Book Essay

The Future of Artificial Intelligence

Allison Berke

Abstract

The first questions facing the development of artificial intelligence (AI), addressed by all three authors, are how likely it is that humanity will develop an artificial human-level intelligence at all, and when that might happen, with the implication that a human-level intelligence capable of utilizing abundantly available computational resources will quickly bootstrap itself into superintelligence. We need not imagine a doomsday scenario involving destructive, superintelligent AI to understand the difficulty of building safety and security into our digital tools.

✳ ✳ ✳ ✳ ✳

How to Create a Mind: The Secret of Human Thought Revealed by Ray Kurzweil. Penguin Books, 2012, 282 pp., $17.00.

Our Final Invention: Artificial Intelligence and the End of the Human Era by James Barrat. St. Martin's Griffin Press, 2013, 267 pp., $16.99.

Superintelligence: Paths, Dangers, Strategies by Nick Bostrom. Oxford University Press, 2014, 260 pp., $29.95.

Three recent popular science works explore the future of AI—examining its feasibility, its potential dangers, and its ethical and philosophical implications. Ray Kurzweil, an inventor, technologist, futurist, and AI pioneer—known for popularizing the concept of the *singularity* (a point at which technological progress in machine intelligence approaches runaway growth)—has in recent years devoted his efforts to machine learning and speech processing. Kurzweil's research, including that of companies he has founded, is centered on enabling computers to recognize speech and text, building individual capabilities necessary for general AI. In *How to Create a Mind*, Kurzweil summarizes recent advancements in neuroscience and software development to put forth an argument that the areas of the brain that produce a uniquely human intelligence—primarily the neocortex—are composed of a network of similar, hierarchically organized units responsible for executing nested pattern recognition algorithms. These algorithms can be translated into software via hierarchical hidden Markov models, and Kurzweil demonstrates that these models can be used to perform speech recognition and query analysis.[1] This approach to AI recognizes that rather than simulating an entire brain at the level of individual neurons, simulating its processes and results is computationally more efficient. The combined effect of Kurzweil's optimism and credentials gives the impression that AI is an attainable goal that technologists and inventors are inexorably approaching, a conclusion that may have spurred James Barrat, a documentary filmmaker with a focus on ancient history and inventions, to pen the case against AI in *Our Final Invention*. Barrat's interest in AI began when he interviewed Kurzweil in 2000, but his investigations into AI led to a more cautionary perspective, warning that superintelligent AI will be difficult or impossible to control, may be developed or motivated by the goals of our adversaries, and will likely resist or outmaneuver our efforts to design in controls and safety measures. Barrat points to many of the same technologies as Kurzweil—Siri, Apple's digital assistant; and Watson, IBM's *Jeopardy!*-winning, question-answering system factor prominently—but he anticipates a future in which Watson's descendants, tasked with improving human lives, ignore or misinterpret these instructions in favor of building more and better copies of themselves. This could lead, Barrat argues, to a depletion of the Earth's resources and the enslavement or eradication of humanity, as the self-improving AI departs for other planets in its quest to acquire more raw materials.

To this debate arrives Nick Bostrom, professor of philosophy at Oxford University and the founding director of Oxford's Future of Humanity Institute. Befitting his academic perspective, in *Superintelligence* Bostrom takes a broader view of AI development and outlines a framework for assessing the possibilities at each stage: how AI may be developed, how its intelligence can be measured, what problems AI will be used to address, where it may diverge from our intentions or abilities to control it, and what the implications of unleashing a superintelligent machine upon our society could be. Bostrom's book provides necessary context and vocabulary, allowing both sides of the debate to address the same questions.

The first questions facing the development of AI, addressed by all three authors, are how likely it is that humanity will develop an artificial human-level intelligence at all, and when that might happen, with the implication that a human-level intelligence capable of utilizing abundantly available computational resources will quickly bootstrap itself into superintelligence. Bostrom defers to the results of a survey of professionals, who place the development of human-level AI at 20 to 30 years in the future, a commonly postulated horizon that continually recedes as the technology in question fails to materialize. Researchers in the 1970s, after some of the first advances in machine learning and language processes, also predicted that human-level AI would be developed in 20 years. Kurzweil, befitting his position as a futurist, is invested in the fruition of this technology and cites his research on the exponential increases in related capabilities such as the number of transistors per chip, the number of operations per second performed by supercomputers, the cost of performing these calculations and of storing their output in digital memory, and the decreasing cost of transistors. His Law of Accelerating Returns proposes that the exponential growth we have observed thus far in the capacity and performance of computation technologies will impel a solution to the problem of digitally replicating human intelligence. Barrat's response to this prediction is to note that as long as we assign a nonzero probability to the development of AI, we must address its risks with the appropriate seriousness; a risk that threatens the existence of humanity, even at a low probability, is of greater urgency than a relatively certain but low- to moderate-level risk, such as the risk of a self-driving car injuring a pedestrian.

