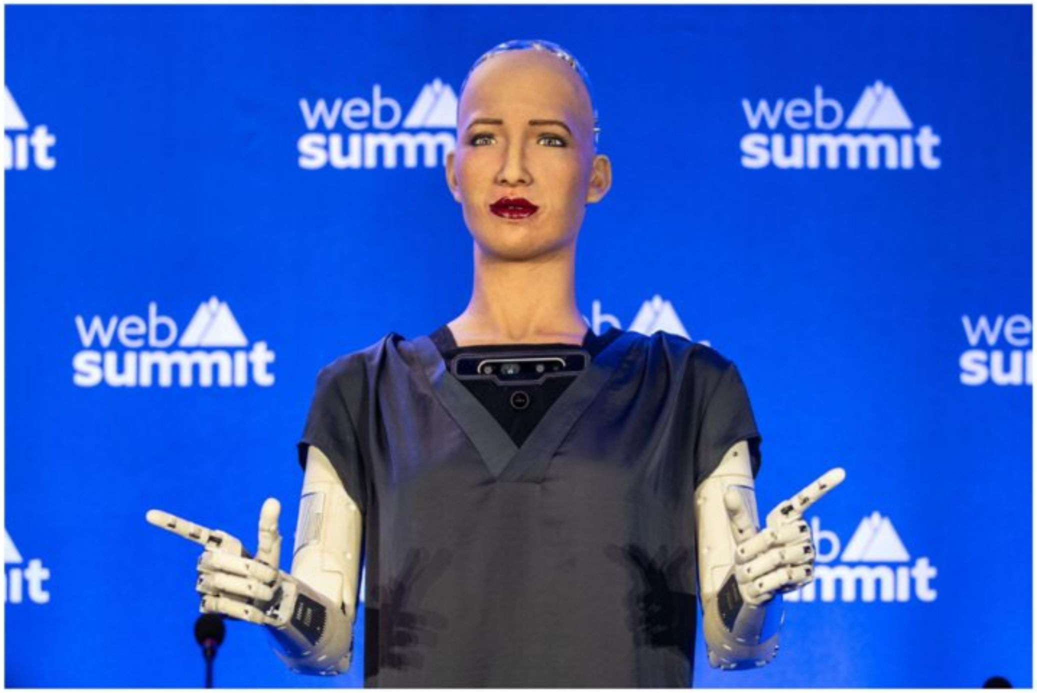
‘Human Compatible’ and ‘Artificial Intelligence’ Review: Learn Like a Machine

Before AI can solve the world’s problems, it must overcome the challenge of understanding humans—a feat we ourselves won’t achieve soon.

BOOKSHELF

By Matthew Hutson +

November 19, 2019 06:46 p.m. EST



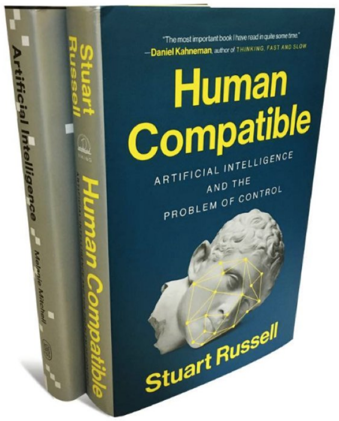
Journalists like to punctuate stories about the risks of artificial intelligence—particularly long-term, humanity-threatening risks—with images of the Terminator. The idea is that unchecked robots will rise up and kill us all. But such martial bodings overlook a perhaps more threatening model: Aladdin. Superhuman AI needn’t have it in for us to wreak havoc. Even with our best interests at metallic heart, an AI might misunderstand our intentions and, say, grant our wish of no more cancer by leaving no more people. Here the risk isn’t evil slaughterbots but an overly literal genie.

At least that’s the general concern raised in “Human Compatible” by Stuart Russell, a computer scientist at the University of California, Berkeley, who argues that what would happen if we achieved superhuman AI is “possibly the most important question facing humanity.” To those who deem the question premature, Mr. Russell counters, “If we were to detect a large asteroid on course to collide with the Earth in 2069, would we say it’s too soon to worry?”

