Nick Bostrom, *Superintelligence: Paths, Dangers, Strategies*, Oxford: Oxford University Press, 2014, xvi+324, 18.99£, ISBN 978-0-19-967811-2.

Paul D. Thorn
Philosophy Department
Heinrich-Heine-Universitaet Duesseldorf
Universitaetsstr. 1, Duesseldorf
40225 Deutschland
thorn@phil.hhu.de

Nick Bostrom's book *Superintelligence: Paths, Dangers, Strategies* is a systematic and scholarly study of the possible dangers issuing from the development of artificial intelligence. The book is relatively comprehensive, covering a multitude of topics relating to the safe development of superhuman artificial intelligence. If the arguments presented in Bostrom's book are correct, then the book's subject matter represents a very important field of study, inasmuch as (1) the creation of a superintelligent being represents a possible means to the extinction of mankind, and (2) there are actions that can be taken to reduce this risk. According to other arguments that Bostrom offers, the creation of a 'friendly' superintelligent being might, on the other hand, lead to the rapid development of many beneficial technologies, e.g., technologies that eliminate death by aging.

Bostrom asserts that his book is written *as if* its target audience were earlier time-slices of himself, assuming that he wanted to quickly bring the past time-slices of Bostrom 'up to speed', regarding the present Bostrom's thoughts concerning the topic of superintelligence. Past time-slices of Bostrom received a PhD in Philosophy at the London School of Economics, with studies prior to that in the areas of Physics, Mathematics, Mathematical Logic, Computational Neuroscience, and Artificial Intelligence. One will not need such an extensive background to catch up with the views of current time-slice Bostrom. In fact, I would say that the book is optimized for persons with training in philosophy, including some familiarity with decision theory. For persons with such a background, such as myself, the book is an interesting and relatively easy read. Computer scientists working in logic and artificial intelligence will also have an easy time with the book. Absent the optimal background, the most important arguments of the book will still be accessible, though some of the technical vocabulary (which is by no means heavy) may be a source of frustration for some. Although the book is more accessible than typical articles appearing in academic

philosophy journals, I do not think that this should detract from the book's interest for academic philosophers.

Bostrom's book could also be used as the main text for an intermediate level course in philosophy, covering the prospects and dangers of machine intelligence, an exciting topic which promises to attract many students. In fact, a colleague and I will use the book for just such a course, in the months following the completion of this review. The book divides into 15 chapters, which is almost perfect for a semester long course that will proceed at a pace of one chapter per week. What will we be covering?

Chapter 1 provides a concise history of artificial intelligence, with a focus on describing the predictions of experts, at various stages in the development of artificial intelligence research, concerning when *human-level* artificial intelligence is likely to be achieved. While confident that the creation of human-level artificial intelligence is inevitable, barring a global catastrophe, Bostrom acknowledges that it is difficult to judge how long it will take to develop this technology. Indeed, Bostrom observes that, in the 1940s, expert predictions placed the landmark 20 years in the future, and that, since then, the expected arrival date has receded at a rate of one year per year. Chapter 2 surveys the different forms that artificial intelligence might take, including Good Old-Fashioned Artificial Intelligence and whole brain emulation. Bostrom discusses the relative dangers represented by different forms of artificial intelligence. Various means of improving human intelligence (e.g., through smart-drugs, and genetic selection) are also discussed. Such ways of augmenting human intelligence may be instrumental to the development of human-level artificial intelligence, and to the safe development of non-human superintelligent beings. The latter point illustrates Bostrom's idea that it may be desirable to influence the order in which we develop certain technologies.

