From Mostly Harmless to Civilization-Threatening: Pathways to Dangerous Artificial Intelligences

Kaj Sotala
MIRI Visiting Fellow

Abstract

The term "superintelligence" comes up frequently in discussions about the Singularity. This is loosely understood to mean a general intelligence vastly greater than that of humans, and therefore disproportionately powerful (Vinge 1993; Bostrom 1998; Yudkowsky 2008), but the exact details of how this might be achieved are frequently left open. This paper seeks to outline some possible paths to superintelligence, and therefore attempts to estimate the degree to which we should actually be worried about artificial intelligences.

The main reason to be worried about greater-than-human intelligence is because it is hard for humans to anticipate and control. Keeping this in mind, we also cover non-intelligence-related factors that might advantage an AI over human decision-makers. We discuss four categories of such advantages: hardware advantages, self-improvement

Sotala, Kaj. 2010. "From Mostly Harmless to Civilization-Threatening: Pathways to Dangerous Artificial Intelligences." In *ECAP10: VIII European Conference on Computing and Philosophy*, edited by Klaus Mainzer. Munich: Dr. Hut.

This version contains minor changes.

and architectural advantages, other advantages from software, and human handicaps. We also try to estimate how quickly such advantages may be relevant. Several methods outlined here also apply to other digital intelligences, such as human brain emulations (Sandberg and Bostrom 2008).

1. Hardware Advantages

An AI running on a system with more processing power might think faster than humans. This is particularly relevant in crisis situations, where everything may hinge on rapidly made decisions, where a faster rate of thought will also allow for more effective long-term planning and scientific discovery. The human brain is estimated to carry out about 10^{11} OPS worth of calculations (Moravec 1998), while the laws of physics theoretically allow for computers in the 10^{40} OPS range (Lloyd 2000).

How likely is a near-term hardware advantage? A full software emulation of the human brain, implementing biological details, is estimated to require somewhere in the region of 10^{18} to 10^{25} FLOPS and to be doable within the near future with computing power as the main constraint (Sandberg and Bostrom 2008). This amount of computing power is estimated to become available for $1 million between 2019 and 2044. Since the brain is estimated to carry out 10^{11} OPS worth of calculations, a full biological replica seems unnecessarily wasteful. If researchers studied simulated models of the brain and then learned to abstract away the principles to more efficient algorithms, we might suddenly have AI systems that could think twenty five years' worth of thoughts in about eight seconds. The more advanced our hardware gets before we have digital intelligences, the more of an advantage they will have over us once they become operational.

A possible limitation is that advances in computing power have been increasingly parallel. A future computer may not be able to use its superior processing power to gain a direct increase in speed over the human brain if the tasks in question do not parallelize well. Yet increases in computing power may increase the amount of data that can be processed at once. For many problems, the easily parallelizable part is the one that grows as data is added, and the serial part remains constant (Gustafson 1988). The parallel nature of the human brain implies that general intelligence does parallelize well.

Human performance in a variety of domains is also correlated with a general intelligence factor, *g* (Gottfredson 1997), theorized to be related to working memory capacity (Oberauer et al. 2008). It seems that differences in *g* can to some degree be predicted from differences in what might correspond to computing power and memory. For instance, the size of the brain correlates with *g* (McDaniel 2005). The neural efficiency hypothesis, which has been supported by research, suggests that people with a higher *g* need to employ less neural resources for individual tasks (Micheloyannis et al. 2006). An AI might have a working memory equivalent far surpassing that of humans.

2. Self-Improvement and Architectural Advantages

Human intelligence might also be improved in a qualitative manner. Several biases and failures of reasoning have been identified in the heuristics and biases literature (see for instance Tversky and Kahneman [1986]; or, for a more recent overview see Stanovich [2008]). Such failures of reasoning have an enormous negative impact on society. Among other things they cause people to suffer from a worse standard of living due to status quo bias, make bad investments, become more easily manipulated, end up falsely accused by the authorities or even imprisoned, make bad decisions leading to an increased death rate, or even fall prey to scams serious enough to crash a national economy (Stanovich 2008). A mind that was immune to such biases would reason more reliably than we do, while possibly exploiting our biases.

