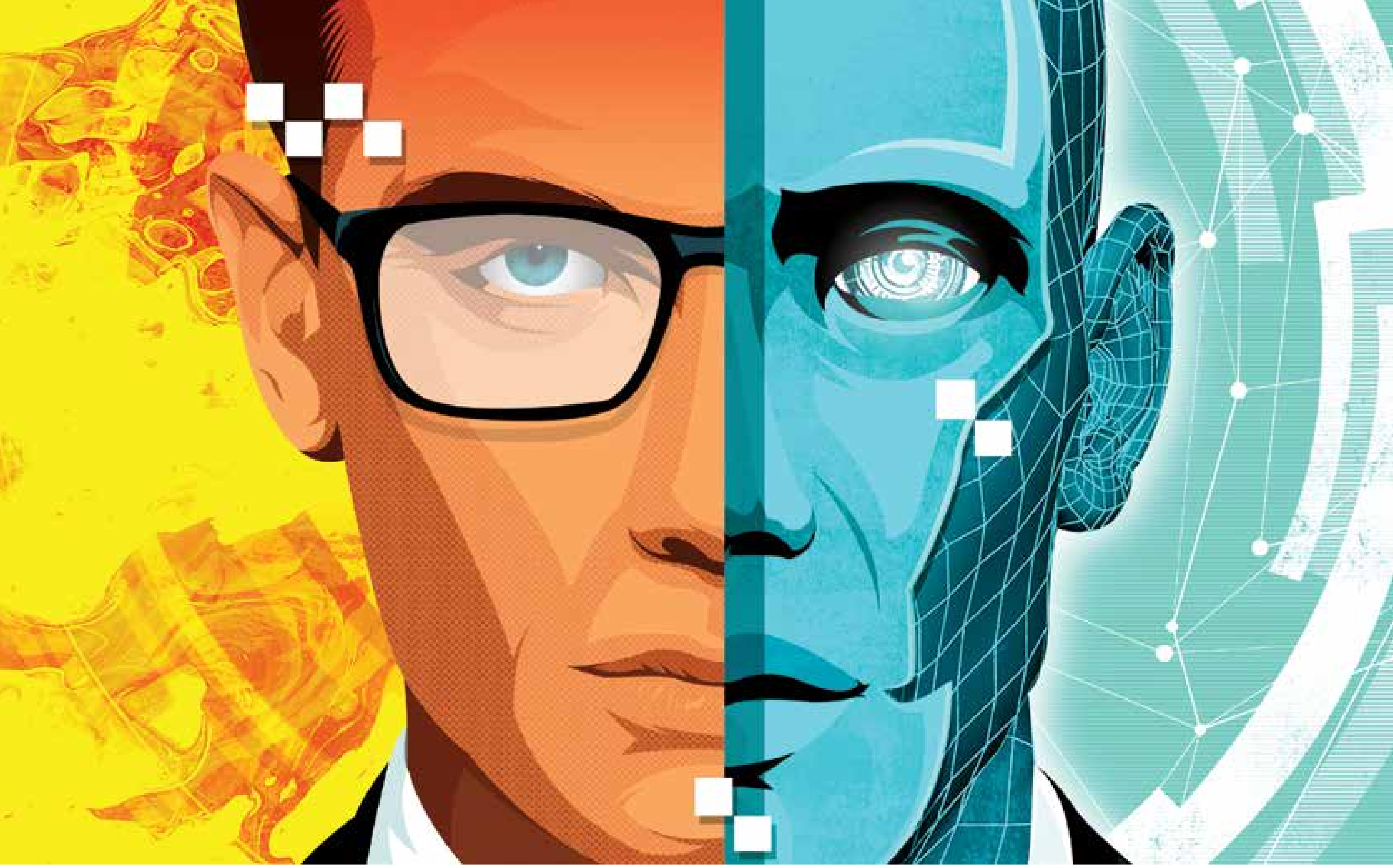


Artificial Intelligence & Ethics

Beyond engineering at the dawn of decision-making machines

by Jonathan Shaw

On March 18, 2018, at around 10 p.m., Elaine Herzberg was wheeling her bicycle across a street in Tempe, Arizona, when she was struck and killed by a self-driving car. Although there was a human operator behind the wheel, an autonomous system—artificial intelligence—was in full control. This incident, like others involving interactions between people and AI technologies, raises a host of ethical and proto-legal questions. What moral obligations did the system's programmers have to prevent their creation from taking a human life? And who was responsible for Herzberg's death? The person in the driver's seat? The company testing the car's capabilities? The designers of the AI system, or even the manufacturers of its onboard sensory equipment?

