Taking Superintelligence Seriously

Miles Brundage

Book Review:
Superintelligence: Paths, Dangers, Strategies by Nick Bostrom (Oxford University Press, 2014).

Philosopher Nick Bostrom's latest book, *Superintelligence: Paths, Dangers, Strategies*, is a seminal contribution to several important research areas including catastrophic risk analysis, the future of artificial intelligence (AI), and safe AI design. It has also received enthusiastic recommendations from high profile figures in technology such as Bill Gates and Elon Musk; and Bostrom noted in a recent interview (Simonite 2015) that his ideas have gained traction at major technology companies. Despite all this attention to Bostrom's book, however, mischaracterizations of his ideas and those of his endorsers abound in the media. For example, numerous op-eds have been written in recent months (e.g. Harris, 2014; Edwards, 2015) purporting to debunk the concerns of Gates, Musk, and others by, for example, highlighting the limitations of present AI technologies. Such articles often miss a key point of *Superintelligence* and those referencing it, which is that even though we may have decades or more before highly intelligent AI systems exist, the challenge of keeping such systems safe is sufficiently large that serious research on the topic is warranted today. Indeed, Musk recently donated $10 million for precisely that purpose, $1.5 million of which will go toward a center led by Bostrom. While *Superintelligence* is not a perfect book (indeed, Bostrom writes in the preface that he expects to have made many mistakes in it [p. viii]), and I will comment on some of its shortcomings below, it is by far the best work on these issues to date. The publication of *Superintelligence* substantially raises the bar for thinking and writing on long-term AI risks, a topic that has previously been mostly confined to conference papers (e.g. Omohundro, 2008), wide-ranging edited volumes (e.g. Eden, Moor, Soraker, and Steinhart eds., 2012), and journalistic books (e.g. Barrat, 2013). Compared to this prior literature, *Superintelligence* is markedly more up-to-date, clear, rigorous, and comprehensive. These attributes, and the deep significance of Bostrom's chosen topic, make the book a must-read for anyone interested in the long-term future of humanity.

Superintelligence is roughly organized into three sections, as suggested by its subtitle ("*Paths, Dangers, Strategies*"): first, Bostrom discusses paths by which superintelligence (a system that vastly exceeds human levels of intelligence in virtually all areas) might be obtained. Next, he argues that the default outcome of developing superintelligence is likely to be catastrophic, motivating the need for substantial care in such systems' design and governance. Finally, he critically analyzes possible strategies for ensuring that the development of superintelligence, if it does occur, proceeds safely and maximally benefits humanity. Cumulatively, these sections of the book constitute a persuasive demolition of what Bostrom calls the "null hypothesis" (p. viii), namely that the future risks of AI needn't be taken seriously today. Steering AI development toward safe and broadly beneficial outcomes is, to Bostrom, the "essential task of our age." (p. 260), and *Superintelligence* puts forward numerous helpful ideas, terms, and principles to assist with addressing it.

In the first few chapters of the book, Bostrom outlines the history of AI research and efforts to anticipate its progress over time. He also defends his focus on AI (as opposed to, e.g. the enhancement of human cognition) as the most plausible route to superintelligence. While noting the limitations of expert prognostications, Bostrom carefully summarizes current expert opinion in the field as follows: "it may be reasonable to believe that human-level machine intelligence has a fairly sizeable chance of being developed by mid-century, and that it has a non-trivial chance of being developed considerably sooner or much later; that it may perhaps fairly soon thereafter result in superintelligence; and that a wide range of outcomes may have a significant chance of occurring, including extremely good outcomes and outcomes that are as bad as human extinction." (p. 21) In light of possible AI-related outcomes ranging from the end of poverty, disease, and other forms of suffering to human extinction this century, Bostrom reasonably concludes that "the topic is worth a closer look." (p. 21) The analysis of expert opinions on AI and the limitations thereof here is notably clear and level-headed, making it a useful introduction to the topic for readers new to AI.

