Futures 72 (2015) 32–35

Contents lists available at ScienceDirect

Futures

journal homepage: www.elsevier.com/locate/futures

Taking superintelligence seriously
Superintelligence: Paths, dangers, strategies by Nick Bostrom (Oxford University Press, 2014)

Miles Brundage

Consortium for Science, Policy, and Outcomes, Arizona State University, Tempe, AZ 85287, United States

ARTICLE INFO

Article history:
Received 4 September 2014
Received in revised form 11 July 2015
Accepted 18 July 2015
Available online 8 August 2015

Keywords:
Existential risk
Artificial intelligence
Superintelligence
Responsible innovation

ABSTRACT

A new book by Nick Bostrom, Superintelligence: Paths, Dangers, Strategies, is reviewed. Superintelligence explores the future of artificial intelligence and related technologies and the risks they may pose to human civilization. The book ably demonstrates the potential for serious thinking aimed at the long-term future. Bostrom succeeds in arguing that the development of superintelligent machines will, if not properly managed, create catastrophic risks to humanity. The book falls short in some respects, and some sections are more compelling and novel than others. Overall, however, Bostrom's book succeeds in demolishing the "null hypothesis" according to which the possibility and risks of superintelligence can continue to be ignored, and is a must-read for those interested in the long-term future of humanity.

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; Edward, 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, & Steinhart, 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.

E-mail address: miles.brundage@asu.edu (M. Brundage).

http://dx.doi.org/10.1016/j.futures.2015.07.009
0016-3287/

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 need not 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)