Mr. Russell’s first few chapters outline the past, present and near future of AI. Broadly, the field has moved from hand-coded rules and symbols to software that collects data and finds patterns—so-called machine learning. Current systems can recognize images and spoken words after training on labeled examples without being given detailed instructions. An area of particular interest to Mr. Russell is reinforcement learning, in which software “agents” are set loose in the world (or a virtual world, such as a videogame) and learn by being rewarded for desirable behavior.

The trick is defining the right rewards so that the lessons the agent learns align with what you really want—the spirit, not the letter, of the mandate. One virtual agent racked up points in a racing game not by finishing the course but by spinning in circles crashing into things. You could pre-emptively spell out, "Do X, but not Y, unless Z" for every conceivable condition, watching the loopholes multiply, or you could use a technique that Mr. Russell has pioneered called inverse reinforcement learning. Instead of learning behavior to achieve given goals, as in reinforcement learning, an agent infers people's goals by observing their behavior.

Key to Mr. Russell's plan for beneficial AI is that agents should remain somewhat uncertain about our preferences. Like an eager-to-please dog, an AI should keep looking back to see whether we approve of its stick-chasing. That solves the off-switch problem.


"The most important book I have read in quite some time."
—Daniel Kahneman, author of THINKING, FAST AND SLOW
Human Compatible
ARTIFICIAL INTELLIGENCE AND THE PROBLEM OF CONTROL
Stuart Russell
Stuart Russell Human Compatible
Artificial Intelligence
Melanie Mitchell

A robot certain you want coffee above all else, by contrast, might kill you before letting you interfere with its directive.

Superhuman AI, Mr. Russell writes, could banish disease, empower people with personalized tutors and accountants, lift everyone out of poverty, and find solutions to climate change. But first it must overcome the challenge of understanding humans, a feat we ourselves are in no danger of achieving. Mr. Russell makes a delicious excursion into the philosophy of utilitarianism and the psychology of preferences. How do we balance the wants and well-being of different people, or the same person across time and mindsets? Is there even a singular self with coherent desires? Mr. Russell's exciting book goes deep, while sparkling with dry witticisms.

Melanie Mitchell, a computer scientist at Portland State University, is in the too-soon-to-worry camp. "My own opinion is that too much attention has been given to the risks from superintelligent AI," she writes in "Artificial Intelligence," "and far too little to deep learning's lack of reliability and transparency and its vulnerability to attacks."

After providing a history of AI, including quotes from experts who have both over- and underestimated computing's prospects to a laughable degree, Ms. Mitchell explores some of AI's main domains: visual recognition, reinforcement learning and language processing. In each area, she explains the nuts and bolts, praises headline-grabbing breakthroughs, and then gives a reality check to those who might see human-like general intelligence in narrow exploits.

Object-recognition software, for instance, can track pedestrians, detect tumors and sort photo libraries. But it doesn't understand the content the way we do. Its obtuseness becomes sharply apparent in so-called adversarial attacks, in which only minimal changes to an image (or a sound or text file) can fool an AI into misidentifying it. Such attacks even transfer to the real world. A stop sign with a few innocuous stickers becomes a speed-limit sign.

The researchers first elucidating such vulnerabilities in neural networks—machine-learning programs inspired by the brain's wiring—called them an "intriguing property." Ms. Mitchell writes, "Calling this an 'intriguing property' of neural networks is a little like calling a hole in the hull of a fancy cruise liner a 'thought-provoking facet' of the ship."

Ultimately, these systems lack common sense, a broad and often unspoken understanding of how the world works. Common sense, in turn, might require embodied experience in the world, plus the ability to abstract from it and form analogies. Much of Ms. Mitchell's academic work concerns helping AI form analogies. It hasn't progressed far. (No fault of hers.)

The problems of control and common sense that Mr. Russell and Ms. Mitchell lucidly portray—with significant overlap—come into play well before superintelligence arrives. When you ask Alexa or Siri to clear your schedule for a meeting, you don't want your child's dance recital deleted.

Just as we train machines to do our bidding, they must also train us to expect what they can and cannot do. Ms. Mitchell quotes the economist Sendhil Mullainathan: "I am far more afraid of machine stupidity than of machine intelligence." It's a good line, but it's not an either-or choice. The problem is both in combination.

Mr. Hutson is the author of "The 7 Laws of Magical Thinking: How Irrational Beliefs Keep Us Happy, Healthy, and Sane."