In subsequent chapters on "Paths to superintelligence" and "Forms of superintelligence," Bostrom describes multiple possible pathways (and corresponding timelines) to the existence of systems with vastly greater than human intelligence, including sped up human brain emulations, genetic enhancement of humans, brain-machine interfaces, and vast networks of humans and machines. Bostrom concludes that AI will ultimately have "enormous" advantages over biological intelligences in terms of both hardware and software (p. 61), including its duplicability, editability, and the ease of expanding its memory and processing power. To illustrate the magnitude of these differences, Bostrom follows researcher Eliezer Yudkowsky in suggesting that the difference between a village idiot and Einstein is likely to be smaller than the difference between Einstein and a superintelligence (p. 70). We are simply the dumbest animals capable of building a global community, Bostrom thinks, rather than the pinnacle of evolution.

Given the ways that humans seem improvable, and the fundamental advantages of digital intelligences, Bostrom thinks that, although the possibility of "a slow takeoff [of intelligence, and by extension, a superintelligence] cannot be excluded." (p. 77), the possibility of a fast or moderate takeoff is one that should be taken quite seriously. This consideration is especially critical in light of the content of Bostrom's fifth chapter, wherein he argues that a single AI project might become sufficiently advanced relative to others that it could achieve its goals on a global scale, whatever those goals might be.

While Bostrom musters many provocative analogies and arguments for the plausibility of a rapid intelligence explosion and subsequent global takeover by a superintelligence, *Superintelligence* hardly settles the issue, which depends on many complex considerations about the future of innovation and society. Indeed, some researchers remain unconvinced that a single AI project could rapidly outstrip human capabilities and those of all other competing projects.

Economist Robin Hanson, for example, points to the distributed nature of innovation (in general, and in the context of AI) as a reason to think that there won't be an AI takeover of the sort Bostrom is concerned with. "As 'intelligence' is just the name we give to being better at many mental tasks by using many good mental modules, there's no one place to improve it. So I can't see a plausible way one project could increase its intelligence vastly faster than could the rest of the world" (Hanson, 2014).

Following a different line of thought, in her analysis of Bostrom's book, researcher Katja Grace raises doubts about Bostrom's notion of a "crossover point" at which an AI system becomes able to improve itself directly (by, e.g. conducting its own AI research), rather than being primarily improved by its developers. In Bostrom's formulation of intelligence growth as a function of "optimization power," or effort, divided by "recalcitrance," or the difficulty of intelligence improvement (p. 65), such a crossover point could portend a rapid increase in system intelligence. Grace, however, finds surprisingly little empirical evidence that there is a strong relationship between technology funding (a proxy for effort at improving it) and its subsequent rates of progress (Grace, 2014).

Finally, adding additional skepticism to *Superintelligence*'s perspective on the plausibility of an intelligence explosion, AI researcher Yoshua Bengio questions Bostrom's assumption that a large increase in computational power and knowledge would lead to enormously higher levels of intelligence:

> "I see a lot of good mathematical and computational reasons why A.I. research could one day face a kind of wall (due to exponentially growing complexities) that human intelligence may also face—which could also explain why whales and elephants, which have bigger brains than ours, are not super-intelligent. We just don't know enough to be able to make anything but informed guesses, regarding this question. If this wall-of-complexity hypothesis is true, we might one day have computers that are as smart as humans but have quick access to a lot more knowledge. But by that time, individual humans might have access to that kind of knowledge too (we already do, but slowly, via search engines). That would be very different from the super-intelligence notion." (Sofge, 2015)

Resolving these types of disagreements about the future distribution and types of intelligence would depend on rigorous formulations of the nature of intelligence and detailed accounts of how and to what extent it can be improved (accounts which may not be forthcoming, as the field of AI currently lacks a robust consensus on such topics). Alas, while *Superintelligence* marks a substantial advance over the prior literature in rigorously exploring such topics, it remains highly tentative and caveated in important respects, which is appropriate given the early state of research on long-term AI safety. Insofar as Bostrom has demonstrated that (artificial) intelligence explosion is a possibility worth strongly considering and researching today (if not necessarily the "essential task of our age" [p. 260]), the reader will be rewarded by the many insights that follow from his exploration the topic. Indeed, even if one dominant superintelligence does not rapidly arise, or none arise, Bostrom provides useful insights into how to design less powerful AI systems to be ethical (and perhaps more importantly, how *not* to do so), and provides a compelling model of how to carefully analyze future technological risks.

