Implications of a Software-Limited Singularity

Carl Shulman
MIRI Visiting Fellow

Anders Sandberg
Future of Humanity Institute

Abstract

A number of prominent artificial intelligence (AI) researchers and commentators (Moravec 1999a; Solomonoff 1985; Vinge 1993) have presented versions of the following argument:

1. Continued exponential improvement in computer hardware will deliver inexpensive processing power exceeding that of the human brain within the next several decades.

2. If human-level processing power were inexpensive, then the software for AI with broadly human-level (and then superhuman) cognitive capacities would probably be developed within two decades thereafter.

Therefore,

3. There will probably be human-level AI before 2060.

Shulman, Carl, and Anders Sandberg. 2010. "Implications of a Software-Limited Singularity." In *ECAP10: VIII European Conference on Computing and Philosophy*, edited by Klaus Mainzer. Munich: Dr. Hut.

This version contains minor changes.

Call this the Hardware Argument. If sound, it is of great importance: human-level AI would likely be capable of developing still more sophisticated AIs soon thereafter, resulting in an "intelligence explosion" or "technological singularity" with potentially enormous impact (Chalmers 2010; Good 1965; Yudkowsky 2008). Further, the first premise has substantial empirical support, as computing price-performance has demonstrated fast exponential improvement since pre-transistor technologies such as vacuum tubes (Kurzweil 2005). However, the second premise is much more controversial, and our focus in this paper.

We discuss several possible hardware-to-software "transmission mechanisms," routes from abundant processing power to expedited progress in AI, that might be held to underly the second premise. Powerful hardware may improve performance simply by allowing existing "brute force" solutions to run faster (Moravec 1976). Where such solutions do not yet exist, researchers might be incentivized to quickly develop them given abundant hardware to exploit. Cheap computing may enable much more extensive experimentation in algorithm design, tweaking parameters or using methods such as genetic algorithms. Indirectly, computing may enable the production and processing of enormous datasets to improve AI performance (Halevy, Norvig, and Pereira 2009), or result in an expansion of the information technology industry and the quantity of researchers in the field.

Are these transmission mechanisms sufficient to support the second premise of the Hardware Argument? A key counterargument is that theoretical advances in AI often require many sequential steps in which ideas are developed, disseminated, and then built upon by other scientists (Malcolm 2000). To the extent that many such advances must be made in order to develop even relatively "brute force" methods of achieving human-level AI, the speedup from human-level processing power will be limited.

We argue that if one accepts this objection, one should not only reduce one's confidence in the Hardware Argument for the medium-term development of human-level AI, but also increase one's confidence in a relatively rapid and disruptive "intelligence explosion" if and when human-level AI is eventually developed. First, ceteris paribus, this should lead one to expect increased hardware capacity relative to the requirements of human-level AI when it is developed. Second, the development of AI systems capable of conducting theoretical AI research at greater serial speeds than humans would more powerfully accelerate AI research if serial research is the bottleneck in AI progress.

1. Introduction

The human brain has a large but finite capacity to process information, a capacity that can be compared in various ways to the processing power of electronic computers. Eventually, advancing computer technology may lead to cheap computing devices with human-level or greater processing power, and several authors have argued that human-level artificial intelligence (AI) will be developed soon after cheap human-level processing power is achieved. Mathematician and novelist Vernor Vinge (1993) wrote, "Progress in computer hardware has followed an amazingly steady curve in the last few decades. Based largely on this trend, I believe that the creation of greater than human intelligence will occur during the next thirty years." AI and information theory pioneer Solomonoff (1985) similarly argued that human-level processing power would expedite the development of human-level AI shortly thereafter. Machine vision pioneer Hans Moravec has advanced this view for several decades in multiple venues (Moravec 1976, 1999a, 1999b). Using Moravec's estimates, we could formulate the argument as follows:

1. Continued exponential improvement in computer hardware will deliver inexpensive processing power exceeding that of the human brain within the next several decades.

2. If human-level processing power were inexpensive, then the software for AI with broadly human-level (and then superhuman) cognitive capacities would probably be developed within two decades thereafter.