"Artificial intelligence" refers to systems that can be designed to take cues from their environment and, based on those inputs, proceed to solve problems, assess risks, make predictions, and take actions. In the era predating powerful computers and big data, such systems were programmed by humans and followed rules of human invention, but advances in technology have led to the development of new approaches. One of these is machine learning, now the most active area of AI, in which statistical methods allow a system to "learn" from data, and make decisions, without being explicitly programmed. Such systems pair an algorithm, or series of steps for solving a problem, with a knowledge base or stream—the information that the algorithm uses to construct a model of the world.

Ethical concerns about these advances focus at one extreme on the use of AI in deadly military drones, or on the risk that AI could take down global financial systems. Closer to home, AI has spurred anxiety about unemployment, as autonomous systems threaten to replace millions of truck drivers, and make Lyft and Uber obsolete. And beyond these larger social and economic considerations, data scientists have real concerns about bias, about ethical implementations of the technology, and about the nature of interactions between AI systems and humans if these systems are to be deployed

Illustrations by Taylor Callery

properly and fairly in even the most mundane applications.

Consider a prosaic-seeming social change: machines are already being given the power to make life-altering, everyday decisions about people. Artificial intelligence can aggregate and assess vast quantities of data that are sometimes beyond human capacity to analyze unaided, thereby enabling AI to make hiring recommendations, determine in seconds the creditworthiness of loan applicants, and predict the chances that criminals will re-offend.

But such applications raise troubling ethical issues because AI systems can reinforce what they have learned from real-world data, even amplifying familiar risks, such as racial or gender bias. Systems can also make errors of judgment when confronted with unfamiliar scenarios. And because many such systems are "black boxes," the reasons for their decisions are not easily accessed or understood by humans—and therefore difficult to question, or probe.

Examples abound. In 2014, Amazon developed a recruiting tool for identifying software engineers it might want to hire; the system swiftly began discriminating against women, and the company abandoned it in 2017. In 2016, ProPublica analyzed a commercially developed system that predicts the likelihood that criminals will re-offend, created to help judges make better sentencing decisions, and found that it was biased against blacks. During the past two years, self-driving cars that rely on rules and training data to operate have caused fatal accidents when confronted with unfamiliar sensory feedback or inputs their guidance systems couldn't interpret. The fact that private commercial developers generally refuse to make their code available for scrutiny, because the software is considered proprietary intellectual property, is another form of non-transparency—legal, rather than technical.

Barbara Grosz

Meanwhile, nothing about advances in the technology, per se, will solve the underlying, fundamental problem at the heart of AI, which is that even a thoughtfully designed algorithm must make decisions based on inputs from a flawed, imperfect, unpredictable, idiosyncratic real world.

Computer scientists have perceived sooner than others that engineering can't always address such problems post hoc, *after* a system has been designed. Despite notable advances in areas such as data privacy (see "The Privacy Tools Project," January-February 2017), and clear understanding of the limits of algorithmic fairness (see page 49), the realization that ethical concerns must in many cases be considered *before* a system is deployed has led to formal integration of an ethics curriculum—taught by philosophy postdoctoral fellows and graduate students—into many computer-science classes at Harvard. Far-reaching discussions about the social impact of AI on the world are taking place among data scientists across the University, as well as in the Ethics and Governance of AI Initiative launched by Harvard Law School's Berkman Klein Center, together with the MIT Media Lab. This intensifying focus on ethics originated with a longtime member of the computer science faculty.

