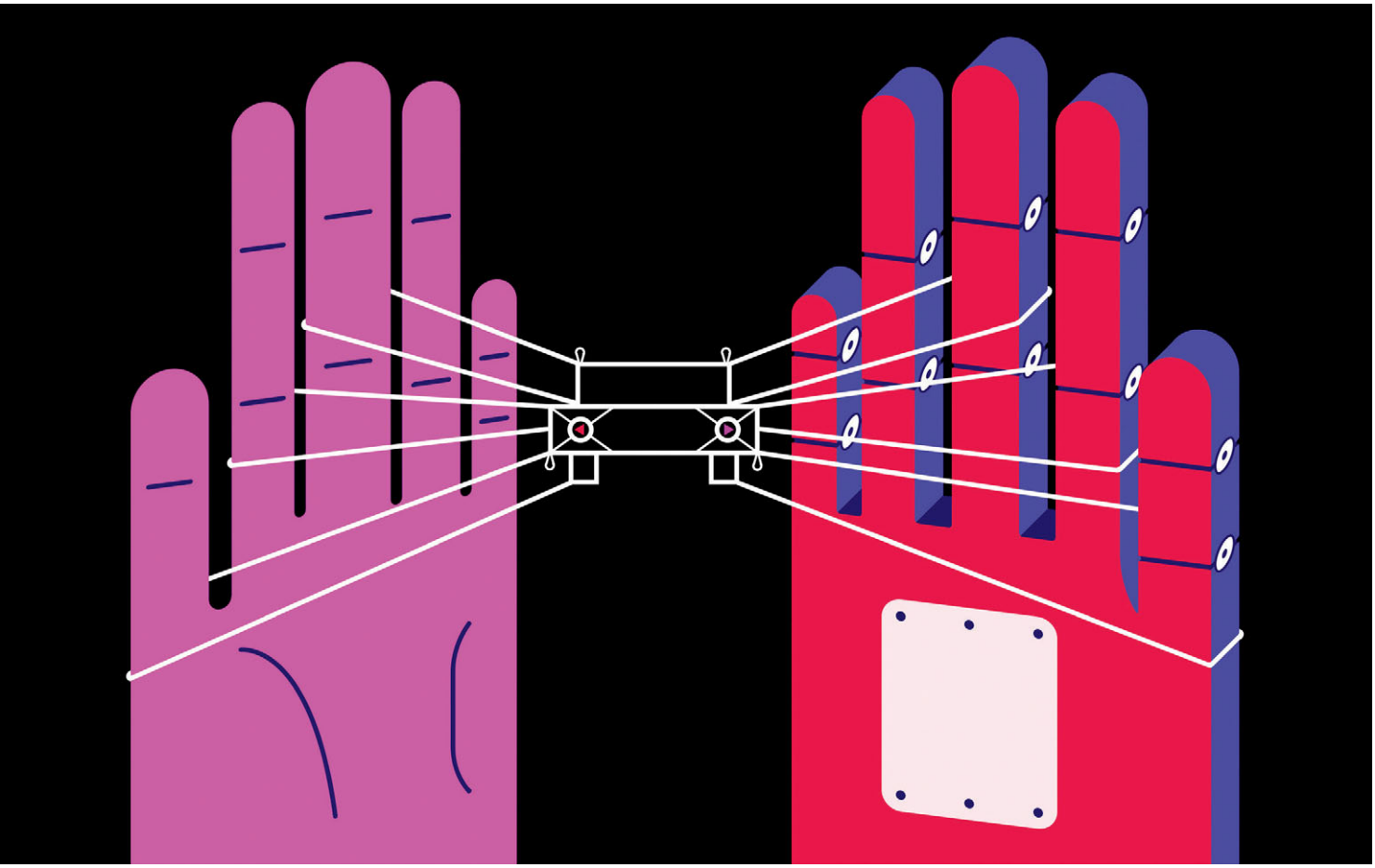

Humans and Computers Learn to Trust Each Other

Parallel neural networks may be key / BY DAVID LEVINE

In an age of self-driving cars and autonomous drones, artificial intelligence (AI) is becoming a bigger part of our lives. It's also becoming increasingly savvy. Today, AI can recognize text, distinguish people by their faces, and identify some physical objects. But even the best AI systems still make mistakes.

