Reducing Long-Term Catastrophic Risks from Artificial Intelligence

Eliezer Yudkowsky, Anna Salamon
Machine Intelligence Research Institute

Carl Shulman, Steven Kaas, Tom McCabe
MIRI Visiting Fellows

Rolf Nelson

Abstract

In 1965, I. J. Good proposed that machines would one day be smart enough to make themselves smarter. Having made themselves smarter, they would spot still further opportunities for improvement, quickly leaving human intelligence far behind (Good 1965). He called this the "intelligence explosion." Later authors have called it the "technological singularity" or simply "the Singularity" (Kurzweil 2005; Vinge 1993).

The Singularity Institute aims to reduce the risk of a catastrophe resulting from an intelligence explosion. We do research, education, and conferences. In this paper, we make the case for taking artificial intelligence (AI) risks seriously, and suggest some strategies to reduce those risks.

Yudkowsky, Eliezer, Carl Shulman, Anna Salamon, Rolf Nelson, Steven Kaas, Steve Rayhawk and Tom McCabe. 2010. *Reducing Long-Term Catastrophic Risks from Artificial Intelligence.*
The Singularity Institute, San Francisco, CA.

The Machine Intelligence Research Institute was previously known as the Singularity Institute.

1. What We're (Not) About

The Singularity Institute doesn't know exactly when the intelligence explosion will occur, but we'd like to figure out how to make its consequences good rather than bad. We do not see ourselves as having the job of foretelling that it will go well or poorly. If the outcome were predetermined there would be no point in trying to intervene.

We suspect that AI is primarily a software problem that will require new insight, not a hardware problem that will fall to Moore's Law. We are interested in rational analyses of AI risks, not storytelling.

2. Indifference, Not Malice

Notions of a "robot rebellion," in which AIs spontaneously develop primate-like resentment for their low tribal status, are the stuff of science fiction. The more plausible danger stems not from malice, but from the fact that human survival requires scarce resources: resources for which AIs may have other uses (Omohundro 2008, 2007). Superintelligent AIs with access to pervasive data networks and autonomous robotics could radically alter their environment. For example, they could harvest all available solar, chemical, and nuclear energy. If such AIs found uses for this energy that better furthered their goals than supporting human life, human survival would become unlikely.

Many AIs will converge toward a tendency to maximize some goal (Omohundro 2008). For instance, AIs developed under evolutionary pressures would be selected for values that maximized reproductive fitness, and would prefer to allocate resources to reproduction rather than to supporting humans (Bostrom 2004). Such unsafe AIs might actively mimic safe benevolence until they became powerful, since being destroyed would prevent them from working toward their goals. Thus, a broad range of AI designs may initially appear safe, but if developed to the point of an intelligence explosion could cause human extinction in the course of optimizing the Earth for their goals.

3. An Intelligence Explosion May Be Sudden

The pace of an intelligence explosion depends on two conflicting pressures. Each improvement in AI technology increases the ability of AIs to research more improvements, but an AI may also face the problem of diminishing returns as the easiest improvements are achieved first.

The rate of improvement is hard to estimate, but several factors suggest it would be high. The predominant view in the AI field is that the bottleneck for powerful AI is software, not hardware. Continued rapid hardware progress is expected in coming

decades (ITRS 2007). If and when the powerful AI software is developed, there may by that time be a glut of hardware available to run many copies of AIs, and to run them at high speeds. This could amplify the effects of AI improvements (Hanson, forthcoming).

Humans are not optimized for intelligence. Rather, we are the first and possibly dumbest species capable of producing a technological civilization. The first AI with humanlike AI research abilities might be able to reach superintelligence rapidly—in particular, more rapidly than researchers and policy-makers can develop adequate safety measures.

4. Is Concern Premature?

We don't know how to build an AI with human-level intelligence, so we can't have much confidence that it will arrive in the next few decades. But we also can't rule out unforeseen advances. Past underestimates of the difficulty of AI (perhaps most infamously, those made for the 1956 Dartmouth Conference [McCarthy et al. 1955]) do not guarantee that AI will never succeed. We need to take into account both repeated discoveries that the problem is more difficult than expected and incremental progress in the field. Advances in AI and machine learning algorithms (Russell and Norvig 2010), increasing R&D expenditures by the technology industry, hardware advances that make computation-hungry algorithms feasible (ITRS 2007), enormous datasets (Halevy, Norvig, and Pereira 2009), and insights from neuroscience give us advantages that past researchers lacked. Given the size of the stakes and the uncertainty about AI timelines, it seems best to allow for the possibility of medium-term AI development in our safety strategies.