From Communication to Cooperation—and Ethics

"A few years ago," says Higgins professor of natural sciences Barbara Grosz, "I was visiting friends at Microsoft—the husband develops computer-vision systems—and we drove somewhere to go walking. On the freeway in front of us was a truck, with a porta-potty on the back, and a bicycle attached to the porta-potty. 'What would my system do with this thing?' the husband wondered. 'Would it know how to react to that?'" The answer is, probably not. Such an image is unlikely to be part of its "experience"—the vast collection of images, laboriously tagged by humans, that form a system's training data.

The fragility of current AI systems stands in stark contrast to human intelligence, which is robust—capable of learning something in one context and swiftly applying it to another. Even if computers can distinguish bikes from trucks from porta-potties, they have difficulty recognizing how they might have been joined together to travel down the freeway, with the bicycle sideways, at 60 miles an hour. (Exploitation of this input vulnerability is the subject of "AI and Adversarial Attacks," on page 48.) In other words, AI lacks common sense and the ability to reason—even if it can also make incredible discoveries that no human could, such as detecting third- or higher-order interactions (when three or more variables must interact in order to have an effect) in complex biological networks. "Stop thinking about robots taking over," is how Grosz sums it up. "We have more to fear from dumb systems that people *think* are smart than from intelligent systems that know their limits."

Grosz, who studied mathematics at Cornell and then computer science at Berkeley, has worked on problems in AI since 1973, when she was hired as a research mathematician at the Artificial Intelligence Center of SRI International. She is considered an architect of the AI subfield devoted to how computers generate and interpret human speech and text—she won the Lifetime Achievement Award of the Association for Computational Linguistics in 2017—and can rattle off a litany of ways that language-capable systems such as Alexa, Siri, and Google fall short. They know where the nearest

Photograph by Eliza Grinnell/SEAS Communications

emergency room is, for example, but not that it might be useful to direct someone with a broken ankle to go there.

Grosz envisions a future that combines the speed and statistical prowess of intelligent computers with innate human talents....

Because her AI work in language predates data-driven approaches to natural language processing (see "Language as a Litmus Test for AI," page 47), Grosz developed a model-based approach to represent human discourse in a way that computers could understand. This has proved especially valuable to the field because it led her to reflect deeply on the nature of human-computer interaction, and later, in the course of imagining a future when computers and humans might work together, to propose theoretical models for collaborative AI systems designed to work on teams with people.

Her work on computational models of discourse goes far beyond the programming of grammatical rules. Understanding speaker intention, in order to determine the structure of a dialogue and thus to decipher meaning in human speech, was one key strategy she pioneered. Real speech, she points out, is full of digressions and shifts of focus, citing a notable example: her recording of the spontaneous dialogue as one person tries to tell another via teletype how to assemble an air compressor. (Well into the conversation, one speaker uses the pronoun "it" to refer to an object that has not been mentioned for half an hour—and both people understand exactly what is meant.) Intonation, she adds, is also key to understanding otherwise ambiguous phrases. "You're a real prince" might be said literally or sarcastically, in ways that a computer must be taught to understand.

From this interdisciplinary research flowed general principles about the nature of human-computer interaction. Grosz, with doctoral student Ece Kamar (now a senior researcher at Microsoft Research) developed a theory of "interruption management," for instance, for guiding information exchange between a human and a computer in order to make such communication exponentially more efficient. And she has come to believe, during the course of a long career, that the best of use of AI involves integrating such systems with human teams. She envisions a future that combines the speed and statistical prowess of intelligent computers with innate human talents, not one that pits machines and humans against each other—the way the relationship is often framed in descriptions of AI systems beating world champions in chess and go, or replacing people in the workplace. Such an integrated approach arguably represents the frontier in AI systems.

When Grosz began experimenting with team-based AI systems in health care, she and a Stanford pediatrician started a project that coordinates care for children with rare diseases who are tended by many people besides parents, including medical experts, home-care aides, physical therapists, and classroom teachers. The care spans years, she says, and "no human being *I've* ever encountered can keep track of 15 other people and what they are doing over long periods of time."

Grosz, with doctoral student Ofra Amir (now a faculty member at the Technion) began by analyzing how the patient-care teams worked, and developed a theory of teamwork to guide interactions between the human members and an AI system designed to coordinate information about the children's care. As she had done with language, she started with general principles. "What we're trying to do, on the theoretical end, is to understand better how to share information" in that multi-member team environment, "and then build tools, first for parents, and then for physicians."

