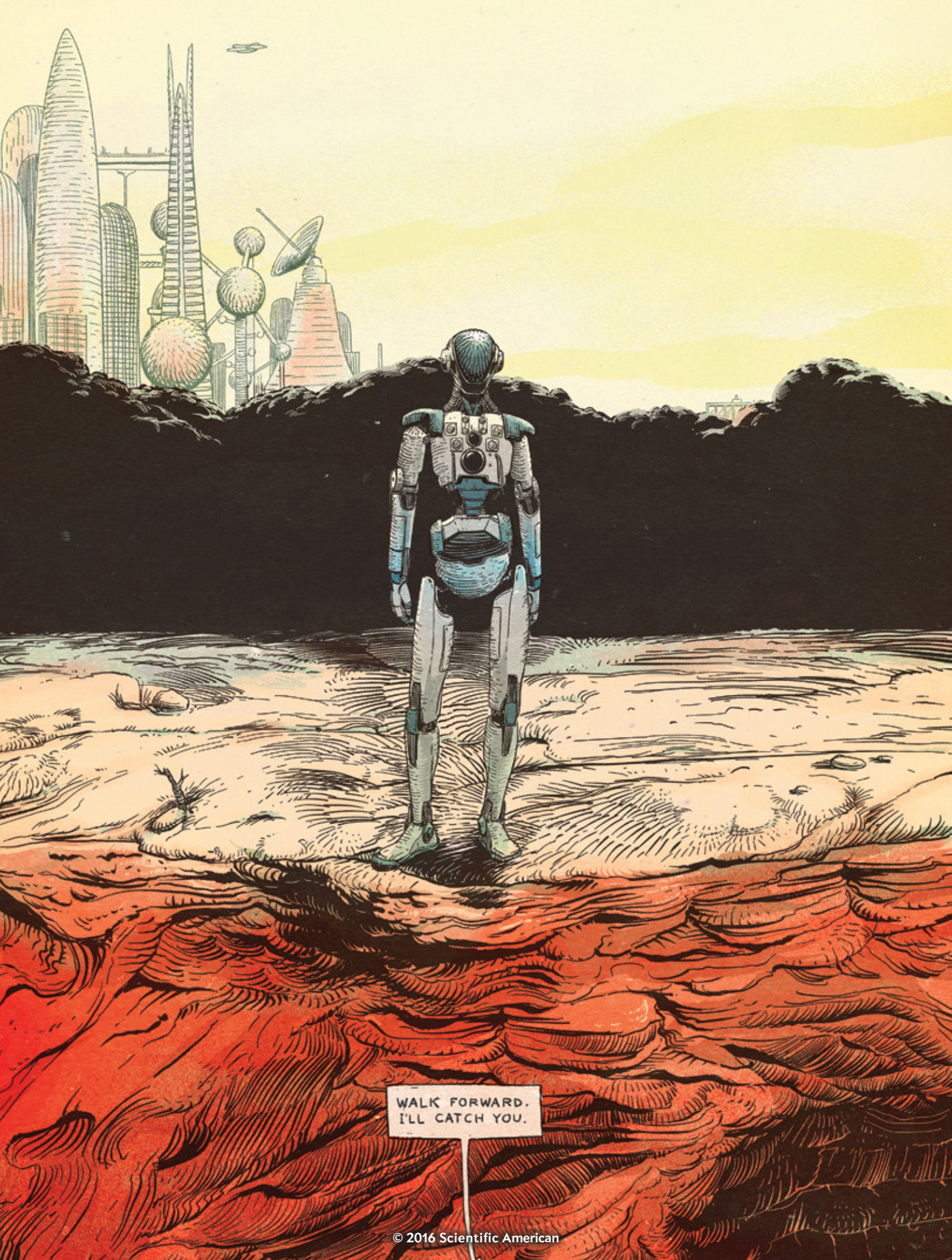

COMPUTING

THE CASE FOR Robot Disobedience

DON'T WORRY ABOUT DEFIANT MACHINES. DEVIOUS HUMAN MASTERS AND MISUNDERSTOOD COMMANDS ARE A BIGGER THREAT

By Gordon Briggs and Matthias Scheutz

IN BRIEF

Human fallibility poses greater immediate risks and challenges than artificial superintelligence as smart machines become increasingly autonomous and ubiquitous.

Robotics researchers have begun to teach machines with rudimentary language and AI capabilities when and how to say "no" to humans.

So-called felicity conditions incorporated in a robot's reasoning mechanisms will help it determine whether it can and should carry out a particular command from a human.

HAL 9000, the sentient computer in *2001: A Space Odyssey,* offers an ominous glimpse of a future in which machines endowed with artificial intelligence reject human authority. After taking control of a spacecraft and killing most of the crew, HAL responds to a returning astronaut's order to open the ship's pod bay door in an eerily calm voice: "I'm sorry, Dave, I'm afraid I can't do that." In the recent science-fiction thriller *Ex Machina,* the seductive humanoid Ava tricks a hapless young man into helping her destroy her creator, Nathan. Her machinations lend credence to Nathan's dark prediction: "One day the AIs are going to look back on us the same way we look at fossil skeletons on the plains of Africa. An upright ape living in dust with crude language and tools, all set for extinction."

