How AI Will Rewire Us

For better and for worse, robots will alter humans' capacity for altruism, love, and friendship.

NICHOLAS A. CHRISTAKIS APRIL 2019 ISSUE of *The Atlantic*

Fears about how robots might transform our lives have been a staple of science fiction for decades. In the 1940s, when widespread interaction between humans and artificial intelligence still seemed a distant prospect, Isaac Asimov posited his famous Three Laws of Robotics, which were intended to keep robots from hurting us. The first—"a robot may not injure a human being or, through inaction, allow a human being to come to harm"—followed from the understanding that robots would affect humans via direct interaction, for good and for ill. Think of classic sci-fi depictions: C-3PO and R2-D2 working with the Rebel Alliance to thwart the Empire in *Star Wars*, say, or HAL 9000 from *2001: A Space Odyssey* and Ava from *Ex Machina* plotting to murder their ostensible masters. But these imaginings were not focused on AI's broader and potentially more significant *social* effects—the ways AI could affect how we humans interact with one another.

Radical innovations have previously transformed the way humans live together, of course. The advent of cities sometime between 5,000 and 10,000 years ago meant a less nomadic existence and a higher population density. We adapted both individually and collectively (for instance, we may have evolved resistance to infections made more likely by these new circumstances). More recently, the invention of technologies including the printing press, the telephone, and the internet revolutionized how we store and communicate information.

As consequential as these innovations were, however, they did not change the fundamental aspects of human behavior that comprise what I call the "social suite": a crucial set of capacities we have evolved over hundreds of thousands of years, including love, friendship, cooperation, and teaching. The basic contours of these traits remain remarkably consistent throughout the world, regardless of whether a population is urban or rural, and whether or not it uses modern technology.

But adding artificial intelligence to our midst could be much more disruptive. Especially as machines are made to look and act like us and to insinuate themselves deeply into our lives, they may change how loving or friendly or kind we are—not just in our direct interactions with the machines in question, but in our interactions with one another.

Consider some experiments from my lab at Yale, where my colleagues and I have been exploring how such effects might play out. In one, we directed small groups of people to work with humanoid robots to lay railroad tracks in a virtual world. Each group consisted of three

people and a little blue-and-white robot sitting around a square table, working on tablets. The robot was programmed to make occasional errors—and to acknowledge them: "Sorry, guys, I made the mistake this round," it declared perkily. "I know it may be hard to believe, but robots make mistakes too."

As it turned out, this clumsy, confessional robot helped the groups perform *better*—by improving communication among the humans. They became more relaxed and conversational, consoling group members who stumbled and laughing together more often. Compared with the control groups, whose robot made only bland statements, the groups with a confessional robot were better able to collaborate.

In another, virtual experiment, we divided 4,000 human subjects into groups of about 20, and assigned each individual "friends" within the group; these friendships formed a social network. The groups were then assigned a task: Each person had to choose one of three colors, but no individual's color could match that of his or her assigned friends within the social network. Unknown to the subjects, some groups contained a few bots that were programmed to occasionally make mistakes. Humans who were directly connected to these bots grew more flexible, and tended to avoid getting stuck in a solution that might work for a given individual but not for the group as a whole. What's more, the resulting flexibility spread throughout the network, reaching even people who were not directly connected to the bots. As a consequence, groups with mistake-prone bots consistently outperformed groups containing bots that did not make mistakes. The bots helped the humans to help themselves.

Both of these studies demonstrate that in what I call "hybrid systems"—where people and robots interact socially—the right kind of AI can improve the way humans relate to one another. Other findings reinforce this. For instance, the political scientist Kevin Munger directed specific kinds of bots to intervene after people sent racist invective to other people online. He showed that, under certain circumstances, a bot that simply reminded the perpetrators that their target was a human being, one whose feelings might get hurt, could cause that person's use of racist speech to decline for more than a month.

But adding AI to our social environment can also make us behave less productively and less ethically. In yet another experiment, this one designed to explore how AI might affect the "tragedy of the commons"—the notion that individuals' self-centered actions may collectively damage their common interests—we gave several thousand subjects money to use over multiple rounds of an online game. In each round, subjects were told that they could either keep their money or donate some or all of it to their neighbors. If they made a donation, we would match it, doubling the money their neighbors received. Early in the game, two-thirds of players acted altruistically. After all, they realized that being generous to their neighbors in one round might prompt their neighbors to be generous to them in the next one, establishing a norm of reciprocity. From a selfish and short-term point of view, however, the best outcome would be to keep your own money *and* receive money from your neighbors. In this experiment, we found that by adding just a few bots (posing as human players) that behaved

in a selfish, free-riding way, we could drive the group to behave similarly. Eventually, the human players ceased cooperating altogether. The bots thus converted a group of generous people into selfish jerks.

Let's pause to contemplate the implications of this finding. Cooperation is a key feature of our species, essential for social life. And trust and generosity are crucial in differentiating successful groups from unsuccessful ones. If everyone pitches in and sacrifices in order to help the group, everyone should benefit. When this behavior breaks down, however, the very notion of a public good disappears, and everyone suffers. The fact that AI might meaningfully reduce our ability to work together is extremely concerning.

Already, we are encountering real-world examples of how AI can corrupt human relations outside the laboratory. A study examining 5.7 million Twitter users in the run-up to the 2016 U.S. presidential election found that trolling and malicious Russian accounts—including ones operated by bots—were regularly retweeted in a similar manner to other, unmalicious accounts, influencing conservative users particularly strongly. By taking advantage of humans' cooperative nature and our interest in teaching one another—both features of the social suite—the bots affected even humans with whom they did not interact directly, helping to polarize the country's electorate.