One of the key tenets she and her colleague, Bar-Ilan University professor Sarit Kraus, developed is that team members should not take on tasks they lack the requisite knowledge or capability to accomplish. This is a feature of good human teamwork, as well as a key characteristic of "intelligent systems that know their limits." "The problem, not just with AI, but a lot of technology that is out in the world, is that it can't do the job it has been assigned"—online customer service chatbots interacting via text that "are unable to understand what you want" being a case in point. Those systems could have been designed differently, she says, so that the first interactions are with a person *aided by* a computer; the person would be building a relationship with the customer, while vetting what the computer was clearly misunderstanding, and the system, meanwhile, would enable the person to provide an answer

Language as a Litmus Test

Language, which clearly played an important role in human evolution, has long been considered a hallmark of human intelligence, and when Barbara Grosz started working on problems in artificial intelligence (AI) in the 1970s, it was the litmus test for defining machine intelligence. The idea that language could be used as a kind of Occam's razor for identifying intelligent computers dates to 1950, when Alan Turing, the British scientist who cracked Nazi Germany's encrypted military communications, suggested that the ability to carry on a conversation in a manner indistinguishable from a human could be used as a proxy for intelligence. Turing raised the idea as a philosophical question, because intelligence is difficult to define, but his proposal was soon memorialized as the Turing test. Whether it is a reasonable test of intelligence is debatable. Regardless, Grosz says that even the most advanced, language-capable AI systems now available—Siri, Alexa, and Google—fail to pass it.

The Higgins professor of natural sciences has witnessed a transformation of her field. For decades, computers lacked the power, speed, and storage capacity to drive neural networks—modeled on the wiring of the human brain—that are able to *learn* from processing vast quantities of data. Grosz's early language work therefore involved developing formal models and algorithms to create a computational model of discourse: telling the computer, in effect, how to interpret and create speech and text. Her research has led to the development of frameworks for handling the unpredictable nature of human communication, for modeling one-on-one human-computer interactions, and for advancing the integration of AI systems into human teams.

The current ascendant AI approach—based on neural networks that learn—relies instead on computers' ability to sample vast quantities of data. In the case of language, for example, a neural network can sample a corpus—extending even to everything ever written that's been posted online—to learn the "meaning" of words and their relationship to each other. A dictionary created using this approach, explains assistant professor of computer science Alexander "Sasha" Rush, contains mathematical representations of words, rather than language-based definitions. Each word is a vector—a relativistic definition of a word in relation to other words. Thus the vectors describing the relationship between the words "man" and "woman" would be mathematically analogous to those describing the relationship between words such as "king" and "queen."

This approach to teaching language to computers has tremendous potential for translation services, for developing miniaturized chips that would allow voice control of all sorts of devices, and even for creating AIs that could write a story about a sporting event based purely on data. But because it captures all the human biases associated with culturally freighted words like "man" and "woman," and what the ensuing mathematical representations might embody with respect to gender, power dynamics, and inequality when confronted with the associations of a word such as "CEO," it can lead neural-network based AI systems to produce biased results.

Rush considers his work—developing language capabilities for microscopic computer chips—to be purely engineering, and his translation work to be functional, not literary, even though the goal of developing an AI that can pass the Turing test is undoubtedly being advanced by work like his. But significant obstacles remain.

