ISSUES
IN SCIENCE AND TECHNOLOGY

NATIONAL ACADEMY OF SCIENCES
NATIONAL ACADEMY OF ENGINEERING
INSTITUTE OF MEDICINE
THE UNIVERSITY OF TEXAS AT DALLAS
ARIZONA STATE UNIVERSITY
SPRING 2015

A New Model for the American Research University

Clean Energy Diplomacy from Bush to Obama

Physics Envy: Get Over It

The Limitations of Climate Models as Guides for Policy

Welcome to the Anthropocene

Empowering Social Science

An Excess of Research Space?

First Science Fiction Contest Winner

USA $12.00 / CAN $13.00
5 1>
0 56698 57433 0

Machine smart

Superintelligence: Paths, Danger, Strategies

by Nick Bostrom. Oxford, England: Oxford University Press, 2014, 352 pp.

Dan Gordon

The subject of intelligent machines that decide that they don't have much use for us has haunted our species at least since golems first were mentioned in the Talmud. And more recently, the issue of superintelligence has been worked over by science fiction authors from Isaac Asimov to Vernor Vinge and beyond. We've thought about this a lot.

Now philosophers have their turn. Oxford University philosopher Nick Bostrom's book *Superintelligence* gives the subject a thorough treatment. His conclusion? We better be damn careful what kind of intelligent machines we build.

Bostrom's erudition bursts from every page. He has a background in physics, computational neuroscience, and mathematical logic, as well as philosophy. He uses all of these disciplines, and more, to advance his argument, which has four main parts.

Part 1: Machine intelligence is feasible. Bostrom reviews the current approaches to computer-based intelligence and divides them roughly into *brain emulation* and *pure artificial intelligence (AI)* approaches, with hybrids and mongrels in between.

Brain emulation intelligence works by completely emulating a human brain—down to the level of neurons and dendrites and cortical columns—in such detail that the person instantiated in that brain comes to life in the artificial medium of computer hardware and software.

Pure AI takes a different course, attempting to build in software a pure artifact that acts intelligently but not in any way that traces a heritage to our native wetware (other than the important detail that we designed the artifact in the first place.)

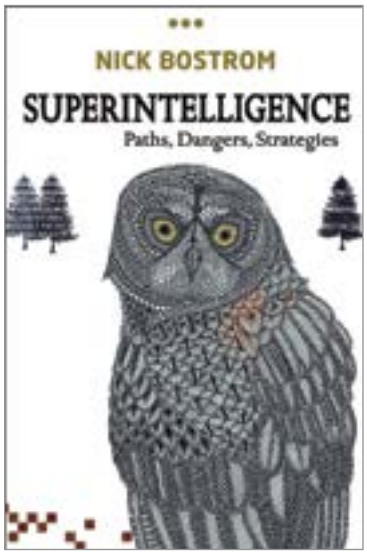
Bostrom maintains that both approaches could feasibly lead to AIs, although he believes that the two approaches have different strengths and weaknesses, and may lead to different future scenarios. Because we are presumably just "running mind software" on a different hardware platform, Bostrom believes that brain emulation AIs are more likely to "be like us," whatever that means, but pure AIs, because all of the design elements are explicit, may be easier for our minds to comprehend and predict. He concludes that brain-emulation AIs are likely to come on the scene sooner, but that either form may arrive by mid-century.

Part 2: Bostrom then argues that once an AI exists, it may (and likely will) rapidly improve its own intelligence. By "rapidly," he means within seconds or hours or days, not months or years. He believes that there may be no limit to this self-improvement, to the point where an AI develops what Bostrom calls "decisive strategic advantage" and is able to neuter potential alternatives or adversaries and, rather rapidly, consolidate its power as what he calls a "singleton." Such a singleton would, in effect, control the future of humanity, what Bostrom calls its "cosmic endowment."

Part 3: There is no special reason to believe that a singleton's intentions would be benign. Bostrom discusses at length what might be the "final purposes" (his term; we might call them "ultimate goals" or "life purpose") of such an all-powerful superintelligence, and how we might influence those purposes. This line of inquiry, which occupies most of the book, is a hash of game theory considerations and speculations about the nature of an AI and its capabilities. How might we, for example, prevent a singleton AI from converting the entire observable universe into paperclips if that were its final purpose?

Part 4: In the final chapters, Bostrom discusses what is to be done. How should we act in the face of what he considers the practical certainty that a superintelligence will be developed—if not within decades, perhaps within a century or two—whose motives might not be benign and whose ability to act on its motives might be unstoppable?