Chapter 3 surveys several means by which a being might manifest superintelligence. For example, superintelligence may manifest itself in a being capable of the same operations as a human intellect, but at speeds much faster than those possible for a human being. Once we develop the technology to implement human-level artificial intelligence within a digital computer, it would, it seems, be a short step to superintelligence via a super-fast implementation of human-level artificial intelligence. Chapter 4 presents an extended discussion of what Bostrom calls the "speed of takeoff", i.e., the speed at which the development of human-level artificial intelligence would lead to the development of 'extreme' superintelligence, i.e., the kind of superintelligence whose capability in developing and deploying new technologies, e.g.,

nano-technologies, would constitute a potential threat to humanity. The topic of Chapter 4 is important inasmuch as a slow takeoff would, presumably, give the human beings involved in the development of an extreme superintelligence the opportunity to influence the goals and character of the superintelligent being, or to avert the process altogether. So, other things being equal, we ought to pursue AI research in a way that tends to a slow takeoff.

Chapters 5 and 6 discuss the likelihood that a single superintelligent being will come into existence, as opposed to several, and the means by which a superintelligent being could gain (what Bostrom calls) a *decisive strategic advantage*, i.e., a level of technological and material advantage sufficient to enable it to achieve complete world domination. If we jump ahead to Chapter 11, we find a discussion of the possible effects on humanity in the case where multiple superintelligent beings come to exist, and no one of these beings achieves ascendancy.

Within Chapter 7, Bostrom advances and defends two theses regarding the motives and likely actions of a superintelligent being. According to the first, the *orthogonality thesis*, superintelligence is compatible with almost any *final goal*. The crucial consequence of the orthogonality thesis is that the possession of superintelligence (in a manner that would enable a decisive strategic advantage) does not imply being wise or benevolent. According to the second thesis, *instrumental convergence*, superintelligent beings with a wide variety of final goals would pursue the same *intermediate goals*. The argument for instrumental convergence appeals to the fact that there are certain intermediate goals whose satisfaction tends to enable the satisfaction of almost any final goal. For example, self-preservation tends to be instrumental to the achievement of one's goals, assuming that one will be in a position to promote one's goals in the future. Similarly, for a superintelligent being, the acquisition of additional physical and computational resources will be instrumental to almost any final goal.

Chapter 8 outlines why it is that, despite the absence of malevolence, a superintelligent being could act in a way that results in human extinction, or in the end of humanity. For example, charged with the goal of computing the numeric value of *pi*, or manufacturing many paperclips, a superintelligent being might proceed by an unbridled acquisition of physical resources, in order to facilitate its computations or manufacturing capacity, thereby appropriating our bodies as a convenient source of atoms, or modifying our environment in a way that results in our extinction.

Chapters 9 and 10 introduce, and discuss, strategies and problems associated with limiting or controlling a superintelligent AI. Two broad

categories of strategy are considered: *capability control* and *motivation selection*. Among the sorts of capability control discussed by Bostrom, we find *boxing* (a precaution aimed at the physical and informational containment of an AI), and *tripwires* (mechanisms that execute the shutdown of an AI, if indicators of dangerous behavior are detected). Bostrom makes a convincing case that various approaches to capability control may be unsuccessful, and thus should be combined with measures that are designed to ensure that a respective superintelligent being has a concern for the interests of human beings.

Following up on chapters 9 and 10, chapters 12 and 13 outline the problem of creating a superintelligent being that has appropriate final goals (i.e., goals that manifest a sufficient concern for the interests of human beings), and is governed by an adequate decision theory, and appropriate epistemological standards, i.e., standards that specify what it is rational to believe, and to what degree, under respective conditions. The technical problems here are considerable, since, to date, no unproblematic accounts have been articulated in any of the key areas: There is no received view regarding how to measure the satisfaction of human interests, regarding what the correct decision theory is, or what the correct epistemological principles are. Given the preceding, Bostrom discusses the possibility of *indirect-normativity*, i.e., the possibility of specifying the process by which a superintelligent being will determine what decision-theoretic and epistemological standards it will adopt, along with its (appropriate) final goals. One idea is that of *coherent extrapolated volition*, which involves having a superintelligent being determine what our considered judgment would be, regarding how to measure the satisfaction of human interests, for example, in the case where we were smarter, more knowledgeable, and more the persons we wished we were, etc.

