Book Review: Reframing Superintelligence

Posted on August 27, 2019 by Scott Alexander

📖

I

Ten years ago, everyone was talking about superintelligence, the singularity, the robot apocalypse. What happened?

I think the main answer is: the field matured. Why isn't everyone talking about nuclear security, biodefense, or counterterrorism? Because there are already competent institutions working on those problems, and people who are worried about them don't feel the need to take their case directly to the public. The past ten years have seen AI goal alignment reach that level of maturity too. There are all sorts of new research labs, think tanks, and companies working on it – the Center For Human-Compatible AI at UC Berkeley, OpenAI, Ought, the Center For The Governance Of AI at Oxford, the Leverhulme Center For The Future Of Intelligence at Cambridge, etc. Like every field, it could still use more funding and talent. But it's at a point where academic respectability trades off against public awareness at a rate where webzine articles saying CARE ABOUT THIS OR YOU WILL DEFINITELY DIE are less helpful.

One unhappy consequence of this happy state of affairs is that it's harder to keep up with the field. In 2014, Nick Bostrom wrote *Superintelligence: Paths, Dangers, Strategies*, giving a readable overview of what everyone was thinking up to that point. Since then, things have been less public-facing, less readable, and more likely to be published in dense papers with a lot of mathematical notation. They've also been – no offense to everyone working on this – less revolutionary and less interesting.

This is one reason I was glad to come across Reframing Superintelligence: Comprehensive AI Services As General Intelligence by Eric Drexler, a researcher who works alongside Bostrom at Oxford's Future of Humanity Institute. This 200 page report is not quite as readable as *Superintelligence*; its highly-structured outline form belies the fact that all of its claims start sounding the same after a while. But it's five years more recent, and presents a very different vision of how future AI might look.

Drexler asks: what if future AI looks a lot like current AI, but better?

For example, take Google Translate. A future superintelligent Google Translate would be able to translate texts faster and better than any human translator, capturing subtleties of language beyond what even a native speaker could pick up. It might be able to understand hundreds of languages, handle complicated multilingual puns with ease, do all sorts of amazing things. But in the end, it would just be a translation app. It wouldn't want to take over the world. It wouldn't even "want" to become better at translating than it was already. It would just translate stuff really well.

The future could contain a vast ecosystem of these superintelligent services before any superintelligent agents arrive. It could have media services that can write books or generate movies to fit your personal tastes. It could have invention services that can design faster cars, safer rockets, and environmentally friendly power plants. It could have strategy services that can run presidential campaigns, steer Fortune 500 companies, and advise governments. All of them would be far more effective than any human at performing their given task. But you couldn't ask the presidential-campaign-running service to design a rocket any more than you could ask Photoshop to run a spreadsheet.

In this future, our AI technology would have taken the same path as our physical technology. The human body can run fast, lift weights, and fight off enemies. But the automobile, crane, and gun are three different machines. Evolution had to cram running-ability, lifting-ability, and fighting-ability into the same body, but humans had more options and were able to do better by separating them out. In the same way, evolution had to cram book-writing, technology-inventing, and strategic-planning into the same kind of intelligence – an intelligence that also has associated goals and drives. But humans don't have to do that, and we probably won't. We're not doing it today in 2019, when Google Translate and AlphaGo are two different AIs; there's no reason to write a single AI that both translates languages and plays Go. And we probably won't do it in the superintelligent future either. Any assumption that we will is based more on anthropomorphism than on a true understanding of intelligence.

These superintelligent services would be safer than general-purpose superintelligent agents. General-purpose superintelligent agents (from here on: agents) would need a human-like structure of goals and desires to operate independently in the world; Bostrom has explained ways this is likely to go wrong. AI services would just sit around algorithmically mapping inputs to outputs in a specific domain.

Superintelligent services would not self-improve. You could build an AI researching service – or, more likely, several different services to help with several different aspects of AI research – but each of them would just be good at solving certain AI research problems. It would still take human researchers to apply their insights and actually build something new. In theory you might be able to automate every single part of AI research, but it would be a weird idiosyncratic project that wouldn't be anybody's first choice.

Most important, superintelligent services could help keep the world safe from less benevolent AIs. Drexler agrees that a self-improving general purpose AI agent is possible, and assumes someone will build one eventually, if only for the lulz. He agrees this could go about the way Bostrom expects it to go, ie very badly. But he hopes that there will be a robust ecosystem of AI services active by then, giving humans superintelligent help in containing rogue AIs. Superintelligent anomaly detectors might be able to notice rogue agents causing trouble, superintelligent strategic planners might be able to develop plans for getting rid of them, and superintelligent military research AIs might be able to create weapons capable of fighting them off.

Drexler therefore does not completely dismiss Bostromian disaster scenarios, but thinks we should concentrate on the relatively mild failure modes of superintelligent AI services. These may involve normal bugs, where the AI has aberrant behaviors that don’t get caught in testing and cause a plane crash or something, but not the unsolveable catastrophes of the Bostromian paradigm. Drexler is more concerned about potential misuse by human actors – either illegal use by criminals and enemy militaries, or antisocial use to create things like an infinitely-addictive super-Facebook. He doesn’t devote a lot of space to these, and it looks like he hopes these can be dealt with through the usual processes, or by prosocial actors with superintelligent services on their side (thirty years from now, maybe people will say “it takes a good guy with an AI to stop a bad guy with an AI”).