He advises us to, in effect, form a League of Extraordinary Humans whose purpose is to systematically and strategically discuss the emergence of a superintelligence. Not to utterly make fun of Bostrom's approach, we might call this an Iron Rice Bowl (the Chinese term for occupation-for-life) for Philosophers.



What are we to make of Bostrom's case?

In the first place, it is a serious argument. If we might in the relatively near future invent our cosmic replacement, then we are required, in the name of humanity's cosmic endowment (which Bostrom calculates to comprise some 10^{58} real or virtual future lives), to give the matter some thought. And Bostrom is quite correct that this kind of problem might benefit from long study. But what are our chances of affecting the outcome?

The core problem is that the leap between today's "intelligent" software and a superintelligence is unknown, and our

temptation is to mystify it. Whether we are building brain emulations or pure AIs, we don't understand what would make them "come to life" as intelligent beings, let alone superintelligent.

"Machine learning" software today uses a statistical model of a subject area to "master" it. Mastering consists of changing the weights of the various elements in the model in response to a set of training instances (situations where human trainers grade the instances: "yes, this is credit card fraud," "no, this is not a valid English sentence," etc.). Clear enough, but it just doesn't seem very much like what our minds do.

And the path from this kind of "learning" (it is an anthropomorphism even to call it learning) to what "human-intelligent" agents do is completely unclear.

It might require nothing but simple scale. A small "machine learning" system may be subintelligent, and at some size, if we had enough computing power and enough elements in the model and enough training instances and enough support, intelligence might "emerge."

This has certainly been the mantra of AI for some decades, and it may have been what technophiles hoped for when IBM's Watson software beat two Jeopardy champions a couple of years back.

Sadly, Watson has not gone on to master, on its own or even with expert human help, any general corpus of knowledge. At a Watson showcase event last year, the demo apps were all mired in the swamp of endless training and re-training that I recall from my AI days in the '80's. There was no indication that unleashing Watson on different domains and at different scales was going to lead to general intelligence, although one is free to hope.

Another path to general intelligence, as some Husserlians, such as Hubert Dreyfus or other more anthropologically-inclined researchers think, may involve human feelings, purposes, or drives. If the AI wanted something badly enough (not to be shut off, for example), the argument goes, then it would learn from its "experiences" and get smarter. Combine "desires" like this with natural selection at scale via a genetic-selection or evolutionary approach, and you might gradually enhance the intelligence of primitive agents. With machine speeds, this could happen quickly.

The core problem is that the leap between today's "intelligent" software and a superintelligence is unknown, and our temptation is to mystify it.

The problem with this approach has been coming up with a mechanical definition of "feelings," "purposes," or "drives." We can write some software that is aimed at doing something, but it is missing something of what we associate with a drive: urgency, existential angst, whatever. Maybe we are confusing the qualia of purpose with the essence of it, and maybe a human-infused purpose can launch software on the road to agency. But at some point it has to have "its own" purposes, whatever that means.

A third approach has been to insist that there is something implicit in our brains that is unique, whether we call this uniqueness "embodied-ness" (with Dreyfus) or "bearing human motivational ancestry" (with Bostrom). Is there something implicit in the organization of our brains that renders us intelligent? If so, then emulating a brain should supply it, unless an emulated brain is like a silk flower. As Dreyfus remarked at one point, we don't think that the software simulation of a thunderstorm should get us wet, do we? Why should the software emulation of a brain embody whatever makes us intelligent?

This "missing link" between AI software today and general intelligence tomorrow wouldn't be so important if it weren't at the heart of Bostrom's argument about how to control emerging AIs. If intelligence emerges from scale or from endogenous machine "drives" or for embodied-ness, how can we hope to put a governor on the motives of machine intelligences? They would toss our flimsy moral strictures aside as easily as adult humans toss away Santa Claus.

But talking about children does give us some suggestions about an approach to making AIs moral. Sigmund Freud believed that children form a superego at an age when they are "impressionable" but not yet adult in their reasoning. A superego, in his theory, is a moral mechanism that functions imperfectly (filled with demons and fascists as well as avatars of light and Christ figures) but is good enough to guide most adults to a reasonable course of moral behavior. Maybe we can fashion a superego for our young AIs and give them enough guidance to allow them to muddle through when they reach adulthood without turning the entire universe into paperclips or destroying us so we don't ask them tough questions.

That is Bostrom's great hope, that we can issue a suitable instruction to emerging AIs (something along the lines of "do the best thing we mean for you to do, even if we can't say it precisely") that will constrain their range of possibilities when they become fully superintelligent. All of us would benefit.

Dan Gordon is a technology partner with Valhalla Partners, a venture capital firm in Vienna, Virginia.