WHY WE NEED FRIENDLY AI

Luke Muehlhauser and Nick Bostrom

Humans will not always be the most intelligent agents on Earth, the ones steering the future. What will happen to us when we no longer play that role, and how can we prepare for this transition?

The human level of intelligence is an evolutionary accident – a small basecamp on a vast mountain side, far below the highest ridges of intelligence allowed by physics. If we were visited by extraterrestrials, these beings would almost certainly be *very much* more intelligent and technologically advanced than we are, and thus our future would depend entirely on the content of *their* goals and desires.

But aliens are unlikely to make contact anytime soon. In the near term, it seems more likely we will *create* our intellectual successors. Computers far outperform humans in many narrow niches (e.g. arithmetic and chess), and there is reason to believe that similar large improvements over human performance are possible for general reasoning and technological development.

Though some doubt that machines can possess certain mental properties like consciousness, the absence of such mental properties would not prevent machines from becoming vastly more able than humans to efficiently steer the future in pursuit of their goals. As Alan Turing wrote, '. . .it seems probable that once the machine thinking method has started, it would not take long to outstrip our feeble powers. . . At some stage therefore we should have to expect the machines to take control. . .'

There is, of course, a risk in passing control of the future to machines, for they may not share our values. This risk is increased by two factors that may cause the transition from

doi:10.1017/S1477175613000316
Think 36, Vol. 13 (Spring 2014)

human control to machine control to be quite sudden and rapid: the possibilities of *computing overhang* and *recursive self-improvement*.

What is computing overhang? Suppose that computing power continues to double according to Moore's law, but figuring out the algorithms for human-like general intelligence proves to be fiendishly difficult. When the software for general intelligence is finally realized, there could exist a 'computing overhang': tremendous amounts of cheap computing power available to run human-level artificial intelligences (AIs). AIs could be copied across the hardware base, causing the AI population to quickly surpass the human population. These digital minds might run thousands or millions of times faster than human minds. AIs might have further advantages, such as superior communication speed, transparency and self-editability, goal coordination, and improved rationality.

And what is recursive self-improvement? We can predict that advanced AIs will have instrumental goals to preserve themselves, acquire resources, and self-improve, because those goals are useful intermediaries to the achievement of almost any set of final goals. Thus, when we build an AI that is as skilled as we are at the task of designing AI systems, we may thereby initiate a rapid, AI-motivated cascade of self-improvement cycles. Now when the AI improves itself, it improves the intelligence that does the improving, quickly leaving the human level of intelligence far behind.

A superintelligent AI might thus quickly become superior to humanity in harvesting resources, manufacturing, scientific discovery, social aptitude, and strategic action, among other abilities. We might not be in a position to negotiate with it or its descendants, just as chimpanzees are not in a position to negotiate with humans.

At the same time, the convergent instrumental goal of acquiring resources poses a threat to humanity, for it means that a superintelligent machine with almost *any* final goal (say, of solving the Riemann hypothesis) would want to take the resources we depend on for its own use. Such

https://doi.org/10.1017/S1477175613000316 Published online by Cambridge University Press

an AI 'does not love you, nor does it hate you, but you are made of atoms it can use for something else'.[1] Moreover, the AI would correctly recognize that humans do not want their resources used for the AI's purposes, and that humans therefore pose a threat to the fulfillment of its goals – a threat to be mitigated however possible.

But because we will create our own successors, we may have the ability to influence their goals and make them friendly to our concerns. The problem of encoding human (or at least *humane*) values into an AI's utility function is a challenging one, but it may be possible. If we can build such a 'Friendly AI,' we may not only avert catastrophe, but also use the powers of machine superintelligence to do enormous good.

Many scientific naturalists accept that machines can be far more intelligent and powerful than humans, and that this could pose a danger for the things we value. Still, they may have objections to the line of thought we have developed so far. Philosopher David Chalmers has responded to many of these objections;[2] we will respond to only a few of them here.

First: why not just keep potentially dangerous AIs safely confined, e.g. without access to the internet? This may sound promising, but there are many complications.[3] In general, such solutions would pit human intelligence against superhuman intelligence, and we shouldn't be confident the former would prevail. Moreover, such methods may only delay AI risk without preventing it. If one AI development team has built a human-level or superintelligent AI and successfully confined it, then other AI development teams are probably not far behind them, and these other teams may not be as cautious. Governments will recognize that human-level AI is a powerful tool, and the race to be the first nation with such a great advantage may incentivize development *speed* over development *safety*. (Confinement measures may, however, be useful as an extra precaution *during the development phase* of safe AI.)

https://doi.org/10.1017/S1477175613000316 Published online by Cambridge University Press

Second: some have suggested that advanced AIs' greater intelligence will cause them to be more moral than we are; in that case, who are we to protest when they do not respect *our* primitive values? That would be downright *immoral!*

Intelligent search for instrumentally optimal plans, however, can be performed in the service of any goal. Intelligence and motivation are in this sense logically orthogonal axes along which possible artificial intellects can vary freely. The imputed connection between intelligence and morality is therefore sheer anthropomorphism. (It is an anthropomorphism that does not even hold true for *humans:* it is easy to find humans who are quite intelligent but immoral, or who are unintelligent but thoroughly decent.)