Therefore,

3. There will probably be human-level AI before 2060.

Call this the Hardware Argument (with alternate versions available for varying estimates). Premise (2) claims that if, counterfactually, we had inexpensive human-level processing power today, human-level AI software would be developed within two decades of the present. The Hardware Argument is thus distinct from arguments that human-level processing power will be followed by human-level AI because of advances in other fields, such as neuroimaging, which would not be generated by increased processing power alone (Kurzweil 2005). It is also distinct from claims that the software for AI will coincidentally be developed by the time human-level processing power is reached, but not because of processing power advances.

If sound, the Hardware Argument is of great importance: human-level AI would likely be capable of developing still more sophisticated AIs soon thereafter, resulting in an "intelligence explosion" or "technological singularity" with potentially enormous positive or destructive consequences (Chalmers 2010; Yudkowsky 2008). Further, the first

premise has substantial empirical support, as computing price-performance has demonstrated fast exponential growth since pre-transistor technologies such as vacuum tubes (Kurzweil 2005). However, the second premise is much more controversial, and calls for an account of the mechanisms by which improved processing power would drive improved software: call these "transmission mechanisms."

We begin with a taxonomy of proposed transmission mechanisms, and then discuss obstacles that could prevent them from delivering the necessary software advances, even given human-level processing power. We argue that the most significant challenge to the Hardware Argument is the possibility that serial cognitive steps by human researchers will act as a bottleneck to progress. However, if this objection to the Hardware Argument holds, it has the surprising implication that we should not only expect that human-level AI will take longer to develop than we would otherwise have thought, but also that if and when human-level AI is created, the transition will be more disruptive and abrupt.

2. Hardware to Software Transmission Mechanisms

We can roughly order the possible mechanisms from most to least direct. At the most direct, sometimes existing algorithms scale well with improved hardware. Moravec (1976) claimed that "although there are known brute force solutions to most AI problems, current machinery makes their implementation impractical." For instance, alpha-beta pruning in computer chess and many aspects of machine vision processing showed this pattern, although these fields also gained from software improvements that provided performance gains comparable to those from hardware (Moravec 1999a; Malcolm 2000). At a slightly further remove, the availability of improved processing power may motivate computer scientists to develop new processing-intensive algorithms. That is, there may be AI-relevant algorithms that could be relatively quickly discovered if researchers made an effort, but researchers do not expend the effort, since the algorithm would be impractically demanding in any case.

Processing power may also be used in the software research process by facilitating experiments. If a test of a new algorithm takes 20 minutes instead of 20 days, researchers can adjust their research paths in response to much more frequent and detailed feedback. Cheap computing can allow the testing of many variations of an algorithm with slightly different parameter values in parallel, as in various forms of evolutionary programming, the approach advocated in (Moravec 1999b).

More indirect benefits flow from the above, as improved hardware capacities and software performance bring additional inputs to the field. For instance, cheap computation helped lead to the development of the World Wide Web and document scanning tech-

nologies, producing enormous language corpora that have invigorated machine translation (Halevy, Norvig, and Pereira 2009). Processing these datasets is in turn a hardware intensive task, as is experimentation. Perhaps more importantly, improved computer performance, whether driven by hardware or software advance, expands the scope of the computing industry and the number of potential innovators. Since 1970, the number of computer scientists in the United States has increased twentyfold, although the formerly exponential growth pattern has slowed in the 21st century (Ruggles et al. 2010). If continued hardware advances create new markets for software advance, e.g., by enabling more autonomous robots, this may further multiply researcher numbers (Moravec 1999a).

3. Bottlenecks and Diminishing Returns

Clearly, increased processing power has the potential to accelerate improvements in many aspects of AI performance and research. However, if other areas do not so benefit, they may become bottlenecks to further progress. Several challenges of this type threaten the second premise of the Hardware Argument.

First, critics note that hardware-driven gains have been strongest in domains historically thought to be unusually hardware constrained, with established hardware-hungry solutions: vision, computer chess, articulated motion, etc. (Malcolm 2000). Insofar as those algorithms were visible decades in advance of human-competitive performance, a lack of similarly identified algorithms for other types of reasoning and learning would suggest that existing or easily discoverable "brute force" solutions will not suffice for human-level AI.

Second, while computer hardware may enable faster and more numerous experiments, the human labor to understand the results and design new experiments may become a limiting factor. In theory, extremely parallel methods such as genetic algorithms could avert this problem, but for that approach human-level processing power is not the relevant benchmark: instead one should consider the vastly larger computational resources required to simulate the evolutionary process that produced the human brain (Baum 2004).

