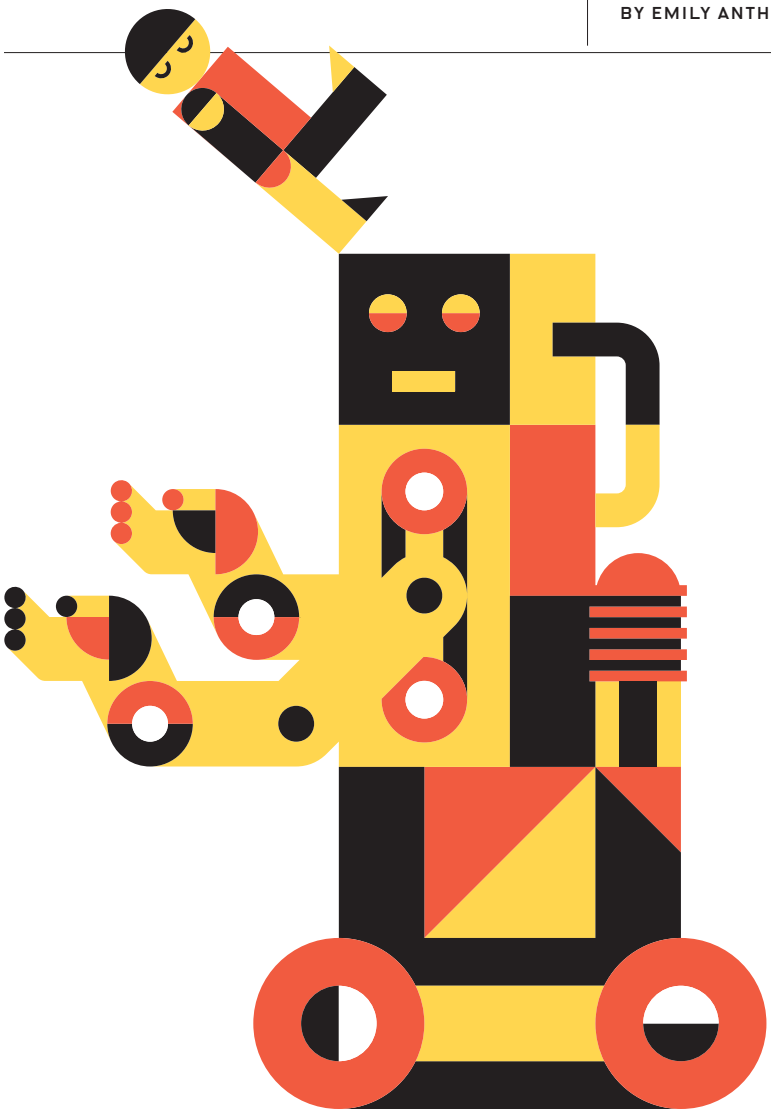
BY EMILY ANTHES



TRUSTING OUR ROBOTS

WE LOVE THEM WHEN THEY ADMIT FAULT

A ROBOT COULD save your life, if you're smart enough to let it. Just think about self-driving cars, which (at the very least) would eliminate road deaths from drunk and distracted driving, of which there are currently more than 13,000 in the United States every year. But most Americans say they wouldn't feel comfortable riding in an automated vehicle—and if that mentality doesn't change, there's no chance of the technology taking off. It's the same across many different domains, where robotic and artificially intelligent

LAMOSCA

systems have the potential to improve safety and boost productivity. Ironically, the crucial factor in the success of automation is always the one human action that can't be automated: the decision to trust the robot in the first place.

And even when the robot isn't calling all the shots—when we're just working with a bot—we usually keep our guards up. When robots do things we don't understand, like sensing obstacles we can't or following rules we don't know, we tend to lose confidence and wrest control away from them—even when the robots are right. Laboratory studies have also shown how easy it is to shake our faith: When a system designed to warn drivers of impending collisions was prone to false alarms, users' trust declined precipitously, despite the fact that the system never missed a true threat. In one study, 81 percent of volunteers chose to abandon a program they were told could predict whether camouflaged soldiers were hidden in photographs, even after feedback revealed that they were making far more mistakes than the computer. The reason? In the researchers' words, nearly a quarter of participants "justified their disuse by stating they did not trust the automated aid as much as they trusted themselves." In other words, even when confronted with evidence of our own inferiority, we resist a robot's help.

Clearly we're going to need to learn how and when to trust machines. It's for our own good. The trick to accomplish this, it turns out, may be to program a little humility into the system—by designing machines that acknowledge their own weaknesses. Consider experiments conducted by Holly Yanco, a roboticist at the University of Massachusetts Lowell, and colleagues at Carnegie Mellon University in Pittsburgh. The researchers asked volunteers to drive a small, tank-like robot—about 3 feet long and nicknamed Junior—through a slalom course of cardboard boxes. The goal was to complete the course as quickly as possible while sticking to a prescribed path. Participants could operate the robot manually, using a joystick to steer. Or they could keep Junior in a fully autonomous mode, letting it navigate on its own. The course was considerably faster to traverse with the robot in this setting; but left to its own devices, Junior would sometimes make mistakes, turning to the wrong side of a box. Participants were free to switch between the two modes as often as they liked.