Human biases can be looked at either as *ad hoc* heuristics that fail to reason correctly in a modern environment, or as satisficing algorithms that do the best possible job given human computational resources (Gigerenzer and Brighton 2009). An AI could potentially overcome most if not all of the biases that plague human reasoning, either by rewriting its algorithms to better suit the environment or to better take into account growing computational resources. One's suspectibility to several other biases correlates negatively with one's general intelligence, suggesting that computational limitations cause at least some of the flaws in human reasoning (Stanovich and West 2000).

Considering the amount of cognitive flaws we have, even better than human reasoning might be done using suboptimal algorithms. Improving the algorithms might allow for pure speed advantage, but it might also allow for qualitative improvements. For instance, an ability to visualize things in 10,000 dimensions might make some mathematical results easier to understand and build intuitions on. This might allow an AI to think thoughts that we are literally incapable of thinking, and therefore develop strategies we never could.

It is not clear how susceptible the first AIs would be to human biases, nor how easy it would be for them to self-improve to get rid of them. It might be that the very first AIs could be programmed with all of these advantages from the start, or they might be plagued with even more severe limitations and require much time and effort to improve. It needs to be noted that for an AI doing self-improvement, each improvement in reasoning capability could spark off further improvements, resulting in a chain reaction that might or might not "go critical" and lead to an intelligence much greater than a human's (Yudkowsky 2008).

3. Other Advantages from Software

Humans are limited by the fact that they can only be in one place and do one thing at a time, but copyable workers could rapidly come to dominate major portions of the economy (Hanson 1994, 2008). An AI might spawn a number of copies of itself, each copy constantly exchanging information with the other copies. This exchange of information might be more comparable to the way that different parts of our brain communicate with each other, rather than the way human individuals communicate with each other.

The appropriate analogy for an AI might therefore not be that of a single human genius pitted against the whole rest of humanity, but that of an entire society of agents working in perfect coordination. The feasibility of this again depends on hardware trends and the amount of computing power an instance of the AI needs. An AI might simply buy large amounts of hardware, or acquire processing resources illegally. Botnets are networks of computers that have been compromised by outside attackers and are used for illegitimate purposes. Estimates range from one study saying the effective sizes of botnets rarely exceed a few thousand bots, to a study saying that botnet sizes can reach 350,000 members (Rajab et al. 2007). Modern top-of-the-line personal computers can reach 10^{11} FLOPS (Shah 2009). Currently, the distributed computing project Folding@home, with 290,000 active clients, can reach speeds in the 10^{15} FLOPS range (Folding@home 2010). The amount of coordination that can be done also depends on the bandwidth available, but the requirements for this are difficult to estimate.

4. Human Handicaps

People have a demonstrated tendency to think of the capabilities of minds unlike themselves as if they were humans, even if explicitly instructed otherwise (Barrett and Keil 1996). The intuitive faculties we employ for understanding others work on the assumption that we're modeling other humans. The neural systems we use for modeling others overlap with those related to self-related processing (Uddin et al. 2007). An AI with a different cognitive architecture from ours would be difficult or even impossible to intuitively model. The difficulty would likely be mutual at first, but with time the AI could self-improve to have customized cognitive modules for modeling humans.

5. Conclusion

The above analysis suggests that an artificial intelligence can become close to impossible for humanity to effectively control. Improving hardware poses a serious risk for such

attempts, for it provides clear advantages as well as making various software advantages stronger.

It has been argued (Yudkowsky 2001, 2008) that we need a firm theoretical grounding for building safe AIs. Hard to control AIs are a risk, because even seemingly benign goals can soon become contrary to humanity's interests (Omohundro 2008). An AI does not need to be outright hostile towards humanity to be a threat: it might simply have a need for our resources (Yudkowsky 2008; Omohundro 2008). If we cannot control an agent bent on confiscating our resources, we might very quickly end up without them. It seems clear that caution is warranted.

References

Barrett, Justin L., and Frank C. Keil. 1996. "Conceptualizing a Nonnatural Entity: Anthropomorphism in God Concepts." *Cognitive Psychology* 31 (3): 219–247. doi:10.1006/cogp.1996.0017.

Bostrom, Nick. 1998. "How Long Before Superintelligence?" *International Journal of Futures Studies* 2.

Folding@home. 2010. "Client Statistics by OS." Stanford University. Accessed August 10, 2010. http://fah-web.stanford.edu/cgi-bin/main.py?qtype=osstats.