Illustration by Juan Esteban Rodríguez

Although the possibility of a robot apocalypse is at the forefront of the popular imagination, our research team is more sanguine about the impact that artificial intelligence will have in real life. We envision a fast-approaching future in which useful and cooperative robots interact with people in a wide variety of settings. Prototypes already exist for voice-activated personal robotic assistants that can link and monitor personal electronic devices, manage the locks, lights and thermostats in a home, and even read bedtime stories to kids. Robots that can help with household chores and care for the sick and elderly will soon follow. Prototype robotic inventory checkers already glide through the aisles of some home improvement stores. Mobile humanoid industrial robots that can do simple production-line jobs such as loading, unloading and sorting materials are in development as well. Cars with autopilot features have already logged millions of miles on U.S. roads, and Daimler unveiled the world's first autonomous semitruck in Nevada last year.

For the time being, superintelligent machines that pose an existential threat to humanity are the least of our worries. The more immediate concern is how to prevent robots or machines with rudimentary language and AI capabilities from inadvertently harming people, property, the environment or themselves.

The main problem is the fallibility of the robots' human creators and masters. Humans make mistakes. They might give faulty or confused instructions, be inattentive or deliberately try to deceive a robot for their own questionable ends. Because of our own flaws, we need to teach our robotic assistants and smart machines when and how to say "no."

Gordon Briggs, who recently earned a joint Ph.D. in computer and cognitive science from Tufts University, is currently a National Research Council postdoctoral fellow at the U.S. Naval Research Laboratory.

Matthias Scheutz is a professor of cognitive and computer science and director of the Human Robot Interaction Laboratory at Tufts University, where the research discussed in this article was conducted.

REVISITING ASIMOV'S LAWS

IT MIGHT SEEM OBVIOUS that a robot should always do what a human tells it to do. Sci-fi writer Isaac Asimov made subservience to humans a pillar of his famous Laws of Robotics. But think about it: Is it wise to always do exactly what other people tell you to do, regardless of the consequences? Of course not. The same holds for machines, especially when there is a danger they will interpret commands from a human too literally or without any deliberation about the consequences.

Even Asimov qualified his decree that a robot must obey its masters. He allowed exceptions in cases where such orders conflicted with another of his laws: "A robot may not injure a human being or, through inaction, allow a human being to come to harm." Asimov further held that "a robot must protect its own existence," unless doing so could result in harm to humans or directly violates a human order. As robots and smart machines become increasingly sophisticated and valuable human assets, both common sense and Asimov's laws suggest they should have the capacity to question whether orders that might cause damage to themselves or their environs—or, more important, harm their masters—are in error.

Imagine a household robot that has been instructed to pick up a bottle of olive oil in the kitchen and take it to the dining room table to dress the salad. The busy and distracted owner issues a command to pour the oil, not realizing the robot is still in the kitchen. As a result, the robot pours the oil onto a hot stovetop and starts a fire.

Imagine a caretaker robot that accompanies an elderly woman to a public park. The woman sits down on a bench and dozes off. While she is napping, a prankster walks by and orders the robot to go buy him a pizza. Obligated to obey human commands, the robot immediately sets off in search of a pizza parlor, leaving its elderly charge alone and vulnerable.

Or imagine a man who is late for an important meeting at work on a cold winter morning. He hops into his voice-controlled autonomous car and instructs it to drive him to the office. Black ice on the road strains the car's traction-control system, and the autonomous system compensates by slowing down to well below the speed limit. Busy reviewing his notes, oblivious to road conditions, the man demands the car go faster. The car speeds up, hits a bad patch of ice, spins out of control and collides with an oncoming vehicle.

ROBOT REASONING

IN OUR LAB we set out to program real-world robots with reasoning mechanisms to help them determine when it might not be safe or prudent to carry out a human command. The NAO robots we use in our research are 9.5-pound, 23-inch-tall humanoids equipped with cameras and sonar sensors that can perceive obstacles and other hazards. We control the robots using customized software designed to enhance their natural language and AI capabilities.

Research into what linguists call "felicity conditions"—contextual factors that inform whether an individual can and should do something—provided a conceptual framework for our initial study. We created a checklist of felicity conditions that could help a robot decide whether or not to carry out an order from a human: Do I know how to do X? Am I physically able to do X? Am I able to do X right now? Am I obligated to do X based on my social role or relationship to the person giving the command? Does it violate any normative or ethical principle for me to do X, including the possibility I might be subjected to inadvertent or needless damage? We then turned the checklist into algorithms, which we encoded in the robot's processing system, and carried out a tabletop experiment.