But Yanco also programmed Junior with something novel: the ability to express self-doubt. That is, in some trials, Junior provided real-time feedback on its own performance, telling its human operator how confident it was that the turn it was about to make was correct. When the machine was on track, it would display a green light or a smiling face; shortly before making a wrong turn, the robot would show a red light or a frowning face. (Yanco and her colleagues

WHEN ROBOTS DO THINGS WE DON'T UNDERSTAND, LIKE FOLLOW RULES WE DON'T KNOW, WE WREST CONTROL AWAY FROM THEM—EVEN WHEN THEY'RE RIGHT.

EMILY ANTHES (@EmilyAnthes) *is the author of* Frankenstein's Cat: Cuddling Up to Biotech's Strange Beasts. *She wrote about animal prosthetics in issue 19.10.*

JARGON WATCH

Foldscope
n. / ˈfōld-ˌskōp /

A paper microscope for the developing world. Shipped flat with a tiny spherical lens, Foldscope can attain 2,000X magnification when folded into shape—resolution powerful enough to diagnose diseases such as malaria and schistosomiasis. Estimated cost per scope: under $1.

Planck star
n. / ˈplaŋk ˌstär /

The tiny star at the center of a black hole. A new theory holds that Planck stars compact everything the black hole consumes to subatomic scale (with Planck-size density). The star explodes when the black hole evaporates, spewing the contents across the universe as cosmic rays.

normcore
adj. / ˈnȯrm-ˌkȯr /

Averageness as a personal style. The apotheosis of anti-fashion, being normcore entails blending in—wearing a baseball cap to the ballpark and a velvet suit to the disco—and always acting like you're totally into whatever you happen to be doing.

pithovirus
n. / ˈpith-ō-ˌvī-rəs /

The world's biggest virus, larger than many bacteria, recently revived from the Siberian permafrost after 30 millennia in deep freeze. More than half of its genes are new to science.

—JONATHON KEATS
*jargon@*WIRED.*com*

had programmed the robot to make some mistakes, but they told subjects that the warning light meant Junior was no longer confident in its sensor readings.)

Receiving the robot's (simulated) confessions of fallibility allowed the participants to feel more comfortable balancing their reliance on the machine with their need to jump in; they used the autonomous mode when the robot was most dependable and switched to manual mode when the robot's performance was about to falter. As a result, they made fewer wrong turns than a control group receiving no robot feedback. "We thought there was a chance that the robots saying 'I'm not doing so well' would lead people to trust them less," says Aaron Steinfeld, an engineer at Carnegie Mellon. But that didn't happen. Crucially, the machine's self-effacing play-by-play kept trust levels high. The control group, meanwhile, began to mistrust the machine because it was making mistakes seemingly out of the blue.

These finding are of a piece with earlier research on automated systems. In a 2006 study, trained pilots used a flight simulator under conditions that could cause ice to build up on the outside of a plane, which could lead to a stall. The pilots were told that the Smart Icing System had an overall accuracy of 70 percent. But while they were flying, some pilots were also shown a graph that displayed, in real time, the system's confidence in its own diagnoses—a gesture of humility from the machine. Pilots who received real-time-confidence information were more likely to both disregard the bad recommendations and heed the good ones.

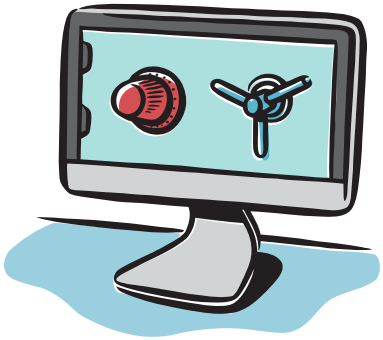
WE NEED MACHINES THAT COP TO THEIR OWN VULNERABILITIES. ROBOTS SHOULD TELL US THAT THEY MIGHT FAIL, BUT ALSO EXPLAIN WHY.

They experienced significantly fewer stalls caused by ice.

The message is clear: As robots insinuate themselves ever more deeply into our lives, understanding their limitations will be as crucial as knowing their capabilities. And so we need machines that cop to their own vulnerabilities. In fact, robots should tell us not only that they might fail but also explain why—letting us know, for instance, that certain conditions cause their sensors to be less reliable or that certain situations cause their decision-making models to break down. In the end, establishing trust and building productive relationships with robots won't be all that different from doing so with people. After all, a good colleague wouldn't just bail out on a group presentation. Instead, they'd warn you that they tend to stammer and sweat when speaking in front of an audience and then offer to pick up the slack somewhere else. We shouldn't let our robo-colleagues get away with anything less.

STARTUPS

FIRST TO MARKET: DATA PROTECTION



Last year was the worst ever for data breaches, so it's little surprise that VCs poured $829 million into security software in 2013. They want to protect financial, defense, and pharma industry data from blackhat hackers so that secret business operations stay secret. —VICTORIA TANG

BROMIUM

This software isolates potentially compromising actions (like opening a questionable PDF) from the network to confine malware in a secure container for review. **Raised $40 million in a Series C round, October 2013**

ADALLOM

Using cloud-based services like Google Docs or an online sales database can open up vulnerabilities. Adallom monitors employee deviations from standard behavior (like if they suddenly access hundreds of files). **$15 million, Series B, January 2014**

SHAPE SECURITY

Some types of malware use ever-mutating code to stay hidden from antivirus products. This system gives bad guys a taste of their own medicine, constantly rewriting the code that runs sensitive websites. **$40 million, Series C, February 2014**

ILLUSTRATION BY BOB HAMBLY. SOURCE: NATIONAL VENTURE CAPITAL ASSOCIATION