Gigerenzer, Gerd, and Henry Brighton. 2009. "Homo Heuristicus: Why Biased Minds Make Better Inferences." *Topics in Cognitive Science* 1 (1): 107–143. doi:10.1111/j.1756-8765.2008.01006.x.

Gottfredson, Linda S. 1997. "Why g Matters: The Complexity of Everyday Life." In "Intelligence and Social Policy." Special issue, *Intelligence* 24 (1): 79–132. doi:10.1016/S0160-2896(97)90014-3.

Gustafson, John L. 1988. "Reevaluating Amdahl's Law." *Communications of the ACM* 31 (5): 532–533. doi:10.1145/42411.42415.

Hanson, Robin. 1994. "If Uploads Come First: The Crack of a Future Dawn." *Extropy* 6 (2). http://hanson.gmu.edu/uploads.html.

———. 2008. "Economics of the Singularity." *IEEE Spectrum* 45 (6): 45–50. doi:10.1109/MSPEC.2008.4531461.

Lloyd, Seth. 2000. "Ultimate Physical Limits to Computation." *Nature* 406 (6799): 1047–1054. doi:10.1038/35023282.

McDaniel, Michael A. 2005. "Big-Brained People are Smarter: A Meta-Analysis of the Relationship between In Vivo Brain Volume and Intelligence." *Intelligence* 33 (4): 337–346. doi:10.1016/j.intell.2004.11.005.

Micheloyannis, Sifis, Ellie Pachou, Cornelis J. Stam, Michael Vourkas, Sophia Erimaki, and Vasso Tsirka. 2006. "Using Graph Theoretical Analysis of Multi Channel EEG to Evaluate the Neural Efficiency Hypothesis." *Neuroscience Letters* 402 (3): 273–277. doi:10.1016/j.neulet.2006.04.006.

Moravec, Hans P. 1998. "When Will Computer Hardware Match the Human Brain?" *Journal of Evolution and Technology* 1. http://www.transhumanist.com/volume1/moravec.htm.

Oberauer, Klaus, Heinz-Martin Süß, Oliver Wilhelm, and Werner W. Wittmann. 2008. "Which Working Memory Functions Predict Intelligence?" *Intelligence* 36 (6): 641–652. doi:10.1016/j.intell.2008.01.007.

Omohundro, Stephen M. 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.

Rajab, Moheeb Abu, Jay Zarfoss, Fabian Monrose, and Andreas Terzis. 2007. "My Botnet is Bigger than Yours (Maybe, Better than Yours): Why Size Estimates Remain Challenging." In *Proceedings of 1st Workshop on Hot Topics in Understanding Botnets (HotBots '07).* Berkeley, CA: USENIX. http://static.usenix.org/event/hotbots07/tech/full_papers/rajab/rajab.pdf.

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.

Shah, Agam. 2009. "Nvidia Closing in on 2 Teraflops with Graphics Card." *Computerworld*, January 8. http://www.computerworld.com/s/article/9125345/Nvidia_closing_in_on_2_teraflops_with_graphics_card.

Stanovich, Keith E. 2008. *What Intelligence Tests Miss: The Psychology of Rational Thought*. New Haven, CT: Yale University Press.

Stanovich, Keith E., and Richard F. West. 2000. "Individual Differences in Reasoning: Implications for the Rationality Debate?" *Behavioral and Brain Sciences* 23 (5): 645–665. http://journals.cambridge.org/abstract_S0140525X00003435.

Tversky, Amos, and Daniel Kahneman. 1986. "Rational Choice and the Framing of Decisions." In "The Behavioral Foundations of Economic Theory." Supplement, *Journal of Business* 59 (4, pt. 2): S251–S278. http://www.jstor.org/stable/2352759.

Uddin, Lucina Q., Marco Iacoboni, Claudia Lange, and Julian Paul Keenan. 2007. "The Self and Social Cognition: The Role of Cortical Midline Structures and Mirror Neurons." *TRENDS in Cognitive sciences* 11 (4): 153–157. doi:10.1016/j.tics.2007.01.001.

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. 2001. *Creating Friendly AI 1.0: The Analysis and Design of Benevolent Goal Architectures.* The Singularity Institute, San Francisco, CA, June 15. http://intelligence.org/files/CFAI.pdf.

———. 2008. "Artificial Intelligence as a Positive and Negative Factor in Global Risk." In *Global Catastrophic Risks*, edited by Nick Bostrom and Milan M. Ćirković, 308–345. New York: Oxford University Press.