

From mostly harmless to civilization-threatening: pathways to dangerous artificial general intelligences

Kaj Sotala

Singularity Institute for Artificial Intelligence

Cognitive Science Unit, Department of Psychology, University of Helsinki
kaj.sotala@singinst.org

Paper presented at the VIII European Conference on Philosophy and Computing (ECAP10).

Introduction. 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 general intelligences (AGIs).

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 AGI over human decision-makers. We discuss four categories of such advantages: hardware advantages, self-improvement 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 & Bostrom, 2008).

Hardware advantages. An AGI 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, but 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 & 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 AGI systems that could think 25 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.

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, Süß, Wilhelm, & Wittmann, 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 AGI might have a working memory equivalent far surpassing that of humans.

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 & Kahneman (1986); or Stanovich (2009) for a more recent overview). Such failures of reasoning have 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, 2009). 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 jury-rigged 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 & Brighton, 2009). An AGI 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 susceptibility 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 & West, 2000).

Considering the amount of cognitive flaws we have, even correct reasoning might be done using suboptimal algorithms. Improving on them 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 AGI to think thoughts that we were literally incapable of thinking, and therefore develop strategies we never could.

It is not clear how susceptible the first AGIs 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 AGIs could be programmed with all of these advantages from the start, or they might be plagued with even more severe limitations and require very much time and effort to improve. It needs to be noted that for an AGI 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).

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; Hanson, 2008). An AGI 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 AGI 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 AGI needs. An AGI might simply buy large numbers of hardware, or acquire it 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 & Zarfoss & Monrose & Terzis, 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 (Pande Lab, 2010). The amount of coordination that can be done also depends on the bandwidth available, but the requirements for this are difficult to estimate.

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 (Barret & 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, Iacoboni, Lange & Keenan, 2007). An AGI with a different cognitive architecture from ours would be impossible to intuitively model. The difficulty would likely be mutual at first, but with time the AGI could self-improve to have customized cognitive modules for modeling humans.

Conclusion. The above analysis suggests that an AGI 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; Yudkowsky 2008) that we need a firm theoretical grounding for building safe AGIs. Hard to control AGIs are a risk, because even seemingly benign goals can soon become contrary to humanity's interests (Omohundro, 2008). An AGI does not need to be outright hostile towards humanity to be a threat: it might simply have a need for our resources (Yudkowsky, 2006; 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.

- Barret, J.L. & Keil, F.C. (1996) Conceptualizing a Nonnatural Entity: Anthropomorphism in God Concepts. *Cognitive Psychology* 31, 219–247.
- Bostrom, N. (1998) How long before superintelligence? *Int. Jour. of Future Studies*, vol. 2.
- Gigerenzer, G. & Brighton, H. (2009) Homo Heuristicus: Why Biased Minds Make Better Inferences. *Topics in Cognitive Science* 1, 107-143.
- Gottfredson, L. (1997) Why g Matters. *Intelligence*, 24, 79-132.
- Gustafson, J.L. (1988) Reevaluating Amdahl's Law. *Communications of the ACM*, 31(5), 532–533.
- Hanson, R. (1994) If uploads come first - The crack of a future dawn. *Extropy* 6(2). Retrieved from <http://hanson.gmu.edu/uploads.html>
- Hanson, R. (2008) Economics of the Singularity. *IEEE Spectrum*, 45(6), 45-50.
- Lloyd, S. (2000) Ultimate physical limits to computation. *Nature* 406, 1047-1054.
- McDaniel, M.A. (2005) Big-brained people are smarter: A meta-analysis of the relationship between in vivo brain volume and intelligence. *Intelligence*, 33, 337-346.
- Micheloyannis, S., Pachou, E. Cornelis, J.S., Vourkas, M., Erimaki, S. & Tsirka, V. (2006) Using graph theoretical analysis of multi channel EEG to evaluate the neural efficiency hypothesis. *Neuroscience Letters*, 402(3), 273-277.
- Moravec, Hans. (1998) "When Will Computer Hardware Match the Human Brain?" *Journal of Evolution and Technology*, vol. 1. Retrieved from <http://www.transhumanist.com/volume1/moravec.htm>
- Oberauer, K., Süß, H-M., Wilhelm, C. & Wittmann, W.W. (2008) Which working memory functions predict intelligence? *Intelligence*, 36, 641-652.

Omohundro, S. (2008) The basic AI drives. P. Wang, B. Goertzel & S. Franklin (Eds.), of the First AGI Conference. *Frontiers in Artificial Intelligence and Applications*, Volume 171 (pp. 483-494) IOS Press.

Pande Lab (2010) Client statistics by OS. Retrieved August 10, 2010, from <http://fah-web.stanford.edu/cgi-bin/main.py?qttype=osstats>

Rajab, M.A., Zarfoss, J., Monroe, F. & Terzis, A. (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)*.

Sandberg, A. & Bostrom, N. (2008) *Whole Brain Emulation: A Roadmap*. Technical Report #2008-3. Future of Humanity Institute, Oxford University. Retrieved from http://www.philosophy.ox.ac.uk/__data/assets/pdf_file/0019/3853/brain-emulation-roadmap-report.pdf

Shah, A. (2009) Nvidia closing in on 2 teraflops with graphics card. *Computerworld*. Retrieved from http://www.computerworld.com/s/article/9125345/Nvidia_closing_in_on_2_teraflops_with_graphics_card

Stanovich, K. E. (2009). *What intelligence tests miss: the psychology of rational thought*. New Haven and London: Yale University Press.

Stanovich, K. E. & West, R.F. (2000) Individual differences in reasoning: Implications for the rationality debate? *Behavioral and Brain Sciences*, 23, 645-726.

Tversky, A. & Kahneman, D. (1986) Rational Choice and the Framing of Decisions. *The Journal of Business*, 59(4), part 2, S251-S278.

Uddin, L.Q., Iacoboni, M., Lange, C. & Keenan, J.P. (2007) The self and social cognition: the role of cortical midline structures and mirror neurons. *Trends in Cognitive Sciences*, 11(4), 153-157.

Vinge, V. (1993) The coming technological singularity: How to survive in the post-human era. *Whole Earth Review*, Winter 1993.

Yudkowsky, E. (2001) *Creating Friendly AI*. <http://www.singinst.org/upload/CFAI/>

Yudkowsky, E. (2008). Artificial Intelligence as a Positive and Negative Factor in Global Risk. In N. Bostrom & M. M. Čirković (Eds.), *Global Catastrophic Risks* (pp. 308–343). Oxford: Oxford University Press.