Other social effects of simple types of AI play out around us daily. Parents, watching their children bark rude commands at digital assistants such as Alexa or Siri, have begun to worry that this rudeness will leach into the way kids treat people, or that kids' relationships with artificially intelligent machines will interfere with, or even preempt, human relationships. Children who grow up relating to AI in lieu of people might not acquire "the equipment for empathic connection," Sherry Turkle, the MIT expert on technology and society, told *The Atlantic*'s Alexis C. Madrigal not long ago, after he'd bought a toy robot for his son.

As digital assistants become ubiquitous, we are becoming accustomed to talking to them as though they were sentient; writing in these pages last year, Judith Shulevitz described how some of us are starting to treat them as confidants, or even as friends and therapists. Shulevitz herself says she confesses things to Google Assistant that she wouldn't tell her husband. If we grow more comfortable talking intimately to our devices, what happens to our human marriages and friendships? Thanks to commercial imperatives, designers and programmers typically create devices whose responses make us feel better—but may not help us be self-reflective or contemplate painful truths. As AI permeates our lives, we must confront the possibility that it will stunt our emotions and inhibit deep human connections, leaving our relationships with one another less reciprocal, or shallower, or more narcissistic.

All of this could end up transforming human society in unintended ways that we need to reckon with as a polity. Do we want machines to affect whether and how children are kind? Do we want machines to affect how adults have sex?

Kathleen Richardson, an anthropologist at De Montfort University in the U.K., worries a lot about the latter question. As the director of the Campaign Against Sex Robots—and, yes, sex robots are enough of an incipient phenomenon that a campaign against them isn't entirely premature—she warns that they will be dehumanizing and could lead users to retreat from real intimacy. We might even progress from treating robots as instruments for sexual gratification to treating other people that way. Other observers have suggested that robots could radically improve sex between humans. In his 2007 book, *Love and Sex With Robots*, the iconoclastic chess master turned businessman David Levy considers the positive implications of "romantically attractive and sexually desirable robots." He suggests that some people will come to prefer robot mates to human ones (a prediction borne out by the Japanese man who "married" an artificially intelligent hologram last year). Sex robots won't be susceptible to sexually transmitted diseases or unwanted pregnancies. And they could provide opportunities for shame-free experimentation and practice—thus helping humans become "virtuoso lovers." For these and other reasons, Levy believes that sex with robots will come to be seen as ethical, and perhaps in some cases expected.

Long before most of us encounter AI dilemmas this intimate, we will wrestle with more quotidian challenges. The age of driverless cars, after all, is upon us. These vehicles promise to substantially reduce the fatigue and distraction that bedevil human drivers, thereby preventing accidents. But what other effects might they have on people? Driving is a very modern kind of social interaction, requiring high levels of cooperation and social coordination. I worry that driverless cars, by depriving us of an occasion to exercise these abilities, could contribute to their atrophy.

Not only will these vehicles be programmed to take over driving duties and hence to usurp from humans the power to make moral judgments (for example, about which pedestrian to hit when a collision is inevitable), they will also affect humans with whom they've had no direct contact. For instance, drivers who have steered awhile alongside an autonomous vehicle traveling at a steady, invariant speed might be lulled into driving less attentively, thereby *increasing* their likelihood of accidents once they've moved to a part of the highway occupied only by human drivers. Alternatively, experience may reveal that driving alongside autonomous vehicles traveling in perfect accordance with traffic laws actually improves human performance.

Either way, we would be reckless to unleash new forms of AI without first taking such social spillovers—or externalities, as they're often called —into account. We must apply the same effort and ingenuity that we apply to the hardware and software that make self-driving cars possible to managing AI's potential ripple effects on those outside the car. After all, we mandate brake lights on the back of your car not just, or even primarily, for your benefit, but for the sake of the people behind you.

In 1985, some four decades after Isaac Asimov introduced his laws of robotics, he added another to his list: A robot should never do anything that could harm humanity. But he struggled with how to assess such harm. "A human being is a concrete object," he later

wrote. “Injury to a person can be estimated and judged. Humanity is an abstraction.”

Focusing specifically on social spillovers can help. Spillovers in other arenas lead to rules, laws, and demands for democratic oversight. Whether we’re talking about a corporation polluting the water supply or an individual spreading secondhand smoke in an office building, as soon as some people’s actions start affecting other people, society may intervene. Because the effects of AI on human-to-human interaction stand to be intense and far-reaching, and the advances rapid and broad, we must investigate systematically what second-order effects might emerge, and discuss how to regulate them on behalf of the common good.

Already, a diverse group of researchers and practitioners—computer scientists, engineers, zoologists, and social scientists, among others—is coming together to develop the field of “machine behavior,” in hopes of putting our understanding of AI on a sounder theoretical and technical foundation. This field does not see robots merely as human-made objects, but as a new class of social actors.

The inquiry is urgent. In the not-distant future, AI-endowed machines may, by virtue of either programming or independent learning (a capacity we will have given them), come to exhibit forms of intelligence and behavior that seem strange compared with our own. We will need to quickly differentiate the behaviors that are merely bizarre from the ones that truly threaten us. The aspects of AI that should concern us most are the ones that affect the core aspects of human social life—the traits that have enabled our species’ survival over the millennia.

The Enlightenment philosopher Thomas Hobbes argued that humans needed a collective agreement to keep us from being disorganized and cruel. He was wrong. Long before we formed governments, evolution equipped humans with a social suite that allowed us to live together peacefully and effectively. In the pre-AI world, the genetically inherited capacities for love, friendship, cooperation, and teaching have continued to help us to live communally.

Unfortunately, humans do not have the time to evolve comparable innate capacities to live with robots. We must therefore take steps to ensure that they can live nondestructively with us. As AI insinuates itself more fully into our lives, we may yet require a new social contract —one with machines rather than with other humans.