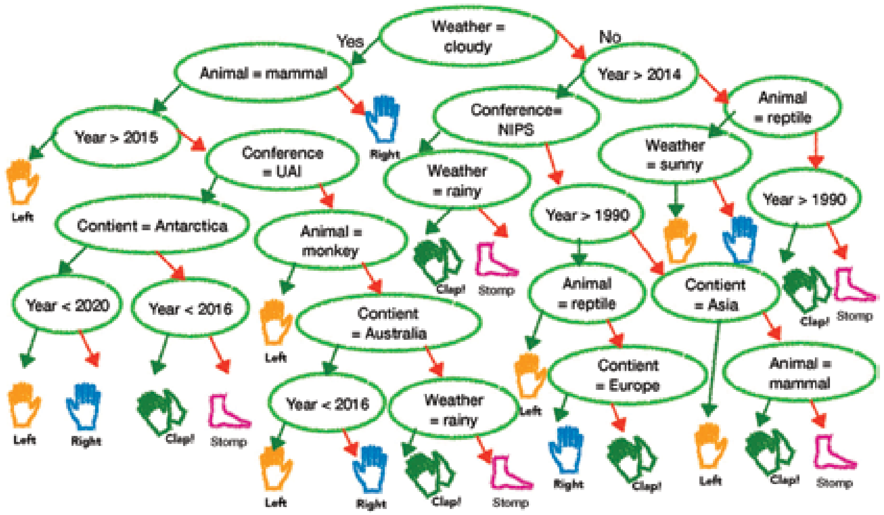
How can a computer be taught to recognize inflection, or the rising tone of words that form a question, or an interruption to discipline kids ("Hey, stop that!"), of the sort that humans understand immediately? These are the kinds of theoretical problems Grosz has been grappling with for years. And although she is agnostic about whatever approach will ultimately succeed in building systems able to participate in everyday human dialogue, probably decades hence, she does allow that it might well have to be a hybrid of neural-network learning and human-developed models and rules for understanding language in all its complexity.

more quickly. When such fundamentals of intelligent-systems design aren't respected, the systems are assumed to be capable of things they can't do, or are used in naïve, inappropriate ways.

Grosz's highly interdisciplinary approach to *research*, informed by linguistics, philosophy, psychology, economics, and even a bit of anthropology and sociology, led her to think also about which of these subjects might best inform the *teaching* of AI systems design. Though she had taught an introductory course on AI from 1987 to 2001, a time when its application remained largely theoretical, the world had changed by the time she rebooted that course in 2013 and 2014, when fully operational AI systems were being deployed. Grosz realized there was a teaching opportunity in the interplay between the ethical challenges presented by AI and good systems design.

This led to one of Grosz's most important contributions to the teaching of computer science at Harvard: the idea that ethics should be tightly integrated into every course. In the fall of 2015, she introduced a new course, "Intelligent Systems Design and Ethical Challenges." By the following year, more than 140 students had applied for the 25 spots in the class, emboldening her to encourage her computer-science colleagues to incorporate some teaching of ethics into their own courses. Because most of them lacked sufficient background to be comfortable teaching ethics, she began a collaboration with Wolcott professor of philosophy Alison Simmons, who chairs the philosophy department. Together, they worked with colleagues in their respective fields, enlisting CS professors willing to include ethics modules in their computer-science courses and philosophy graduate students to teach them.

The aim of this "Embedded EthiCS" initiative, she says, is to instruct the people who will build future AI systems in how to identify and think through ethical questions. (Computer science is now the second largest concentration among Harvard undergraduates; if students from related fields such as statistics or applied mathematics are included, the total enrollment substantially exceeds that of top-ranked economics.) "Most of these ethical challenges have no single right answer," she points out, "so just as [the students] learn fundamental computing skills, I wanted them to learn fundamental ethical-reasoning skills." In the spring of 2017, four computer-science courses included some study of ethics. That fall, there were five,



COURTESY OF FINALE DOSHI-VELEZ

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

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.

hour on city streets to the hospital, while all the other cars part like the Red Sea?"

Students in Zittrain's class thought they knew how the discussion about autonomous vehicles would unfold. But when he posed a very simple question—"Should the driver be able to instruct the car to go 80 miles per hour?"—they were confronted with a designer's moral dilemmas. If yes, and the car were involved in an accident at that speed, would the driver be responsible? Or would the carmaker be liable for *allowing* the car to speed? "People speed all the time, but we have the implicit comfort of knowing that there is roughly nothing we can do about it," Zittrain notes. "The understandable initial premise [with autonomous vehicles] is that, gosh, there's no driver, and we can't blame an inanimate object like a car. It looks as though there is a paucity of responsibility"—whereas in fact, "there's a surfeit of responsibility." The manufacturers, the AI designers, the policymakers, *and* the driver could all be held accountable.