Some of Bostrom's critical insights relate to possible solutions to concerns that arise from the combination of two important (and often misunderstood) theses that Bostrom presents in his chapter, "The superintelligent will." The first of these, the *orthogonality thesis* claims that, with some caveats, more or less any level of intelligence can be paired with more or less any set of final goals, since intelligence measures an agent's ability to achieve its goals (p. 107). For instance, there is nothing logically contradictory about an agent as intelligent as Einstein having the final goal of counting blades of grass in a yard for the rest of his life, even if it is implausible in a human context. In the context of a designed artifact like AI, we have control over its motivational system and can mix and match things like knowledge, cognitive processes, and final goals. This doesn't matter that much in and of itself, but it is important when paired with Bostrom's second claim in this chapter: the *instrumental convergence thesis*. The instrumental convergence thesis states that, again with some caveats and complexities, there are likely several sub-goals that would be instrumentally valuable for a wide variety of agents to pursue regardless of their final goals. Bostrom argues for the instrumental convergence of a wide variety of agents on instrumental goals such self-preservation, goal-content integrity, cognitive enhancement, technological perfection, and resource acquisition.

Combining these two claims, Bostrom argues persuasively in the next chapter, "Is the default outcome doom?," that great care is needed to design the final goals of a superintelligent agent in a way that is not catastrophic for humanity. Much like a genie in a bottle, a superintelligence would by default do what we asked, not what we wished we had asked, motivating the detailed exploration of safe superintelligence design in many of the remaining chapters. An AI instructed to keep humans "safe and happy" might, for example, "entomb everyone in concrete coffins on heroin drips," Bostrom's colleague Stuart Armstrong noted (Sawer, 2015). In search of solutions to such difficulties, Bostrom explores options such as "oracle" AIs that could only answer questions, finding that most intuitively appealing approaches to AI safety would turn out to be surprisingly dangerous. While none of the better solutions he explores are developed in enough detail to be implementable anytime soon, Bostrom sheds considerable light on why exactly naïve approaches wouldn't work, and points toward many promising areas of research that might be needed to prepare for and shape an intelligence explosion, should that become necessary. Here, Bostrom's background as a philosopher shines through, as he rigorously examines the merits of a variety of possible approaches to superintelligence control and pokes holes in superficially promising solutions sometimes put forward by proponents of the null hypothesis.

In the final chapters of *Superintelligence*, Bostrom broadens his scope beyond individual superintelligent systems and considers what the foregoing analysis implies for society as a whole. Here, the analysis is lucid, but fairly abstract, and the book offers little in the way of concrete suggestions for acting on the concerns raised in earlier chapters. In "The strategic picture," he analyzes issues such as the relationship between advances in different technological domains and cooperation between different AI development projects. With regard to paces of technological development, Bostrom posits a "principle of differential technological development," which is that society should "retard the development of dangerous and harmful technologies, especially ones that raise the level of existential risk, and accelerate the development of beneficial technologies, especially those that reduce the existential risks posed by nature or by other technologies" (p. 230). Such a principle is hard to disagree with, though Bostrom's book provides little insight into which technologies to differentially accelerate/decelerate or how to do so (he considers how whole brain emulation research fits into the overall AI risk picture, but stops short of any strong conclusion in light of the uncertainties involved [p. 245]). Indeed, the literature on responsible innovation (Guston 2014) testifies to the complexity of analyzing and constructively intervening in innovation systems in real-time.

While the principle of differential technological development is well-stated and germane to the discussion at hand, Bostrom here adds little in the way of specific proposals, in contrast to his more significant contributions to the analysis of long-term AI safety. His analysis also usefully identifies the critical importance of collaboration between AI projects, but falls short of suggesting how approaches like financial cross-investment between AI companies can be elicited in a competitive market economy, or how scientific and engineering cultures can be improved with respect to their safety consciousness (p. 258).

Bostrom's "common good principle," which states that "[S]uperintelligence should be developed only for the benefit of all of humanity and in the service of widely shared ideals," (p. 254) does indeed seem like a worthy principle to inculcate in the AI research community, as he suggests. But, the massive social stakes of research and innovation have long been recognized, and calls for incorporating values into the design and governance of technologies in general and AI in particular are not new (Guston, 2014; Brundage, forthcoming), so it's not obvious how much progress Bostrom's suggested starting point represents. Indeed, AI already touches the lives of billions of people around the world on a daily basis. Thus, designing and governing AI more democratically now may be important regardless of whether superintelligence comes to fruition.