5. Friendly AI

Concern about the risks of future AI technology has led some commentators, such as Sun co-founder Bill Joy, to suggest the global regulation and restriction of such technologies (Joy 2000). However, appropriately designed AI could offer similarly enormous benefits.

An AI smarter than humans could help us eradicate diseases, avert long-term nuclear risks, and live richer, more meaningful lives. Further, the prospect of those benefits along with the competitive advantages from AI would make a restrictive global treaty difficult to enforce.

The Singularity Institute's primary approach to reducing AI risks has thus been to promote the development of AI with benevolent motivations that are reliably stable under self-improvement, what we call "Friendly AI" (Yudkowsky 2008a).

To very quickly summarize some of the key ideas in Friendly AI:

1. We can't make guarantees about the final outcome of an agent's interaction with the environment, but we may be able to make guarantees about what the agent is trying to do, given its knowledge. We can't determine that Deep Blue will win against Kasparov just by inspecting Deep Blue, but an inspection might reveal that Deep Blue searches the game tree for winning positions rather than losing ones.

2. Because code executes on the almost perfectly deterministic environment of a computer chip, we may be able to make strong guarantees about an agent's motivations (including how that agent rewrites itself), even though we can't logically prove the outcomes of particular tactics chosen. This is important, because if the agent fails with a tactic, it can update its model of the world and try again. But during self-modification, the AI may need to implement a million code changes, one after the other, without any of them having catastrophic effects.

3. Gandhi doesn't want to kill people. If someone offers Gandhi a pill that he knows will alter his brain to make him want to kill people, then Gandhi will likely refuse to take the pill. In the same way, most utility functions should be stable under reflection, provided that the AI can correctly project the result of its own self-modifications. Thus, the problem of Friendly AI is not in creating an extra conscience module that constrains the AI despite its preferences. Rather, the challenge is in reaching into the enormous design space of possible minds and selecting an AI that prefers to be Friendly.

4. Human terminal values are extremely complicated. This complexity is not introspectively visible at a glance. The solution to this problem may involve designing an AI to learn human values by looking at humans, asking questions, scanning human brains, etc., rather than an AI preprogrammed with a fixed set of imperatives that sounded like good ideas at the time.

5. The explicit moral values of human civilization have changed over time, and we regard this change as progress. We also expect that progress may continue in the future. An AI programmed with the explicit values of 1800 might now be fighting to reestablish slavery. Static moral values are clearly undesirable, but most random changes to values will be even less desirable. Every improvement is a change, but not every change is an improvement. Perhaps we could program the AI to "do what we would have told you to do if we knew everything you know" and "do what we would've told you to do if we thought as fast as you do and could consider many more possible lines of moral argument" and "do what we would tell you to do if

we had your ability to reflect on and modify ourselves." In moral philosophy, this approach to moral progress is known as reflective equilibrium (Rawls 1971).

6. Seeding research programs

As we get closer to advanced AI, it will be easier to learn how to reduce risks effectively. The interventions to focus on today are those whose benefits will compound over time. Possibilities include:

Friendly AI: Theoretical computer scientists can investigate AI architectures that self-modify while retaining stable goals. Theoretical toy systems exist now: Gödel machines make provably optimal self-improvements given certain assumptions (Schmidhuber 2007). Decision theories are being proposed that aim to be stable under self-modification (Drescher 2006). These models can be extended incrementally into less idealized contexts.

Stable brain emulations: One route to safe AI may start with human brain emulation. Neuroscientists can investigate the possibility of emulating the brains of individual humans with known motivations, while evolutionary theorists can investigate methods to prevent dangerous evolutionary dynamics, and social scientists can investigate social or legal frameworks to channel the impact of emulations in positive directions (Sandberg and Bostrom 2008).

Models of AI risks: Researchers can build models of AI risks and of AI growth trajectories, using tools from game theory, evolutionary analysis, computer security, or economics (Bostrom 2004; Hall 2007; Hanson, forthcoming; Omohundro 2007; Yudkowsky 2008a). If such analysis is done rigorously it can help to channel the efforts of scientists, graduate students, and funding agencies to the areas with the greatest potential benefits.

