Superintelligence or *The Technological Singularity*: A Critical Review

Nick Bostrom's *Superintelligence: Paths, Dangers, and Strategies* and Murray Shanahan's *The Technological Singularity*, a volume in the MIT Press Essential Knowledge Series, offer very different ways into the concepts of artificial superintelligence. Shanahan's book presents in brief several scenarios leading toward the technological singularity including some of Bostrom's, whereas Bostrom includes both an overview of artificial intelligence and more targeted arguments about the dangers inherent in the prospect of machine superintelligence. For the reader with a general interest in potential advancements in artificial intelligence, Shanahan lays out the main potential scenarios, while Bostrom delves deeper into the dystopian potential of recursive self-improvement and the reward functions of a potential machine superintelligence as well as offering a more in-depth look into the larger history of technological change as well as the history of AI research.

In addition to the differences in content and focus, Shanahan and Bostrom reveal starkly different attitudes toward humanity. Shanahan conveys wonder at the parallel processing power of the human brain and its ability to operate with very small amounts of energy (pp. 29-31). He marvels about the density and functions of neurons in the human brain and even the mouse brain. Indeed, the mouse brain serves as the basis for many of the thought experiments that Shanahan invites his readers to participate in. These invitations lend his introduction to the concept of technological singularity a welcoming and human-centric feeling even when he discusses troubling dystopian scenarios and skeptical view of artificial intelligence including Bostrom's.

Bostrom, on the other hand, sees human cognition as limited and shortsighted. In discussing the potential processing power of digital artificial intelligence, he enumerates features of it including processing and storage potential of computer hardware and finds that it could outstrip human cognition in nearly every respect (pp. 71-74). Bostrom's discussion of the "singleton," a potential single world power whether it be a country, a coalition, or a single superintelligent AI, begins with the historical example of post-World War II nuclear proliferation, and, it is both nuanced and allows for a reader for a non-technical background to understand the potential complexities as intelligent AIs begin to be developed by different organizations and world powers (pp. 106-109).

This discussion also illustrates one of the off-putting and potentially problematic facets of Bostrom's arguments; he often presents current patterns of human behavior and organization with an air of disdain. In describing potential organizational or political maneuvering, he chooses to describe them as "members of a conspiracy" rather than a faction, a subgroup, a political party, or any less loaded term (p. 108). In his afterword, he describes an overall improvement in the current research landscape for superintelligence, and even then, he describes popular interest and fiction about superintelligence as "popular cacophony" in contrast to a "more grownup conversation," of which he is presumably a part (p. 321). He even attacks his assumed readers (or in his framing, his assumed book buyers). He directs "one little remark- directed to new owners of the paperback edition" and goes on to assert that a large subset of them will not actually read the book but rather will glance at the front matter and the conclusion (p. 324). In doing so, he attempts to defend himself against a potentially charge of being overly pessimistic regarding superintelligence based purely on the page count of various sections of the book (p. 324).

This not too far off from how his original conclusion ends bemoaning “the fog of everyday trivialities” that obscures the truly important work of superintelligence research (p. 320). To borrow a phrase from popular culture that might rankle Bostrom, insulting the intelligence and intellectual stamina of your readers is not a good look. Lest one thinks that we as potential academic readers are a step above these imagined cases, Bostrom also lets it be known what he thinks of self-promoting academics like ourselves: “Information continence may be especially challenging for academic researchers, accustomed as they are to constantly disseminating their results on every available lamppost and tree” (p. 318). A “we” may have gone a long way to lessening the distance between Professors Bostrom and those other academics.

Perhaps my argument thus far has been too harsh on Bostrom; his discussion of human cognitive enhancement suggests that his harshest criticisms are reserved for current society rather than humanity overall (pp. 286-290). He presents potential cognitive acceleration as advantageous to human and technological development only under very specific circumstances, whereas he explains deceleration positively both by analogy and hypothetically. He describes life in the Pleistocene as “the kaleidoscope of human affairs” which “churned at a reasonable rate with births, deaths, and other personally and locally significant events” (p. 287). Similarly, he imagines a slowed future in which “international relations around the globe come to resemble those between the countries of the European Union” (p. 289). Clearly, this is a pre-Brexit argument.

Returning to Shanahan, his chapter “AI and Consciousness” offers several examples of a more charitable view of humanity, machine intelligence, and even animal consciousness (pp. 117-150). He adds a new element to the mouse thought experiment mentioned above wherein each neuron in the mouse’s brain is replaced until it is all synthetic and then the processed is gradually reversed (pp. 118-124). Shanahan argues that if consciousness persists throughout the process then synthetic consciousness and biological consciousness should considered equivalent. He also raises questions as to whether human use of conscious artificial intelligence would constitute slavery. Similarly, in his concluding chapter, Shanahan discusses the potential of instilling (or installing) morality in superintelligent machines (pp. 217-222). In keeping with his appreciation for humanity, Shanahan describes the complexity of human morality and the difficulty in adequately conveying it through computer programming. While some of the ideas presented in his conclusion are adapted directly from Bostrom’s work, Shanahan’s continual appreciation of humanity broadly speaking shifts the emphasis from skepticism and extreme caution to one caution mixed with optimism.

References

Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies*. OUP Oxford.

Shanahan, M. (2015). The technological singularity. MIT Press.