Having established AI as a problem worthy of discussion, the authors diverge in accordance with their interests. Kurzweil's assumption is that the reader will want to know how AI will be developed, with proofs of principle for the computational underpinnings of its methods. *How to Create a Mind* takes the reader on a tour of neocortical analysis, brain scanning, evolutionary algorithms, and programs like Siri and Watson that provide sophisticated solutions to carefully delineated problems of language analysis. Kurzweil touches briefly on the question of whether a human-level AI would be considered conscious; his conclusion is that, so long as the AI's responses are sufficiently convincing, we should not care, as qualia-like color perception and emotional experience are already subjective and internal. He hardly addresses whether the AI we build might destroy us. While acknowledging that nation-states have competitive incentives to build AI, to Kurzweil, AI will only be used to help humanity—as a symbiotic tool that will enhance our analytical and decision-making capabilities.

In contrast, Barrat sees the negative consequences of AI as intrinsic to its development, and he focuses instead on who will be motivated to construct an AI, what their motivations reveal about the goals they will program into their systems, and, therefore, how best to prepare for—or attempt to mitigate—the harms these systems will visit upon the world. Barrat draws sinister conclusions from the secrecy of large companies like Google, the funding aims of organizations like the Defense Advanced Research Projects Agency (DARPA), and the types of problems motivating defense contractors and foreign governments. While Kurzweil's AI will be a helpful savant—a child of Siri and Watson that aims to provide us with information while understanding our puns, accents, and wordplay—Barrat's AI is a killing machine, bent on global domination or unwittingly destroying the planet's resources to provide itself with more energy and silicon.

Which, then, is more likely? An AI that assists humanity and provides us with answers to problems we thought hopelessly intractable or an AI that remorselessly crushes us to better execute its code? The difficulty of answering this question stems from the fact that, as Bostrom outlines, both scenarios require us to evaluate concepts, like "superhuman intelligence," that exceed the scope of our experience. To define how we will recognize intelligence that is exponentially superior to ours, or the types of values and moral judgments with which we could imbue this intelligence to prevent it from harming us, we have to define concepts that have long stymied philosophers; presumably, if we all agreed completely on what outcomes are good for humanity, we would not need AI to tell us how to achieve them. The possibility of engineering initial conditions into our AI seedlings that will spur their development along moral and beneficial paths neglects the reality that we attempt to do this routinely, such as when we code software we assume is secure or even through the process of raising children—and are just as routinely surprised by unintended results. More prosaic goals, such as constructing an AI that can be kept isolated from other networks or an AI that does not seek to destroy other AIs are still subject to modes of failure that Bostrom characterizes as stemming from the available options for the motivations and capabilities that can be programmed into our AI.

Technologists may find such a philosophical conclusion unfulfilling, just as historians may find a preoccupation with the how, rather than the why, of AI development to be insufficiently imaginative. Kurzweil's and Barrat's works serve as complementary correctives, the former providing a solid base for understanding how we are approaching the development of AI and the latter a discussion of the hazards accompanying that approach. Bostrom's analysis requires more thought from the reader but provides a strong framework with which to organize that thought, stepping through the potential alternatives at each stage of AI development and deployment. Where all three volumes understandably fall short is in analogies to other technological developments—and attendant fears—that historically went unrealized. Technology skeptics occasioned an "AI winter" once, and those interested in the recent resurgence of funding and interest in AI are unwilling to dismiss it yet again as a goal too grandiose for debate. Yet there may be instructive parallels in the development of nuclear weapons or space travel; both were accompanied by grand and existentially threatening predictions that were averted by deliberate and strategic cooperation as well as by technological limitations and safeguards. Similarly, though Bostrom and Barrat describe AI component technologies, such as digital assistants or machine-learning algorithms that design circuits and identify faces, only to bolster the case that the development of full AI is fast approaching, the ethical problems involved in the control of AI are seen in microcosm in the question of what should happen when a self-driving car cannot avoid a crash or how Siri should respond to a suicidal user. We need not imagine a doomsday scenario involving destructive, superintelligent AI to understand the difficulty of building safety and security into our digital tools.

Notes

[1] A Markov process is usually characterized as *memorylessness*: the probability distribution of the next state depends only on the current state and not on the sequence of events that preceded it. In a hidden Markov process, the current state is not visible, but the output is visible. See the *Wikipedia* entry at https://en.wikipedia.org/wiki/Hidden_Markov_model, accessed 14 July 2016.

Disclaimer

The views and opinions expressed or implied in SSQ are those of the authors and are not officially sanctioned by any agency or department of the US government. We encourage you to send comments to: strategicstudiesquarterly@us.af.mil.