Institutional improvements: Major technological risks are ultimately navigated by society as a whole. Success requires that society understand and respond to scientific evidence. Knowledge of the biases that distort human thinking around catastrophic risks (Yudkowsky 2008b), improved methods for probabilistic forecasting (Rayhawk et al. 2009) or risk analysis (Matheny 2007), and methods for identifying and aggregating expert opinions (Hanson 1996) can all improve the odds of a positive Singularity. So can methods for international cooperation around AI development, and for avoiding an AI "arms race" that might be won by the competitor most willing to trade off safety measures for speed (Shulman 2009).

7. Our Aims

We aim to seed the above research programs. We are too small to carry out all the needed research ourselves, but we can get the ball rolling.

We have groundwork already. We have: (a) seed research about catastrophic AI risks and AI safety technologies; (b) human capital; and (c) programs that engage outside research talent, including our annual Singularity Summits and our Visiting Fellows program.

Going forward, we plan to continue our recent growth by scaling up our visiting fellows program, extending the Singularity Summits and similar academic networking, and writing further papers to seed the above research programs, in-house or with the best outside talent we can find. We welcome potential co-authors, Visiting Fellows, and other collaborators, as well as any suggestions or cost-benefit analyses on how to reduce catastrophic AI risk.

8. The Upside and Downside of Artificial Intelligence

Human intelligence is the most powerful known biological technology. But our place in history probably rests not on our being the smartest intelligences that could exist, but rather on being the first intelligences that did exist. We probably are to intelligence what the first replicator was to biology. The first single-stranded RNA capable of copying itself was not a sophisticated, robust replicator—but it still had an important place in history, due to being first.

The future of intelligence is, hopefully, *much* greater than its past. The origin and shape of human intelligence may end up playing a critical role in the origin and shape of future civilizations on a much larger scale than one planet. And the origin and shape of the first self-improving Artificial Intelligences humanity builds may have a similarly large impact, for similar reasons. The values of future intelligences will shape future civilizations. What stands to be won or lost are the values of future intelligences, and thus the values of future civilizations.

9. Recommended Reading

This has been a very quick introduction. For more information, please contact louie@intelligence.org or see:

For a general overview of AI catastrophic risks: Yudkowsky (2008a).

For discussion of self-modifying systems' tendency to approximate optimizers and fully exploit scarce resources: Omohundro (2008).

For discussion of evolutionary pressures toward software minds aimed solely at reproduction: Bostrom (2004).

For tools for doing cost-benefit analysis on human extinction risks, and a discussion of gaps in the current literature: Matheny (2007).

For an overview of potential causes of human extinction, including AI: Bostrom (2002).

For an overview of the ethical problems and implications involved in creating a superintelligent AI: Bostrom (2003).

References

Bostrom, Nick. 2002. "Existential Risks: Analyzing Human Extinction Scenarios and Related Hazards." *Journal of Evolution and Technology* 9. http://www.jetpress.org/volume9/risks.html.

———. 2003. "Ethical Issues in Advanced Artificial Intelligence." In *Cognitive, Emotive and Ethical Aspects of Decision Making in Humans and in Artificial Intelligence,* edited by Iva Smit and George E. Lasker, 12–17. Vol. 2. Windsor, ON: International Institute for Advanced Studies in Systems Research / Cybernetics.

———. 2004. "The Future of Human Evolution." In *Two Hundred Years After Kant, Fifty Years After Turing,* edited by Charles Tandy, 339–371. Vol. 2. Death and Anti-Death. Palo Alto, CA: Ria University Press.

Bostrom, Nick, and Milan M. Ćirković, eds. 2008. *Global Catastrophic Risks.* New York: Oxford University Press.

Drescher, Gary L. 2006. *Good and Real: Demystifying Paradoxes from Physics to Ethics.* Bradford Books. Cambridge, MA: MIT Press.

Good, Irving John. 1965. "Speculations Concerning the First Ultraintelligent Machine." In *Advances in Computers,* edited by Franz L. Alt and Morris Rubinoff, 31–88. Vol. 6. New York: Academic Press. doi:10.1016/S0065-2458(08)60418-0.

