Tristan Contant

Prof. Pruim

DATA-303, Calvin University

24 April 2023

Superintelligence: Book Report

Introduction

Published in 2004, *Superintelligence: Paths, Danger Strategies* is the newest book from Oxford University Professor Nick Bostrom. In a world where machine intelligence is rapidly advancing, Bostrom writes to lay a foundation of strategies to mitigate the potential drawbacks of superintelligent agents, intellects that are vastly superior to human intelligence in all nontrivial domains.

Nick Bostrom holds a variety of titles at Oxford. Beyond being a professor of philosophy, he is the director of the Future of Humanity Institute, an interdisciplinary research group that combines mathematics, philosophy, and social sciences to address the future of human society. Additionally, Bostrom is the director of the Strategic Artificial Intelligence Center, another multidisciplinary group at Oxford, this one developing strategies and tools for creating safe and benevolent artificial intelligence.

Superintelligence is structured so that each chapter addresses a different question. However, these questions are not disjoint; instead, the book begins with chapters that develop a framework of terminologies in which major theories are built upon in later chapters.

Summary

The book begins by explaining that superintelligence may not exist today, but it is something that will be coming—and sooner than most would expect. Bostrom compiled four

different surveys where experts were asked when they think human-level machine intelligence will exist. On average, experts said that there was a fifty percent chance that such an intellect would exist by 2050, and a ninety percent chance that it would exist by the end of the century. Including these statistics creates a sense of urgency for this topic. If we don't start preparing today, we could be caught off guard with issues that we are ill-equipped to mitigate.

Next, Bostrom outlines three different ways in which super intelligence can be dangerous. First, Bostrom notes that the direct reach of a superintelligent intellect, the tasks that it is readily able to perform given its structures, is not a bound to its indirect reach, all the things which a super intelligent intellect might be capable of doing. In other words, all superintelligent intellects, regardless of differences in their creation, can ultimately carry out the same tasks, and thus have the same indirect reach.

Second, Bostrom elucidates the different types of intelligence takeoffs, the rate at which an agent begins to acquire more intelligence. He defines this rate of change as a monotonically increasing function of two variables:

$$rate\ of\ change\ in\ intelligence\ = \frac{optimization\ power}{recalcitrance}$$

Optimization power refers to the "quality-rated design effort" to increase the agent's intelligence, and recalcitrance is the level of unresponsiveness to such optimization power. Therefore, a system would have a faster takeoff if it had more optimization power and less recalcitrance. Fast takeoffs, Bostrom notes, are problematic because it leaves humanity with little time to deliberate.

Third, superintelligent agents may pose existential crises based on their goals, or a Bostrom calls them, values. For instance, if a superintelligent intellect has a final goal of creating as many paperclips as possible, there is theoretically nothing to stop it from harvesting everything in its light cone of the universe into paperclips. This example might seem silly, but it

shows that, like a genie, it can be difficult to load the values that we intend into a superintelligent agent.

Another major theme of the book is a warning to not anthropomorphize superintelligence. This means that we should not rely on our own expectations of what a superintelligent agent's capabilities and motivations are. Bostrom explains that we have a very limited and biased outlook when it comes to intelligence. To humanity, a "village idiot" and "Einstein" are the extreme ends of the intelligence spectrum. However, in terms of mental capabilities, the two are relatively very similar. Bostrom claims that it will be much harder to form an intellect that has the intelligence of a village idiot than it will be to enhance such an intellect so that it is smarter than any human.

We can also fall prey to anthropomorphizing a superintelligent intellect's motivations. Bostrom points to our history of science fiction movies where a bud-eyed monster carries of a sexy woman in a torn dress. The creator of these movies never seems to ask why such a monster would even care about the woman in the first place. Instead, the creator placed is own values into the monster, rather than trying to get at what might be going on in the monster's brain. In the same way, we need to be sure not to anthropomorphize the motivations of supintelligence. There is nothing paradoxical about a superintelligent agent whose goal is to maximize the number of paperclips or calculate the digits of pi.

Reflection

While reading *Superintelligence*, I kept thinking about the different ethical frameworks that we learned about at the beginning of class: consequentialism, deontology, and virtue ethics. I think all three are important to consider when thinking about superintelligence and its relationship to future humanity.

Consequentialists like to think about the best solution to a problem by considering the consequences of the various possible actions that could be taken. Given how existentially destructive superintelligence can be, I think that a consequentialist framework would force us to be more cautious about superintelligence, as one small mistake could have massive implications for many people. Therefore, doing what's best for the most people, what's best for humanity, may often be the right things to do when it comes to superintelligence.

However, a deontological approach might work too. There may be some universal rules that should be outlined before people continue developing artificial intelligence. Having these agreed-upon rules, like rules for what values a superintelligent being could have, would be highly beneficial to the safety of humanity. Of course, there will be times when rules should be broken, but systems could be put in place so that one must get permission to break any given rule.

Lastly, I think that virtue ethics could be helpful to our development of superintelligence. Rather than only caring about the final goals a superintelligent agent has, we might consider how we could create an intellect that behaves virtuously. This is arguably harder to think about as most of our artificial intelligent machines are used for things like playing chess, not developing virtues. Nonetheless, this approach could highlight key ideas on the road to superintelligence.

All in all, I would not recommend *Superintelligence* to someone looking for a fun book to read on the beach. Bostrom's writing is highly technical and his ideas often took time for me to fully understand. However, if you want to think about the intersection of the future of technology and humanity, this book is for you.