal to get on with life and get uploaded might be seen as similar to a refusal to emerge from the womb. If by chance my “larva” should not be destroyed by the uploader, people would see the larva as having *no* claim to being me, and society would have no qualms about hunting it down and euthanizing it. In this scenario, I would be among the first to agree with this decision, and so would the larva, who might have an elemental desire to avoid death, but would lay claim to nothing more than being a discarded piece of a growing person.

Since social conventions would evolve along with technological possibility, there would be no controversy regarding personal identity, and people would indeed be transformed into posthuman computational entities. But in spite of these facts, outside observers might be justified in concluding that the human race had ceased to exist, because no one was making adult human beings any more. They might miss us, the way we might wish some *Homo habilis* were still around, even though (let’s suppose) they became us and no one ever regretted it.

And last, let me say why I think it would be a pity for too many smart people to devote time and mental effort to thinking about the Singularity, which is that there are other possible events, unambiguous catastrophes, with higher probability. The most likely one, the one that really scares me, is an environmental collapse followed by a nuclear war as the survivors quarrel over the remaining resources. I can vividly picture this happening within the lifetime of *some of the people reading this,* and if not yours, then *your and my children’s.* I can picture them being refugees fighting for survival — or dead — amidst the ruins of civilization. So I think most of our resources and energy should be directed toward the problems of resource management and nuclear disarmament.

The exponential growth in technology that is the major argument for the Singularity is accompanied by, perhaps made possible by, an exponential growth in exploitation of finite natural resources (including the atmosphere, viewed as a carbon-dioxide sponge). Our civilization’s addiction to a process that simply cannot continue is a sign of insanity, and belief in the Singularity may be one of its most comforting delusions. Even if some of the world's richer citizens get "uploaded," what happens when the power goes off?