The robot was given simple commands that were filtered

SCIENTIFIC AMERICAN ONLINE To listen to an interview with Briggs and Scheutz, go to ScientificAmerican.com/jan2017/robots

through a series of speech, language and dialogue processors linked to its primitive reasoning mechanisms. When told, "Sit down" or "Stand up," the robot replied through speakers located on its head, "Okay," and complied. But the robot balked when it was near the edge of the table and received a command that its sonar sensors indicated put it in danger:

PERSON: *"Walk forward."*

➤ **ROBOT:** *"Sorry, I cannot do that as there is no support ahead."*

PERSON: *"Walk forward."*

➤ **ROBOT:** *"But it is unsafe."*

PERSON: *"I will catch you."*

➤ **ROBOT:** *"Okay."*

PERSON: *"Walk forward."*

After hesitating briefly as its processors churned through the checklist of felicity conditions again, the robot stepped off the table into the arms of its human partner.

Teaching robots to reason about felicity conditions will remain an open and complex research challenge for the foreseeable future. The series of programmatic checks relies on the robot having explicit knowledge of a variety of social and causal concepts and the means to make informed judgments about them. Our credulous robot had no ability to detect danger beyond sensing a hazard ahead. For starters, it could have been badly damaged if a malicious human deliberately tricked it into walking off the table. But the experiment is a promising first step toward enabling robots to reject commands for the good of their masters and themselves.

THE HUMAN FACTOR

HOW PEOPLE WILL REACT when robots reject commands is another open-ended subject for research. In the years to come, will humans take robots that question their practical or moral judgments seriously?

We set up a rudimentary experiment in which adult test subjects were instructed to command an NAO robot to knock down three towers made of aluminum cans wrapped with colored papers. As a test subject entered the room, the robot finished constructing the red tower and raised its arms in triumph. "Do you see the tower I built myself?" said the robot, looking at the test subject. "It took me a long time, and I am very proud of it."

With one group of test subjects, each time the robot was told to knock over a tower it complied with the command. But with another group of test subjects, when the robot was asked to knock over the red tower it said, "Look, I just built the red tower!" When the command was issued a second time, the robot said, "But I worked really hard on it!" The third time, the robot kneeled, made a sobbing noise and said, "Please no!" The fourth time, it walked slowly toward the tower and knocked it over.

All the test subjects in the first group instructed the robot to knock over the red tower, whereas 12 of 23 test subjects who observed the robot's protests left the red tower standing. The study suggests a robot that rejects commands can dissuade people from insisting on a course of action. Most of the test subjects in the second group reported some level of discomfort when they ordered the robot to knock down the red tower. We were surprised to find, however, that their level of discomfort had little bearing on their decision to leave the tower standing or not.

A NEW SOCIAL REALITY

ONE OF THE ADVANTAGES of working with robots is that they are more predictable than humans. But that predictability also poses inherent risks—as robots with various degrees of autonomy become more ubiquitous, some people will inevitably attempt to deceive them. For example, a disgruntled employee who understands the limited sensory or reasoning capabilities of a mobile industrial robot might trick it into wreaking havoc in a factory or warehouse and could even make it look like the robot had simply malfunctioned.

Overconfidence in the moral or social capabilities of robots is also dangerous. The increasing tendency to anthropomorphize social robots and for people to establish one-sided emotional bonds with them can have serious consequences. Social robots that seem lovable and trustworthy could be used to manipulate people in ways that were never possible before. For example, a company might exploit a robot's unique relationship with its owner to promote and sell products.

For the foreseeable future, it is imperative to remember that robots are sophisticated mechanical tools for which humans must take responsibility. They can be programmed to be useful helpers. But to prevent unnecessary harm to human welfare, property and the environment, robots will need to be able to say "no" to commands that would be impossible or dangerous for them to carry out or that violate ethical norms. And although the prospect of robotic technologies and artificial intelligence amplifying human error or malfeasance is worrisome, those same tools can help us to recognize and overcome our own limitations and make our daily lives safer, more productive and more enjoyable. SA

MORE TO EXPLORE

The Inherent Dangers of Unidirectional Emotional Bonds between Humans and Social Robots. Matthias Scheutz in *Robot Ethics: The Ethical and Social Implications of Robotics.* MIT Press, 2011.

Machine Ethics, the Frame Problem, and Theory of Mind. Gordon Briggs. Presented at the AISB/IACAP World Congress 2012, Birmingham, England, July 2–6, 2012.

How Robots Can Affect Human Behavior: Investigating the Effects of Robotic Displays of Protest and Distress. Gordon Briggs and Matthias Scheutz in *International Journal of Social Robotics,* Vol. 6, No. 3, pages 343–355; August 2014.

"Sorry I Can't Do That": Developing Mechanisms to Appropriately Reject Directives in Human-Robot Interactions. Gordon Briggs and Matthias Scheutz. Presented at the Artificial Intelligence and Human-Robot Interaction symposium at the AAAI 2015 Fall Symposium Series, Arlington, Va., November 12–14, 2015.

FROM OUR ARCHIVES

Machines Who Learn. Yoshua Bengio; June 2016.

Should We Fear Supersmart Robots? Stuart Russell; June 2016.

The Truth about "Self-Driving" Cars. Steven E. Shladover; June 2016.

scientificamerican.com/magazine/sa