SAM HARRIS

THE BLOG

Can We Avoid a Digital Apocalypse?

A Response to the 2015 Edge Question

Consciousness | Economics | Ethics | Neuroscience | Philosophy | January 16, 2015

(Photo via Armand Turpel)

It seems increasingly likely that we will one day build machines that possess superhuman intelligence. We need only

continue to produce better computers?which we will, unless we destroy ourselves or meet our end some other way. We already know that it is possible for mere matter to acquire ?general intelligence??the ability to learn new concepts and employ them in unfamiliar contexts?because the 1,200 cc of salty porridge inside our heads has managed it. There is no reason to believe that a suitably advanced digital computer couldn?t do the same.

It is often said that the near-term goal is to build a machine that possesses ?human level? intelligence. But unless we specifically emulate a human brain?with all its limitations?this is a false goal. The computer on which I am writing these words already possesses superhuman powers of memory and calculation. It also has potential access to most of the world?s information. Unless we take extraordinary steps to hobble it, any future artificial general intelligence (AGI) will exceed human performance on every task for which it is considered a source of ?intelligence? in the first place. Whether such a machine would necessarily be conscious is an open question. But conscious or not, an AGI might very well develop goals incompatible with our own. Just how sudden and lethal this parting of the ways might be is now the subject of much colorful speculation.

One way of glimpsing the coming risk is to imagine what might happen if we accomplished our aims and built a superhuman AGI that behaved exactly as intended. Such a machine would quickly free us from drudgery and even from the inconvenience of doing most intellectual work. What would follow under our current political order? There is no law of economics that guarantees that human beings will find jobs in the presence of every possible technological advance. Once we built the perfect labor-saving device, the cost of manufacturing new devices would approach the cost of raw materials. Absent a willingness to immediately put this new capital at the service of all humanity, a few of us would enjoy unimaginable wealth, and the rest would be free to starve. Even in the presence of a truly benign AGI, we could find ourselves slipping back to a state of nature, policed by drones.

And what would the Russians or the Chinese do if they learned that some company in Silicon Valley was about to develop a superintelligent AGI? This machine would, by definition, be capable of waging war?terrestrial and cyber?with unprecedented power. How would our adversaries behave on the brink of such a winner-take-all scenario? Mere rumors of an AGI might cause our species to go berserk.

It is sobering to admit that chaos seems a probable outcome even in the *best-case* scenario, in which the AGI remained perfectly obedient. But of course we cannot assume the best-case scenario. In fact, ?the control problem??the solution to which would guarantee obedience in any advanced AGI?appears quite difficult to solve.

Imagine, for instance, that we build a computer that is no more intelligent than the average team of researchers at Stanford or MIT?but, because it functions on a digital timescale, it runs a million times faster than the minds that built it. Set it humming for a week, and it would perform 20,000 years of human-level intellectual work. What are the chances that such an entity would remain content to take direction from us? And how could we confidently predict the thoughts and actions of an *autonomous* agent that sees more deeply into the past, present, and future than we do?

The fact that we seem to be hastening toward some sort of digital apocalypse poses several intellectual and ethical challenges. For instance, in order to have any hope that a superintelligent AGI would have values commensurate with our own, we would have to instill those values in it (or otherwise get it to emulate us). But whose values should count? Should *everyone* get a vote in creating the utility function of our new colossus? If nothing else, the invention of an AGI would force us to resolve some very old (and boring) arguments in moral philosophy.

However, a true AGI would probably acquire *new* values, or at least develop novel?and perhaps dangerous?near-term goals. What steps might a superintelligence take to ensure its continued survival or access to computational resources? Whether the behavior of such a machine would remain compatible with human flourishing might be the most important question our species ever asks.

The problem, however, is that only a few of us seem to be in a position to think this question through. Indeed, the moment of truth might arrive amid circumstances that are disconcertingly informal and inauspicious: Picture ten young men in a room?several of them with undiagnosed Asperger?s?drinking Red Bull and wondering whether to flip a switch. Should any single company or research group be able to decide the fate of humanity? The question nearly answers itself.

And yet it is beginning to seem likely that some small number of smart people will one day roll these dice. And the temptation will be understandable. We confront problems?Alzheimer?s disease, climate change, economic instability?for which superhuman intelligence could offer a solution. In fact, the only thing nearly as scary as building an AGI is the prospect of *not* building one. Nevertheless, those who are closest to doing this work have the greatest responsibility to anticipate its dangers. Yes, other fields pose extraordinary risks?but the difference between AGI and something like synthetic biology is that, in the latter, the most dangerous innovations (such as germline mutation) are *not* the most tempting, commercially or ethically. With AGI the most powerful methods (such as recursive self-improvement) are precisely those that entail the most risk.

We seem to be in the process of building a God. Now would be a good time to wonder whether it will (or even can) be a good one.

?

Read the other responses at Edge.org

Notes

Find this article online at: https://www.samharris.org/blog/item/can-we-avoid-a-digital-apocalypse