And the situation becomes more complex if the vehicle's AI system dynamically changes its behavior as it "learns" from experiences on the road, Zittrain points out. "Maybe if it drives enough in Boston, it will become a Boston driver!" This applies to many learning systems, and the legal solutions remain unexplored. Maybe, he suggests, if an AI designer or other contributor builds a learning system in which behavior can't always be predicted, there will be a price tag on operating with that uncertainty.

The subject is a nexus of interdisciplinary inquiry, Zittrain continues. At the Berkman Klein Center and MIT's Media Lab, he and his colleagues have created a group called "Assembly" that brings software developers from outside companies in on sabbatical to work with students and one another for a couple of months on some of these puzzles in AI and other data-science fields. "The embedded ethics instruction is part of an effort to create opportunities for students from across the University to encounter one another, and bring the tools they are learning in their respective schools to bear on this kind of stuff in teams.

"I think that's part of what's made Barbara [Grosz]'s teaching and research so influential here. And so timeless. Her teaching is not how to intervene in a computer system or piece of software to fix it. It's really thinking at a broader level about how people and technologies should be interacting." Can they be accountable? Can they be understood? Can they be fair?

Systemic Bias and Social Engineering

The problem of fairness in autonomous systems featured prominently at the inaugural Harvard Data Science Conference (HDSC) in October, where Colony professor of computer science David Parkes outlined guiding principles for the study of data science at Harvard: it should address ethical issues, including privacy (see "The Watchers," January-February 2017, page 56); it should not perpetuate existing biases; and it should be transparent. But to create learning AI systems that embody these principles can be hard. System complexity, when thousands or more variables are in play, can make true understanding almost impossible, and biases in the datasets on which learning systems rely can easily become reinforced.

There are lots of reasons why someone might want to "look under the hood" of an AI system to figure out how it made a particular decision: to assess the cause of biased output, to run safety checks before rollout in a hospital, or to determine accountability after an accident involving a self-driving car.

What might not be obvious is how difficult and complex such an inquiry can be. Assistant professor of computer science Finale Doshi-Velez demonstrated by projecting onscreen a relatively simple decision tree, four layers deep, that involved answering questions based on five inputs (see a slightly more complex example, opposite). If executed correctly, the final instruction was to raise your left hand. A few of the conference attendees were able to follow along. Then she showed a much more complex decision tree, perhaps 25 layers deep, with five new parameters determining the path down through the tree to the correct answer—an easy task for a computer. But when she asked if anyone in the audience could describe in words *why* they had reached the answer they did, no one responded. Even when the correct path to a decision is highlighted, describing the influence of complex interacting inputs on the outcome in layman's terms is extremely difficult. And that's just for simple models such as decision trees, not modern deep architectures with millions of parameters. Developing techniques to extract explanations from arbitrary models—scalable systems with an abitrary number of variables, task, and outputs—is the subject of research in her lab.

Bias poses a different set of problems. Whenever there is a diverse population (differing by ethnicity, religion, or race, for example), explained McKay professor of computer science Cynthia Dwork during a HDSC talk about algorithmic fairness,

(please turn to page 74)

ARTIFICIAL INTELLIGENCE AND ETHICS

(continued from page 49)

an algorithm that determines eligibility for, say, a loan, should treat each group the same way. But in machine-learning systems, the algorithm itself (the step-by-step procedure for solving a particular problem) constitutes only one part of the system. The other part is the data. In an AI system that makes automated loan decisions, the *algorithm* component can be unbiased and completely fair with respect to each group, and yet the overall result, after the algorithm has learned from the *data*, may not be. "Algorithms don't have access to the ground truth" (computer lingo for *veritas*), Dwork explained. If there is bias in the data used to make the decision, the decision itself may be biased.

There are ways to manage this problem. One is to select very carefully the applicant attributes an algorithm is permitted to consider. (Zip codes, as well-known proxies for race, are often eliminated.) But bias has a way of creeping back in through correlations with other variables that the algorithm uses—such as surnames combined with geographic census data.

"Algorithms, which are purely optimization-driven tools, can inherit, internalize, reproduce, and exacerbate existing inequalities."