Given the dangers described in the preceding chapters, chapters 14 and 15 discuss the sorts of treaties and policy measures that might be adopted at the national and international levels as means to reducing the risk of developing and unleashing a malignant superintelligent being. These chapters also discuss some factors that may induce competing projects, aimed at developing superintelligence, to adopt inadequate safety standards.

Assuming that we take Bostrom's warnings seriously, the book outlines many lines of research by which we can increase the likelihood of developing a safe superintelligent being. For example, it looks like developing correct epistemological standards will be essential for creating a safe superintelligent being. Indeed, if we create a superintelligent being with the correct goals (including an appropriate concern for the

interests of human beings) and the correct decision theory, we may still be in trouble if the being is radically deluded about the state of the world. To date, the problem of outlining correct epistemic standards has proven beyond the capacities of mankind – perhaps the problem is ill-defined. If the problem is beyond us, it may be that we can harness indirect-normativity, and allow a prospective superbeing to figure out the correct epistemic standards for itself. If we attempt this route, then we will at least need to articulate, in a precise and unambiguous manner, what problem it is that epistemic standards are intended to solve. Even here, the task has so far proven beyond the capacities of mankind. Beyond developing correct epistemic standards, or specifying the problem for which epistemic standards are the solution, it appears that we would need to complete similar projects, concerning the correct decision theory, and the correct means of evaluating human interests, if we are going to create a safe superintelligent being. A further point, which I did not find in Bostrom's book, is that implementing correct epistemic standards (whether directly or indirectly) will be essential, if we hope to use indirect-normativity in order to impart a superbeing with the correct decision theory, and an adequate means of evaluating human interests.

Bostrom presupposes a Bayesian picture of epistemic rationality, and thus maintains that specifying the correct epistemic standards amounts to selecting a correct prior probability function, at which point an agent's degrees of belief should be updated by conditionalization. Given this Bayesian picture, Bostrom argues that the problem of providing an artificial being with correct (or safe) epistemic standards may be less difficult than providing the being with suitable goals or the correct decision theory. Bostrom's optimisim concerning the provision of correct epistemic standards is based on formal results that show that updating relevantly similar prior probabilities by Bayesian conditionalization, given a sufficiently abundant and varied body of evidence (a body of evidence meeting certain formal conditions), yields converging posterior probabilities. In other words, modest differences in an artificial being's epistemic starting point will not prevent the being from reaching the same (presumed correct) conclusion that it would have reached had it had different prior probabilities (so long as the being's evidence meets certain conditions). Assuming that Bayesianism provides an adequate epistemic framework (which is an assumption that many would be unwilling to grant) and that our prospective superbeing has access to a body of evidence of the sort that would yield convergent posteriors (another big assumption), Bostrom appears to have overlooked the problem of providing adequate criteria concerning

when it is correct to regard a proposition as 'evident', and thereby a proper object upon which to conditionalize. This is a considerable problem, whose difficulty is accentuated when we consider it in the context of designing a safe superintelligent being.

Despite the preceding quibble, I think Bostrom is successful in defending the main thesis of his book, which is that there are credible reasons to believe that the development of artificial intelligence represents a significant risk to the future of humanity. Notice that the objection of the preceding paragraph maintains that Bostrom has *underestimated* the difficulty of solving one problem that is probably essential to the creation of a safe superintelligent being. On the other hand, one might think that Bostrom overestimates the likelihood of one or another scenario by which the creation of a superintelligent being leads to the end of humanity, perhaps because one thinks he overestimates the difficulty of solving or implementing one or another safety measure. Here it should be noted that Bostrom is careful to acknowledge how difficult it is to estimate the likelihood that certain possibilities will come to pass. Given the difficulty of making such judgments, and given the tremendous stakes, it appears that we ought to take Bostrom's warnings seriously. Assuming that we do take his warnings seriously, his book lays an excellent foundation from which future work concerning the creation of safe artificial intelligence may proceed.