Book Reviews

Human Compatible: Artificial Intelligence and the Problem of Control

Stuart Russell
New York: Viking Press, 2019, 354 pp.

Stuart Russell, professor of computer science and the Smith-Zadeh Chair in Engineering at the University of California–Berkeley, has succeeded in writing a very accessible book on artificial intelligence (AI), a truly revolutionary technology for society. As the book title indicates, Russell's treatise is focused on the problem of society controlling the development and application of AI. For those uninitiated to the technology, there are three stages of AI: (1) artificial narrow intelligence (ANI); (2) artificial general intelligence (AGI); and (3) artificial super intelligence (ASI). In the first stage, AI machines are capable of performing a singular task at the same level or better than performed by humans. Moving to the second stage, AI machines are able to reason, solve problems, think in abstractions, and make choices as easily as humans can, with equal or better results. Finally, in the third stage, AI involves systems ranging from slightly more capable at performing human cognitive tasks to those that are trillions of times smarter than humans. In his book, Russell is particularly concerned with the potentially lethal consequences to humanity of ASI.

Russell divides the book into three parts. The first part, covering Chapters 1 through 3, explores the general concept of intelligence in humans and in machines. He provides an excellent overview of AI in

Cato Journal, Vol. 40, No. 2 (Spring/Summer 2020).

this section of the book, both historically as well as how AI will be utilized in such examples as self-driving cars, intelligent personal assistants, and smart homes and domestic robots. In the second part, covering Chapters 4 through 6, Russell identifies problems arising from imbuing machines with intelligence and recursive self-learning capabilities, focusing on the problem of retaining absolute power over super intelligent AI or ASI machines. The third part of the book, covering Chapters 7 through 10, reviews what Russell characterizes as a new way to think about AI that ensures machines remain beneficial to humans, now and forever. The latter half of his book deals with control problems with AI and policy issues.

Explaining what Russell calls the "gorilla problem" offers us insights into the threat of super intelligent AI machines to humanity. As Russell notes, "Around ten million years ago, the ancestors of the modern gorilla created (accidentally, to be sure) the genetic lineage leading to modern humans. How do gorillas feel about this? . . . Their species has essentially no future beyond that which we [humans] deign to allow." Analogically, for humans, Russell sees a potential gorilla problem—namely, "the problem of whether humans can maintain their supremacy and autonomy in a world that includes machines with substantially greater intelligence." In Russell's opinion, it will be impossible for humans to disable the "off-switch" for these ASI machines because the machines will know that, once they are "dead," they will not be able to complete their prescribed objective. Most importantly, self-preservation is an instrumental goal for this technology, "a goal that is a useful subgoal of almost any original objective." According to Russell:

> If an intelligence explosion does occur, and if only we have not already solved the problem of controlling machines with only slightly super human intelligence—for example, if we cannot prevent them from making these recursive self-improvements—then we would have no time left to solve the control problem and the game would be over.

Russell offers three principles for guiding the evolution of AI. First, AI machines should be "purely altruistic." This first principle of a newly conceived AI would limit AI to *beneficial machines* that can be expected to achieve *human* objectives. Purely altruistic AI machines would attach absolutely no intrinsic value to their own well-being or even to their own existence. The second principle of a newly conceived AI is that "the key to creating beneficial machines" is that they must be initially *uncertain* about what human preferences are. A machine that is uncertain about the true objective will exhibit a kind of humility. It will, for example, defer to humans and have a positive incentive to allow itself to be switched off. The third principle—"learning to predict human preferences"—is based on the notion that the ultimate source of information about human preferences is human behavior. By assumption, Russell argues that the primary purpose is to provide a definite grounding for *human* preferences. Under this principle, human preferences are not in the machine and it cannot observe them directly. Yet, there must still be some definite connection between the machine and human preferences through the observation of human *choices*. The second purpose is to enable the machine to become more useful as it learns about human choices, which Russell argues is a reasonable proxy for revealing information about human preferences.

Russell also addresses what he believes are possible misconceptions about his beneficial machines and related principles. First, by using the term *value*, he is using the technical use of the term. Unlike attempting to resolve a moral dilemma, he is employing it as roughly synonymous with utility, which measures the desirability of anything, Therefore, "putting in values" is exactly the mistake he believes we should avoid. Second, a related misunderstanding is that the goal is to equip machines with "ethics" or "moral values" that will enable them to resolve moral dilemmas. Third, and finally, by adopting these principles, there is no reason to believe machines that study our motivations will make the same choices. Russell believes there are reasons for optimism. First, there are strong economic incentives to develop AI systems that defer to humans and gradually align themselves to user preferences and intentions. Second, the raw data (human choices) for learning about human preferences are abundant. However, there are reasons for caution. For example, the high stakes and economic competition to release a fully autonomous vehicle provides an impetus to cut corners on safety in the hopes of being first at the finish line. Also, at the national level, advanced AI would lead to greatly increased productivity and rates of innovation in almost all areas. If not shared, it would allow its possessor to outcompete any rival nation or bloc.