Bias against *groups* can often be addressed through smart algorithm design, Dwork said, but ensuring fairness to *individuals* is much harder because of a fundamental feature of algorithmic decisionmaking. Any such decision effectively draws a line—and as Dwork pointed out, there will always be two individuals from different groups close to the line, one on either side, who are very similar to each other in almost every way. And yet only one will get a loan.

In some cases, correcting bias through system design may be an insufficient approach. Consider a hiring system designed by McKay professor of computer science Yiling Chen and graduate student Lily Hu '15 to eliminate hiring bias against African Americans, historically a disadvantaged group. As Hu puts it, "Algorithms, which are purely optimization-driven tools, can inherit, internalize, reproduce, and exacerbate existing inequalities. Say we have a labor-market disparity that persists without any sort of machine-learning help, and then here comes machine learning, and it learns to re-inscribe those inequalities." Their solution, which uses tools from economics and sociology to understand disparities in the labor market, pushes the thinking about algorithmic fairness beyond computer science to an interdisciplinary, systems-wide view of the problem.

Chen works in social computing, an area of data science that emphasizes the effect of human behavior on inputs to algorithms. Because humans are "self-interested, independent, error-prone, and not predictable" enough to enable design of an algorithm that would ensure fairness in every situation, she started thinking about how to take bias out of the training data—the real-world information inputs that a hiring algorithm would use.

She and Hu focused on the problem of implementing affirmative action in hiring. A straightforward remedy to counteract the historical disadvantage faced by a minority group would be, simply, to favor that group in employment decisions, all other things being equal. (This might itself be deemed unfair to the majority group, but still be considered acceptable until equity in hiring is attained.) But Chen and Hu then considered the human element. Suppose many of the minority group's members don't go to college, reasoning that "it's expensive, and because of discrimination, even if I get a degree, the chances of my getting a job are still low." Employers, meanwhile, may believe that "people from minority groups are less educated, and don't perform well, because they don't try hard." The point Chen and Hu make is that even though a minority-group member's decision not to attend college is rational, based on existing historical unfairness, that decision reinforces employers' preconceived ideas about the group as a whole. This pattern of feedback effects is not just difficult to break—it is precisely the sort of data pattern that an algorithm, looking at past successful hires and associating them with college degrees, will reinforce.

The solution that Chen and Hu propose is not based on math alone: instead, it is social engineering that uses an algorithm to change the ground truth. And it represents an acknowledgment of just how difficult it can be to counteract bias in data. What the researchers propose is the creation of a temporary labor market. Think of it, says Chen, as an internship in which every job candidate must participate for two years before being hired into the permanent workforce. Entry into the internship pool would be subject to a simple "fairness constraint," an algorithm that would *require* employers to choose interns from minority and majority groups in representative numbers. Then, at the conclusion of the internship, hiring from the pool of interns would be based *solely* on performance data, without regard to group membership. Because the groups are equally talented at the population level, explains Chen, the two groups eventually reach parity.

"What this paper's trying to fight back against," Hu explains, "is the perception—still dominant within the machine-learning/AI community—that everything is fundamentally an *optimization* problem, or a *prediction* problem, or a *classification* problem. And when you do that—if you treat it in a standard machine-learning way—you will end up reinforcing those inequalities."

Hu was the teaching fellow for Grosz's course in AI and ethics (co-taught with philosophy fellow Jeffrey Behrends) last year. She says people need to understand that the act of "building technologies, and the way that we implement them, are themselves political actions. They don't exist in a vacuum, as instrumental tools that sometimes are good, sometimes are bad. I think that that's a particularly naïve way of thinking of technology."

Whether the technology is meant to provide facial recognition to identify crime suspects from video footage, or education tailored to different learning styles, or medical advice, Hu stresses, "What we need to think about is how technologies embed particular values and assumptions. Exposing that is a first step: realizing that it's not the case that there are some ethical questions, and some non-ethical questions, but really that, in *everything* we design…there are always going to be normative questions at hand, every step of the way." Integrating that awareness into existing coursework is critical to ensuring that "the world that we're building, with ubiquitous technology, is a world that we want to live in."

Jonathan Shaw '89 is managing editor of this magazine.