Halevy, Alon, Peter Norvig, and Fernando Pereira. 2009. "The Unreasonable Effectiveness of Data." *IEEE Intelligent Systems* 24 (2): 8–12. doi:10.1109/MIS.2009.36.

Hall, John Storrs. 2007. *Beyond AI: Creating the Conscience of the Machine.* Amherst, NY: Prometheus Books.

Hanson, Robin. 1996. "Idea Futures." Unpublished manuscript, June 12. Accessed May 20, 2012. http://hanson.gmu.edu/ideafutures.html.

———. Forthcoming. "Economic Growth Given Machine Intelligence." *Journal of Artificial Intelligence Research.* Preprint at. http://hanson.gmu.edu/aigrow.pdf.

ITRS. 2007. *International Technology Roadmap for Semiconductors: 2007 Edition.* International Technology Roadmap for Semiconductors. http://www.itrs.net/Links/2007ITRS/Home2007.htm.

Joy, Bill. 2000. "Why the Future Doesn't Need Us." *Wired,* April. http://www.wired.com/wired/archive/8.04/joy.html.

Kurzweil, Ray. 2005. *The Singularity Is Near: When Humans Transcend Biology.* New York: Viking.

Matheny, Jason G. 2007. "Reducing the Risk of Human Extinction." *Risk Analysis* 27 (5): 1335–1344. doi:10.1111/j.1539-6924.2007.00960.x.

McCarthy, John, Marvin Minsky, Nathan Rochester, and Claude Shannon. 1955. *A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence.* Formal Reasoning Group, Stanford University, Stanford, CA, August 31.

Omohundro, Stephen M. 2007. "The Nature of Self-Improving Artificial Intelligence." Paper presented at Singularity Summit 2007, San Francisco, CA, September 8–9. http://intelligence.org/summit2007/overview/abstracts/#omohundro.

———. 2008. "The Basic AI Drives." In *Artificial General Intelligence 2008: Proceedings of the First AGI Conference,* edited by Pei Wang, Ben Goertzel, and Stan Franklin, 483–492. Frontiers in Artificial Intelligence and Applications 171. Amsterdam: IOS.

Rawls, John. 1971. *A Theory of Justice.* Cambridge, MA: Belknap.

Rayhawk, Stephen, Anna Salamon, Thomas McCabe, Michael Anissimov, and Rolf Nelson. 2009. "Changing the Frame of AI Futurism: From Storytelling to Heavy-Tailed, High-Dimensional Probability Distributions." Paper presented at the 7th European Conference on Computing and Philosophy (ECAP), Bellaterra, Spain, July 2–4.

Russell, Stuart J., and Peter Norvig. 2010. *Artificial Intelligence: A Modern Approach.* 3rd ed. Upper Saddle River, NJ: Prentice-Hall.

Sandberg, Anders, and Nick Bostrom. 2008. *Whole Brain Emulation: A Roadmap.* Technical Report, 2008-3. Future of Humanity Institute, University of Oxford. `http://www.fhi.ox.ac.uk/Reports/2008-3.pdf`.

Schmidhuber, Jürgen. 2007. "Gödel Machines: Fully Self-Referential Optimal Universal Self-Improvers." In *Artificial General Intelligence,* edited by Ben Goertzel and Cassio Pennachin, 199–226. Cognitive Technologies. Berlin: Springer. doi:`10.1007/978-3-540-68677-4_7`.

Shulman, Carl. 2009. "Arms Control and Intelligence Explosions." Paper presented at the 7th European Conference on Computing and Philosophy (ECAP), Bellaterra, Spain, July 2–4.

Vinge, Vernor. 1993. "The Coming Technological Singularity: How to Survive in the Post-Human Era." In *Vision-21: Interdisciplinary Science and Engineering in the Era of Cyberspace,* 11–22. NASA Conference Publication 10129. NASA Lewis Research Center. `http://ntrs.nasa.gov/archive/nasa/casi.ntrs.nasa.gov/19940022855_1994022855.pdf`.

Yudkowsky, Eliezer. 2008a. "Artificial Intelligence as a Positive and Negative Factor in Global Risk." In Bostrom and Ćirković 2008, 308–345.

———. 2008b. "Cognitive Biases Potentially Affecting Judgment of Global Risks." In Bostrom and Ćirković 2008, 91–119.