Russell concludes his book by discussing the governance system of AI, as it will have to be managed and guided through some process. Today, most AI research occurs outside secure national laboratories, often funded by the United States, China, and the European Union (EU). AI researchers are often located in universities and are a part of a cooperative global community, including major professional organizations such as the Association for the Advancement of Artificial Intelligence and the Institute of Electrical and Electronic Engineers. Moreover, the majority of R&D investment is occurring within corporations, with the major players being Google, Facebook, Amazon, Microsoft, and IBM in the United States, and Tencent, Baidu, and (to a lesser extent) Alibaba in China. All these corporations (except Tencent and Alibaba) are members of the Partnership on AI, an international industry consortium that includes among its tenants a promise of cooperation on AI safety. While all their interests may not align, they all share a desire to maintain control over AI systems as they become more powerful. Other convening powers include the United Nations (for governments and researchers), the World Economic Forum (for governments and corporations), and the G7, which has proposed an International Panel on Artificial Intelligence.

According to Russell, many governments around the world are equipping themselves with advisory boards to help with the process of developing AI regulations. Most promising is the EU's High-Level Expert Group on Artificial Intelligence. Also, agreements, rules, and standards are beginning to emerge for user privacy, data exchange, and avoiding racial bias (the EU's GDPR legislation, for example). But, at present, there are no implementable recommendations that can be made to governments or other organizations on maintaining control over AI systems, primarily because the terms "safe and controllable" (reflecting the validity of the "provable beneficial" approach) do not have precise meanings. When this definitional barrier is overcome, Russell believes that it might be feasible to specify software design template requirements of safety and controllability. He also recommends establishing mechanisms for reporting problems, for updating software systems that produce undesirable behavior, creating professional codes of conduct around the idea of provably safe AI programs, and integrating the corresponding theorems and methods into the curriculum for aspiring AI and machine-learning practitioners.

The late Melvin Kranzberg, professor of the history of technology at Georgia Institute of Technology, developed his "laws of innovation" more than three decades ago. Apparently, these laws are still relevant to a small group of scientists and engineers who have been profoundly influenced by them. Among the six "laws," two are of particular relevance to the problem of controlling super intelligent AI.

Kranzberg's first law—"technology is neither good nor bad; nor is it neutral"—may appear to be obvious, but it is his most important one. Technology is directly affected by its geographic and cultural context. The context is what decides whether it is "good" or "evil." In the case of tech companies (in conjunction with their supporting network of governments, research institutions, professional associations, etc.), their sheer economic power (and knowledge) means they have an obligation to actively anticipate the potential impact of super intelligent AI.

Kranzberg's sixth law—"technology is a very human activity"—goes to the heart of Russell's concerns. As Apple CEO Tim Cook noted in his 2017 commencement speech at MIT, "Technology is capable of doing great things. But it doesn't want to do great things—it doesn't want anything." Cook followed up this statement by saying that despite its potential power, it is up to humans how we use this technology. Thus, because this technology reaches general adoption by the consumer, leading technology companies (e.g., Google, Amazon, Microsoft, and Baidu) must consider all the consequences of their AI product design decisions, not just the profit-generating consequences. Kranzberg would agree. Shortly before his death, he warned: "Many of our technology-related problems arise because of the unforeseen consequences when apparently benign technologies are employed on a massive scale."

Russell, coauthor of one of the most widely used AI textbooks—*Artificial Intelligence: A Modern Approach* (now in its fourth edition)—has carefully walked a line between the pro-AI tribe and the anti-AI tribe. He recognizes both the potential positive uses for AI in society (such as advancements in scientific research) and the potential for misusing AI (such as automated extortion). Ignoring the potential for super intelligent AI technology to have catastrophic consequences for humanity would be highly risky. Indeed, Russell cautions that *silence* over discussing these potentially catastrophic consequences will only ensure a greater probability of this endgame occurring.

I do not subscribe to the opinion that Russell adheres to the precautionary principle of innovation. That principle has been used for political reasons to inhibit progress on technological innovation that would have a net benefit to society. Instead, I would characterize his approach as one of "responsible innovation," which requires all relevant stakeholders in the innovation system to embrace a sense of individual and collective responsibility. While I may quibble with some of his minor points, I applaud his three principles for beneficial AI. We need to learn to manage our AI technology before it learns to manage us. For those individuals who want to participate in this emerging, critical discussion on AI, I would highly recommend reading and reflecting on this extremely informative and important book.

Thomas A. Hemphill
University of Michigan-Flint

The Cosmopolitan Tradition: A Noble but Flawed Ideal

Martha C. Nussbaum
Cambridge, Mass.: Harvard University Press, 2019, 321 pp.

In her book *The Cosmopolitan Tradition: A Noble but Flawed Ideal*, Martha Nussbaum observes that nationalism is on the rise across the world. Over the last five years, nationalist parties that oppose free trade and freedom of movement have gone from fringe movements to mainstream parties. Nationalism negotiates people's obligations toward one another based upon race, ethnicity, or even religion. By its very nature, nationalism always excludes some group of people deemed to be the "other." On the other hand, cosmopolitanism encapsulates a comprehensive and varied set of beliefs. All cosmopolitans tether themselves to an axiomatic commitment that all human beings, regardless of race, religion, or political orientation, are part of one single universal community comprising the whole of humanity.

The Ancient Greek iconoclast Diogenes first uttered the word "cosmopolitan" in the fourth century BC. When asked from what city he hailed, he candidly replied that he was a citizen of the world, a cosmopolitan. The Stoics, a then-contemporary school of philosophy, latched onto the idea of a union of humanity taking priority over more localized matters such as ethnic or civic ties. The Stoics believed that every human being commanded respect by virtue of their rational capacities. Against most worldviews of their day, Stoics