Five ways the superintelligence revolution might happen

September 26 2014, by Nick Bostrom

Biological brains are unlikely to be the final stage of intelligence. Machines already have superhuman strength, speed and stamina – and one day they will have superhuman intelligence. This is of course not certain to occur – it is possible that we will develop some other dangerous technology first that destroys us, or otherwise fall victim to some existential risk.

But assuming that scientific and technological progress continues, human-level machine intelligence is very likely to be developed. And shortly thereafter, superintelligence.

Predicting how long it will take to develop such intelligent machines is difficult. Contrary to what some reviewers of my book seem to believe, I don't have any strong opinion about that matter. (It is as though the only two possible views somebody might hold about the future of artificial intelligence are "machines are stupid and will never live up to the hype!" and "machines are much further advanced than you imagined and true AI is just around the corner!").

A survey of leading researchers in AI suggests that there is a 50% probability that human-level machine intelligence will have been attained by 2050 (defined here as "one that can carry out most human professions at least as well as a typical human"). This doesn't seem entirely crazy. But one should place a lot of uncertainty on both sides of this: it could happen much sooner or very much later.

Exactly how we will get there is also still shrouded in mystery. There are several paths of development that should get there eventually, but we don't know which of them will get there first.

Biological inspiration

We do have an actual example of generally intelligent system – the human brain – and one obvious idea is to proceed by trying to work out how this system does the trick. A full understanding of the brain is a very long way off, but it might be possible to glean enough of the basic computational principles that the brain uses to enable programmers to adapt them for use in computers without undue worry about getting all the messy biological details right.

We already know a few things about the working of the human brain: it is a neural network, it learns through reinforcement learning, it has a hierarchical structure to deal with perceptions and so forth. Perhaps there are a few more basic principles that we still need to discover – and that would then enable somebody to clobber together some form of "neuromorphic AI": one with elements cribbed from biology but implemented in a way that is not fully biologically realistic.

Pure mathematics

Another path is the more mathematical "top-down" approach, which makes little or no use of insights from biology and instead tries to work things out from first principles. This would be a more desirable development path than neuromorphic AI, because it would be more likely to force the programmers to understand what they are doing at a deep level – just as doing an exam by working out the answers yourself is likely to require more understanding than doing an exam by copying one of your classmates' work.

In general, we want the developers of the first human-level machine intelligence, or the first seed AI that will grow up to be superintelligence, to know what they are doing. We would like to be able to prove mathematical theorems about the system and how it will behave as it rises through the ranks of intelligence.

Brute Force

One could also imagine paths that rely more on brute computational force, such by as making extensive use of genetic algorithms. Such a development path is undesirable for the same reason that the path of neuromorphic AI is undesirable – because it could more easily succeed with a less than full understanding of what is being built. Having massive

amounts of hardware could, to a certain extent, substitute for having deep mathematical insight.

We already know of code that would, given sufficiently ridiculous amounts of computing power, instantiate a superintelligent agent. The AIXI model is an example. As best we can tell, it would destroy the world. Thankfully, the required amounts of computer power are physically impossible.

Plagiarising nature

The path of whole brain emulation, finally, would proceed by literally making a digital copy of a particular human mind. The idea would be to freeze or vitrify a brain, chop it into thin slices and feed those slices through an array of microscopes. Automated image recognition software would then extract the map of the neural connections of the original brain. This 3D map would be combined with neurocomputational models of the functionality of the various neuron types constituting the neuropil, and the whole computational structure would be run on some sufficiently capacious supercomputer. This approach would require very sophisticated technologies, but no new deep theoretical breakthrough.

In principle, one could imagine a sufficiently high-fidelity emulation process that the resulting digital mind would retain all the beliefs, desires, and personality of the uploaded individual. But I think it is likely that before the technology reached that level of perfection, it would enable a cruder form of emulation that would yield a distorted human-ish mind. And before efforts to achieve whole brain emulation would achieve even that degree of success, they would probably spill over into neuromorphic AI.

Competent humans first, please

Perhaps the most attractive path to machine superintelligence would be an indirect one, on which we would first enhance humanity's own biological cognition. This could be achieved through, say, genetic engineering along with institutional innovations to improve our collective intelligence and wisdom.

It is not that this would somehow enable us "to keep up with the machines" – the ultimate limits of information processing in machine substrate far exceed those of a biological cortex however far enhanced. The contrary is instead the case: human cognitive enhancement would hasten the day when machines overtake us, since smarter humans would make more rapid progress in computer science. However, it would seem on balance beneficial if the transition to the machine intelligence era were engineered and overseen by a more competent breed of human, even if that would result in the transition happening somewhat earlier than otherwise.

Meanwhile, we can make the most of the time available, be it long or short, by getting to work on the control problem, the problem of how to ensure that superintelligent agents would be safe and beneficial. This would be a suitable occupation for some of our generation's best mathematical talent.

Provided by The Conversation