Can you quickly navigate this simple decision tree? The inputs are: ICML (International Conference on Machine Learning); 2017; Australia; kangaroo; and sunny. Assuming you have done it correctly, imagine trying to explain in words how your decision to clap hands was reached. What if there were a million variables?

then 8 by spring 2018, and now 18 in total, spanning subjects from systems programming to machine learning and its effects on fairness and privacy, to social networks and the question of censorship, to robots and work, and human-computer interaction.

Surveys of students in these classes show that between 80 percent and 90 percent approve of embedded ethics teaching, and want more of it. "My fantasy," says Grosz, "is that every computer-science course, with maybe one or two exceptions, would have an ethics module," so that by graduation, every concentrator would see that "ethics matters everywhere in the field—not just in AI." She and her colleagues want students to learn that in order to tackle problems such as bias and the need for human interpretability in AI, they must design systems with ethical principles in mind from the start.

Becoming a Boston Driver

Bemis professor of international law and professor of computer science Jonathan Zittrain, who is faculty director of the Berkman Klein Center for Internet and Society, has been grappling with this goal from a proto-legal perspective. In the spring of 2018, he co-taught a course with MIT Media Lab director Joi Ito exploring how AI technologies should be shaped to bear the public interest in mind. Autonomous vehicles provided a particularly salient case study that forced students to confront the nature of the complexities ahead, beyond the "runaway trolley problem" of deciding whom to harm and whom to save.

Once a car is truly autonomous, Zittrain explains, "It means that if an arrest warrant is issued for someone, the next time they enter an autonomous vehicle, the doors could lock and the car could just drive them to the nearest police station. Or what if someone in the car declares an emergency? Can the car propel them at 70 miles per

COURTESY OF FINALE DOSHI-VELEZ

AI and Adversarial Attacks

The privacy and security issues surrounding big data, the lifeblood of artificial intelligence, are well known: large streams and pools of data make fat targets for hackers. AI systems have an additional vulnerability: inputs can be manipulated in small ways that can completely change decisions. A credit score, for example, might rise significantly if one of the data points used to calculate it were altered only slightly. That's because computer systems classify each bit of input data in a binary manner, placing it on one side or the other of an imaginary line called a classifier. Perturb the input—say, altering the ratio of debt to total credit—ever so slightly, but just enough to cross that line, and that changes the score calculated by the AI system.

The stakes for making such systems resistant to manipulation are obviously high in many domains, but perhaps especially so in the field of medical imaging. Deep-learning algorithms have already been shown to outperform human doctors in correctly identifying skin cancers. But a recent study from Harvard Medical School coauthored by Nelson professor of biomedical informatics Isaac Kohane (see "Toward Precision Medicine," May-June 2015, page 17), together with Andrew Beam and Samuel Finlayson, showed that the addition of a small amount of carefully engineered noise "converts an image that the model correctly classifies as benign into an image that the network is 100 percent confident is malignant." This kind of manipulation, invisible to the human eye, could lead to nearly undetectable health-insurance fraud in the $3.3-trillion healthcare industry as a duped AI system orders unnecessary treatments. Designing an AI system ethically is not enough—it must also resist *unethical* human interventions.

Yaron Singer, an associate professor of computer science, studies AI systems' vulnerabilities to adversarial attacks in order to devise ways to make those systems more robust. One way is to use multiple classifiers. In other words, there is more than one way to draw the line that successfully classifies pixels in a photograph of a school bus as yellow or not yellow. Although the system may ultimately use only one of those classifiers to determine whether the image does contain a school bus, the attacker can't know which classifier the system is using at any particular moment—and that increases the odds that any attempt at deception will fail.

Singer points out that adding noise (random variations in brightness or color information) to an image is not in itself unethical—it is the uses, not the technology itself, that carry moral force. For example, noise can be used with online postings of personal photographs as a privacy-ensuring measure to defeat machine-driven facial recognition—a self-protective step likely to become more commonplace as consumer-level versions of noise-generating technologies become widely available. On the other hand, as Singer explains, were such identity-obfuscating software already widely available, Italian police would probably not have apprehended a most-wanted fugitive who'd been on the run since 1994. He was caught in 2017, perhaps when a facial recognition program spotted a photo of him at the beach, in sunglasses, on Facebook.