This segues nicely into some similar concerns that OpenAI researcher Paul Christiano has brought up. He worries that AI services will be naturally better at satisfying objective criteria than at “making the world better” in some vague sense. Tasks like “maximize clicks to this site” or “maximize profits from this corporation” are objective criteria; tasks like “provide real value to users of this site instead of just clickbait” or “have this corporation act in a socially responsible way” are vague. That means AI may asymmetrically empower some of the worst tedencies in our society without giving a corresponding power increase to normal people just trying to live enjoyable lives. In his model, one of the tasks of AI safety research is to get AIs to be as good at optimizing vague prosocial tasks as they will naturally be at optimizing the bottom line. Drexler doesn't specifically discuss this in *Reframing Superintelligence*, but it seems to fit the spirit of the kind of thing he's concerned about.

II

I'm not sure how much of the AI alignment community is thinking in a Drexlerian vs. a Bostromian way, or whether that is even a real dichotomy that a knowledgeable person would talk about. I know there are still some people who are very concerned that even programs that seem to be innocent superintelligent services will be able to self-improve, develop misaligned goals, and cause catastrophes. I got to talk to Dr. Drexler a few years ago about some of this (although I hadn't read the book at the time, didn't understand the ideas very well, and probably made a fool of myself); at the time, he said that his work was getting a mixed reception. And there are still a few issues that confuse me.

First, many tasks require general intelligence. For example, an AI operating in a domain with few past examples (eg planning defense against a nuclear attack) will not be able to use modern training paradigms. When humans work on these domains, they use something like common sense, which is presumably the sort of thing we have because we understand thousands of different domains from gardening to ballistics and this gives us a basic sense of how the world works in general. Drexler agrees that we will want

AIs with domain-general knowledge that cannot be instilled by training, but he argues that this is still “a service”. He agrees these tasks may require AI architectures different from any that currently exist, with relatively complete world-models, multi-domain reasoning abilities, and the ability to learn “on the fly” – but he doesn’t believe those architectures will need to be agents. Is he right?

Second, is it easier to train services or agents? Suppose you want a good multi-domain reasoner that can help you navigate a complex world. One proposal is to create AIs that train themselves to excel in world simulations the same way AlphaGo trained itself to excel in simulated games of Go against itself. This sounds a little like the evolutionary process that created humans, and agent-like drives might be a natural thing to come out of this process. If agents were easier to “evolve” than services, agentic AI might arise at an earlier stage, either because designers don’t see a problem with it or because they don’t realize it is agentic in the relevant sese.

Third, how difficult is it to separate agency from cognition? Natural intelligences use “active sampling” strategies at levels as basic as sensory perception, deciding how to direct attention in order to best achieve their goals. At higher levels, they decide things like which books to read, whose advice to seek out, or what subdomain of the problem to evaluate first. So far AIs have managed to address even very difficult problems without doing this in an agentic way. Can this continue forever? Or will there be some point at which intelligences with this ability outperform those without it.

I think Drexler's basic insight is that Bostromian agents need to be really different from our current paradigm to do any of the things Bostrom predicts. A paperclip maximizer built on current technology would have to eat gigabytes of training data about various ways people have tried to get paperclips in the past so it can build a model that lets it predict what works. It would build the model on its actually-existing hardware (not an agent that could adapt to much better hardware or change its hardware whenever convenient). The model would have a superintelligent understanding of the principles that had guided some things to succeed or fail in the training data, but wouldn't be able to go far beyond them into completely new out-of-the-box strategies. It would then output some of those plans to a human, who would look them over and make paperclips 10% more effectively.

The very fact that this is less effective than the Bostromian agent suggests there will be pressure to build the Bostromian agent eventually (Drexler disagrees with this, but I don't understand why). But this will be a very different project from AI the way it currently exists, and if AI the way it currently exists can be extended all the way to superintelligence, that would give us a way to deal with hostile superintelligences in the future.

III

All of this seems kind of common sense to me now. This is worrying, because I didn't think of any of it when I read *Superintelligence* in 2014.

I asked readers to tell me if there was any past discussion of this. Many people brought up Robin Hanson's arguments, which match the "ecosystem of many AIs" part of Drexler's criticisms but don't focus as much on services vs. agents. Other people brought up discussion under the heading of Tool AI. Combine those two strains of thought, and you more or less have Drexler's thesis, minus some polish. I read some of these discussions, but I think I failed to really understand them at the time. Maybe I failed to combine them, focused too much on the idea of an Oracle AI, and missed the idea of an ecosystem of services. Or maybe it all just seemed too abstract and arbitrary when I had fewer examples of real AI systems to think about.

I've sent this post by a couple of other people, who push back against it. They say they still think Bostrom was right on the merits and superintelligent agents are more likely than superintelligent services. Many brought up Gwern's essay on why tool AIs are likely to turn into agent AIs and this post by Eliezer Yudkowsky on the same topic – I should probably reread these, reread Drexler's counterarguments, and get a better understanding. For now I don't think I have much of a conclusion either way. But I think I made a mistake of creativity in not generating or understanding Drexler's position earlier, which makes me more concerned about how many other things I might be missing.