Despite the book's limitations, which stem primarily from the complexity and diversity of the topics that Bostrom navigates, *Superintelligence* is a highly useful contribution to discussions of AI and the future of humanity. Bostrom eloquently articulates the potential upside of well-designed superintelligence, and constructively analyzes the potentially catastrophic downsides. By articulating various principles, terms, and typologies for thinking about superintelligence and technological grand strategy more generally, *Superintelligence* rightfully earns its place on the bookshelf of anyone seriously interested in the future of humanity. If it is as widely read as it deserves to be, *Superintelligence*'s publication is likely to inspire further careful thought from researchers of various disciplines (and the public more broadly) on how to steer AI technologies in a socially positive direction.

Acknowledgments

The author would like to acknowledge the helpful feedback and suggestions from Seth Baum, Jake Nebel, Katja Grace, Adi Wiezel, and two anonymous reviewers on earlier versions of this review. This work was supported by the National Science Foundation under award #1257246 through the Virtual Institute for Responsible Innovation (VIRI). The findings and observations contained in this paper are those of the author and do not necessarily reflect the views of the National Science Foundation.

References

Barrat, J. 2013. *Our Final Invention: Artificial Intelligence and the End of the Human Era*. New York: Thomas Dunne Books.

Brundage, M. Forthcoming. "Artificial Intelligence and Responsible Innovation," chapter in Fundamental Issues of Artificial Intelligence, ed. Vincent C. Müller, Berlin: Springer (Synthese Library), forthcoming.

Eden, Moor, Soraker, and Steinhart (eds.), 2012. *Singularity Hypotheses: A Scientific and Philosophical Assessment*. New York: Springer.

Edward, J. 2015. "7 Reasons Elon Musk Is Wrong to Believe Intelligent Robots Will One Day Kill Us All," Business Insider, January 21, 2015. Accessed July 10, 2015, at http://www.businessinsider.com/why-elon-musk-is-wrong-that-intelligent-robots-will-kill-us-all-2015-1

Future of Life Institute, 2015. "Research Priorities for Robust and Beneficial Artificial Intelligence: An Open Letter." Accessed July 10, 2015 at http://futureoflife.org/AI/open_letter

Grace, K. 2014. "Superintelligence 6: Intelligence explosion kinetics," LessWrong.com, October 21, 2014. Accessed July 10, 2015, at http://lesswrong.com/r/discussion/lw/l4e/superintelligence_6_intelligence_explosion/

Guston, D. 2014. "Responsible innovation: motivations for a new journal." *Journal of Responsible Innovation.* Volume 1, Issue 1, p. 1-8.

Hanson, R. 2014. "I still don't get FOOM," OvercomingBias.com, July 24, 2014. Accessed July 10, 2015, at http://www.overcomingbias.com/2014/07/30855.html

Harris, D. 2014. "When data becomes dangerous: Why Elon Musk is right and wrong about AI," Gigaom, April 4, 2014. Accessed July 10, 2015, at https://gigaom.com/2014/08/04/when-data-becomes-dangerous-why-elon-musk-is-right-and-wrong-about-ai/

Omohundro, S. 2008. "The Basic AI Drives," *Proceedings of the First AGI Conference*, Volume 171, *Frontiers in Artificial Intelligence and Applications,* ed. Wang, Goertzel, and Franklin, 2008.

Sawer, P. 2015. "Threat from Artificial Intelligence not just Hollywood fantasy," *The Telegraph*, June 27, 2015. Accessed July 10, 2015, at http://www.telegraph.co.uk/news/science/science-news/11703662/Threat-from-Artificial-Intelligence-not-just-Hollywood-fantasy.html

Simonite, T. 2015. "AI Doomsayer Says His Ideas Are Catching On," *MIT Technology Review*, April 7, 2015. Accessed July 10, 2015, at http://www.technologyreview.com/news/536381/ai-doomsayer-says-his-ideas-are-catching-on/

Sofge, E. “Bill Gates Fears A.I., But A.I. Researchers Know Better,” *Popular Science*, January 30, 2015. Accessed July 10, 2015, at http://www.popsci.com/bill-gates-fears-ai-ai-researchers-know-better