Superintelligence

Answer to the 2009 EDGE QUESTION: "WHAT WILL CHANGE EVERYTHING?"

Nick Bostrom
www.nickbostrom.com

Intelligence is a big deal. Humanity owes its dominant position on Earth not to any special strength of our muscles, nor any unusual sharpness of our teeth, but to the unique ingenuity of our brains. It is our brains that are responsible for the complex social organization and the accumulation of technical, economic, and scientific advances that, for better and worse, undergird modern civilization.

All our technological inventions, philosophical ideas, and scientific theories have gone through the birth canal of the human intellect. Arguably, human brain power is the chief rate-limiting factor in the development of human civilization.

Unlike the speed of light or the mass of the electron, human brain power is not an eternally fixed constant. Brains can be enhanced. And, in principle, machines can be made to process information as efficiently as — or more efficiently than — biological nervous systems.

There are multiple paths to greater intelligence. By "intelligence" I here refer to the panoply of cognitive capacities, including not just book-smarts but also creativity, social intuition, wisdom, etc.

Let's look first at how we might enhance our biological brains. There are of course the traditional means: education and training, and development of better methodologies and conceptual frameworks. Also, neurological development can be improved through better infant nutrition, reduced pollution, adequate sleep and exercise, and prevention of diseases that affect the brain. We can use biotech to enhance cognitive capacity, by developing pharmaceuticals that improve memory, concentration, and mental energy; or we could achieve these ends with genetic selection and genetic engineering. We can invent external aids to boost our effective intelligence — notepads, spreadsheets, visualization software.

We can also improve our collective intelligence. We can do so via norms and conventions — such as the norm against using ad hominem arguments in scientific discussions — and by improving epistemic institutions such the scientific journal, anonymous peer review, and the patent system. We can increase humanity's joint problem-solving capacity by creating more people or by integrating a greater fraction of the world's existing

population into productive endeavors, and we can develop better tools for communication and collaboration — various internet applications being recent examples.

Each of these ways of enhancing individual and collective human intelligence holds great promise. I think they ought to be vigorously pursued. Perhaps the smartest and wisest thing the human species could do would be to work on making itself smarter and wiser.

In the longer run, however, biological human brains might cease to be the predominant nexus of Earthly intelligence.

Machines will have several advantages: most obviously, faster processing speed — an artificial neuron can operate a million times faster than its biological counterpart. Machine intelligences may also have superior computational architectures and learning algorithms. These "qualitative" advantages, while harder to predict, may be even more important than the advantages in processing power and memory capacity. Furthermore, artificial intellects can be easily copied, and each new copy can — unlike humans — start life fully-fledged and endowed with all the knowledge accumulated by its predecessors. Given these considerations, it is possible that one day we may be able to create "superintelligence": a general intelligence that vastly outperforms the best human brains in every significant cognitive domain.

The spectrum of approaches to creating artificial (general) intelligence ranges from completely unnatural techniques, such as those used in good old-fashioned AI, to architectures modeled more closely on the human brain. The extreme of biological imitation is whole brain emulation, or "uploading". This approach would involve creating a very detailed 3d map of an actual brain — showing neurons, synaptic interconnections, and other relevant detail — by scanning slices of it and generating an image using computer software. Using computational models of how the basic elements operate, the whole brain could then be emulated on a sufficiently capacious computer.

The ultimate success of biology-inspired approaches seems more certain, since they can progress by piecemeal reverse-engineering of the one physical system already known to be capable of general intelligence, the brain. However, some unnatural or hybrid approach might well get there sooner.

It is difficult to predict how long it will take to develop human-level artificial general intelligence. The prospect does not seem imminent. But whether it will take a couple of decades, many decades, or centuries, is probably not something that we are currently in a position to know. We should acknowledge this uncertainty by assigning some non-trivial degree of credence to each of these possibilities.

However long it takes to get from here to roughly human-level machine intelligence, the step from there to superintelligence is likely to be much quicker. In one type of scenario, "the singularity hypothesis", some sufficiently advanced and easily modifiable machine intelligence (a "seed AI") applies its wits to create a smarter version of itself. This smarter version uses its greater intelligence to improve itself even further. The process is iterative, and each cycle is faster than its predecessor. The result is an intelligence explosion. Within some very short period of time — weeks, hours — radical superintelligence is attained.

Whether abrupt and singular, or more gradual and multi-polar, the transition from human-level to superintelligence would of pivotal significance. Superintelligence would be the last invention biological man would ever need to make, since, by definition, it would be much better at inventing than we are. All sorts of theoretically possible technologies could be developed quickly by superintelligence — advanced molecular manufacturing, medical nanotechnology, human enhancement technologies, uploading, weapons of all kinds, lifelike virtual realities, self-replicating space-colonizing robotic probes, and more. It would also be super-effective at creating plans and strategies, working out philosophical problems, persuading and manipulating, and much else beside.

It is an open question whether the consequences would be for the better or the worse. The potential upside is clearly enormous; but the downside includes existential risk. Humanity's future might one day depend on the initial conditions we create, in particular on whether we successfully design the system (e.g., the seed AI's goal architecture) in such a way as to make it "human-friendly" — in the best possible interpretation of that term.