Economist Robin Hanson suggests that inter-generational conflicts analogous to the ones that could arise between humans and machines are common. Generations old and new compete for resources, and the older generation often wants to control the values of the younger generation. The values of the younger generation end up dominating as the older generation passes away. Must we be so selfish as to insist that the values of *Homo sapiens* dominate the solar system forever?

Along a similar line, the roboticist Hans Moravec once suggested that while we should expect that future robotic corporations will eventually overrun humanity and expropriate our resources, we should think of these robotic descendants as our 'mind children.' Framed in this way, Moravec thought, the prospect might seem more attractive.

It must be said that a scenario in which the children kill and cannibalize their parents is not everyone's idea of a happy family dynamic. But even if we were willing to sacrifice ourselves (and our fellow human beings?) for the sake of some 'greater good,' we would still have to put in hard work to ensure that the result would be something more worthwhile than masses of computer hardware used only to evaluate the Riemann hypothesis (or to calculate the decimals of pi, or to manufacture as many paperclips as possible, or some other arbitrary goal that might be easier to specify than what humans value).

https://doi.org/10.1017/S1477175613000316 Published online by Cambridge University Press

There is, however, one good reason not to insist that superhuman machines be made to share all our current values. Suppose that the ancient Greeks had been the ones to face the transition from human to machine control, and they coded their own values as the machines' final goal. From our perspective, this would have resulted in tragedy, for we tend to believe we have seen moral progress since the Ancient Greeks (e.g. the prohibition of slavery). But presumably we are still far from perfection. We therefore need to allow for continued moral progress.

One proposed solution is to give machines an algorithm for figuring out what our values *would* be if we knew more, were wiser, were more the people we wished to be, and so on. Philosophers have wrestled with this approach to the theory of values for decades, and it may be a productive solution for machine ethics.

Third: others object that we are too far from the transition from human to machine control to work on the problem now. But we must remember that economic incentives favor development speed over development safety. Moreover, our scientific curiosity can sometimes overwhelm other considerations such as safety. To quote J. Robert Oppenheimer, the physicist who headed the Manhattan Project: 'When you see something that is technically sweet, you go ahead and do it and you argue about what to do about it only after you have had your technical success. That is the way it was with the atomic bomb.'[4]

Still, one might ask: What can we do about the problem of AI risk when we know so little about the design of future AIs? For a start, we can do the kind of work currently performed by the two research institutes currently working most directly on this difficult problem: the Machine Intelligence Research Institute in Berkeley and the Future of Humanity Institute at Oxford University. This includes:

1. *Strategic research*. Which types of technological development are risk-increasing or risk-decreasing, and how can we encourage

https://doi.org/10.1017/S1477175613000316 Published online by Cambridge University Press

governments and corporations to shift funding from the former to the latter? What is the expected value of certain kinds of research, or of certain kinds of engagement with governments and the public? What can we do to reduce the risk of an AI arms race? How does AI risk compare to risks from nuclear weapons, biotechnology, near earth objects, etc.? Can economic models predict anything about the impact of AI technologies? Can we develop technological forecasting methods capable of giving advance warning of the invention of AI?

2. *Technical research*. Can we develop safe confinement methods for powerful AIs? How can an agent with a desirable-to-humans utility function maintain its desirable goals during updates to the ontology over which it has preferences? How can we extract a coherent utility function from inconsistent human behavior, and use it to inform an AI's own utility function? Can we develop an advanced AI that will answer questions but not manifest the dangerous capacities of a superintelligent agent?
3. *Raising awareness*. Outreach to researchers, philanthropists, and the public can attract more monetary and human capital with which to attack the problem.

The quest for AI has come a long way. Computer scientists and other researchers must begin to take the implications of AI more seriously.

Luke Muehlhauser is Executive Director of the Machine Intelligence Research Institute. Nick Bostrom is Director of the Future of Humanity Institute at the University of Oxford. luke@intelligence.org

https://doi.org/10.1017/S1477175613000316 Published online by Cambridge University Press

Notes

[1] E. Yudkowsky, 'AI as a positive and a negative factor in global risk', *Global Catastrophic Risks* (eds) N. Bostrom and M. Cirkovic (New York: Oxford University Press, 2008).

[2] D. Chalmers, 'The Singularity: a reply to commentators', *Journal of Consciousness Studies*, vol. 19, nos. 7–8 (2012), 141–167.

[3] S. Armstrong, A. Sandberg, N. Bostrom, 'Thinking inside the box: Using and controlling Oracle AI', *Minds and Machines*, vol. 22, no. 4 (2012), 299–324.

[4] Robert Jungk, *Brighter than a Thousand Suns: A Personal History of the Atomic Scientists*, trans. Lames Cleugh (New York: Harcourt Harvest, 1958), 296.

https://doi.org/10.1017/S1477175613000316 Published online by Cambridge University Press