Third, and perhaps most importantly, increasing the number of researchers working in parallel does not appear to increase the pace of innovation at a near-linear rate: few would argue that the pace of computer innovation has multiplied twentyfold in tandem with the quantity of computer scientists. To the extent that research progress demands sequential steps by humans as insights are developed, communicated, and built upon, the relatively stable pace of human thought and communication will remain a core bottleneck, requiring large gains in other areas for the Hardware Argument to go through.

These obstacles suggest a common basis for rejecting the Hardware Argument, namely the claim that learning to exploit hardware resources effectively will require many incremental human theoretical advances building upon one another in succession. Indeed, this seems to be the typical core objection to the Hardware Argument (Malcolm 2000; Hofstadter 1979; Lanier 2000).

4. Rejecting the Hardware Argument Suggests Later, More Abrupt AI

The above objection takes the speed of human thought and communication as a key bottleneck to AI development, a problem not to be circumvented with additional hardware or human researchers. Those who accept it and reject the Hardware Argument should thus think human-level AI further in the future than otherwise. However, if human-level AI is eventually developed, systems capable of conducting AI research themselves could likely be run at higher serial speeds than human brains if given adequate computing power. Furthermore, slower advances in AI software should mean that hardware is more advanced at the time human-level AI is developed. If this makes hardware abundant relative to the demands of early human-level AI designs, the pace of research could increase extremely rapidly.

The primary reason to think otherwise is that researchers could use hardware inefficiently in early AIs to ease design, "bloating" early AIs to the limits of available machines. However, the less important hardware is relative to software design, the less the incentive for such bloating, and the easier it would be to quickly improve hardware efficiency given a prototype AI. Allocating a larger portion of computation to AI research could exacerbate bloat, while it would be reduced if development were more computationally demanding relative to running an AI, e.g., if evolutionary methods are used.

Thus, while the Hardware Argument leads to the apparently extreme prediction that human-level AI will probably arise in the next several decades, rejecting the argument appears to lead to the alternative extreme prediction that if and when AI eventually occurs, subsequent progress will be accelerated, perhaps aptly described as an "intelligence explosion" (Good 1965). This connection should reduce our confidence in a future path to advanced AI that is simultaneously successful, slow, and smooth.

References

Baum, Eric B. 2004. *What Is Thought?* Bradford Books. Cambridge, MA: MIT Press.

Chalmers, David John. 2010. "The Singularity: A Philosophical Analysis." *Journal of Consciousness Studies* 17 (9–10): 7–65. http://www.ingentaconnect.com/content/imp/jcs/2010/00000017/f0020009/art00001.

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.

Hofstadter, Douglas R. 1979. *Gödel, Escher, Bach: An Eternal Golden Braid.* New York: Basic Books.

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

Lanier, Jaron. 2000. "One Half a Manifesto." *Edge,* November 11. http://www.edge.org/conversation/one-half-a-manifesto.

Malcolm, Chris. 2000. "Why Robots Won't Rule the World." January 20. Accessed July 31, 2012. http://web.archive.org/web/20100531093218/http://www.dai.ed.ac.uk/homes/cam/WRRTW.shtml.

Moravec, Hans P. 1976. "The Role of Raw Rower in Intelligence." Unpublished manuscript, May 12. Accessed August 12, 2012. http://www.frc.ri.cmu.edu/users/hpm/project.archive/general.articles/1975/Raw.Power.html.

———. 1999a. *Robot: Mere Machine to Transcendent Mind.* New York: Oxford University Press.

———. 1999b. "Simple Equations for Vinge's Technological Singularity." Unpublished manuscript, February. http://www.frc.ri.cmu.edu/~hpm/project.archive/robot.papers/1999/singularity.html.

Ruggles, Steven, J. Trent Alexander, Katie Genadek, Ronald Goeken, Matthew B. Schroeder, and Matthew Sobek. 2010. *Integrated Public Use Microdata Series: Version 5.0 [Machine-Readable Database].* Minneapolis, NM: University of Minnesota. http://usa.ipums.org.

Solomonoff, Ray J. 1985. "The Time Scale of Artificial Intelligence: Reflections on Social Effects." *Human Systems Management* 5:149–153.

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. 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.