THAT POSES A BIG PROBLEM, SAYS KATE SAENKO, a College of Arts & Sciences assistant professor of computer science. "If an AI tool makes mistakes, human users quickly learn to discount it and eventually stop using it altogether," she says. "I think that humans by nature are not likely to just accept things that a machine tells them."

A further complication, she adds, is that as AI becomes more powerful, the algorithms that drive it have become

ILLUSTRATION BY MENGXIN LI

increasingly opaque to human users. Information goes into one end of a computational "black box," and an answer comes out the other side—yet the set of rules and reasoning used to find that answer are obscured.

Saenko is working to change that. She and collaborator Trevor Darrell at the University of California, Berkeley, are recipients of a four-year, $7.55 million grant from the Defense Advanced Research Projects Agency, or DARPA. Saenko's lab at BU will receive $800,000 for research that seeks to uncover new ways of getting inside the "mind" of AI and creating a translation tool that explains its decision-making process to human users.

That goal may sound trivial. Who really cares how a computer came to a conclusion, as long as it's right? Getting feedback on why an AI device makes a particular decision, however, could improve its accuracy by providing opportunities for humans to offer tiny course corrections, Saenko says. This process could increase the trust that humans put into a machine, making it a better collaborator on complex jobs. Achieving that sort of openness in today's AI, though, may not be so simple.

Deep Neural Networks

It hasn't always been hard to look inside the mind of AI. In the past, many artificial intelligence systems, like facial recognition, used rules and guidelines that programmers identified ahead of time—rules for defining skin color, for what shapes make up a nose, for defining light and shadow. All those user-created concepts had to be hard-coded into AI from the start, giving it a framework to do its job.

That made it easy to figure out how a machine came to its conclusion: just identify which preprogrammed rules it used to get there. But it also limits the abilities of AI. Real life is vastly complex, after all, and even the best human programmers can't come up with every possible rule that a computer might use to make sense of the world. "It's very hard for us to anticipate all possible ways a dog might look in any image anywhere in the world, for example," says Saenko. "If you have enough processing power and data, a better approach would be to show a computer a million pictures of dogs and let it define them itself."

In the past five years or so, that approach has become more widely used. Instead of working with a single template, new systems involve a more iterative approach, modeled on the way that our own nervous system works. These new types of AI, called "deep neural networks," employ huge numbers of interconnected functions, or nodes, arranged in a vast web. Each one is responsible for parsing a tiny amount of information and progressively builds on the work of the nodes before it.

This sort of incremental process, building bit by bit on simple data, is at the core of a deep neural network. It makes AI flexible, fast, and powerful—for some systems, it can operate with more than 95 percent accuracy. In those few cases where AI is not accurate, though, deep neural networks make it extremely hard to figure out why. There are no preset coded definitions to turn to, since a neural network creates those big-picture guidelines as it goes.

An "Interpreter" for AI

Saenko and Darrell are working with Zeynep Akata, a colleague at the University of Amsterdam in the Netherlands, and Kitware, an open-source software company, on ways to make deep neural networks more easily understood. Asking a network like this to explain itself would likely reduce its speed and efficiency, the researchers say, so they're hoping to create a sort of translation tool—a second network that acts alongside the first, interpreting its choices in real time and reporting them to a human user.

This translation is important, Saenko says, because when a deep neural network does make a mistake, it's probably because it found a pattern in the data that doesn't quite match the real world. If it's steering an autonomous car along a poorly maintained road, for instance, it might stop at a shadow, thinking it's a pothole.

If that happens, an "interpreter" AI could prompt a user for more information in plain English. "I want it to be able to say, 'I stopped driving because I'm not sure if that's a pothole or a shadow, so tell me what to do here,'" she says. "In the future, we're going to be using AI as a collaboration between humans and computers. We need to be able to communicate with it, understand its strengths, so it can help us with things we're not so good at."

JACKIE RICCIARDI

▼ Kate Saenko wants to create AI that can explain its decision-making process to human users.