A Study on Driverless-Car Ethics Offers a Troubling Look Into Our Values

The first time Azim Shariff met Iyad Rahwan—the first real time, after communicating with him by phone and e-mail—was in a driverless car. It was November, 2012, and Rahwan, a thirty-four-year-old professor of computing and information science, was researching artificial intelligence at the Masdar Institute of Science and Technology, a university in Abu Dhabi. He was eager to explore how concepts within psychology—including social networks and collective reasoning—might inform machine learning, but there were few psychologists working in the U.A.E. Shariff, a thirty-one-year-old with wild hair and expressive eyebrows, was teaching psychology at New York University's campus in Abu Dhabi; he guesses that he was one of four research psychologists in the region at the time, an estimate that Rahwan told me "doesn't sound like an exaggeration." Rahwan cold-e-mailed Shariff and invited him to visit his research group.

The lab was situated in Masdar City, an experimental planned community in the heart of Abu Dhabi. The city runs entirely on renewable energy and prohibits the use of gas-powered vehicles. Instead, residents travel by "personal rapid transit"—a system of small, driverless cars that snake around the streets on magnetized paths. Rahwan waited for Shariff in a parking lot near the city limits, where commuters transfer from gas-powered cars to the self-driving pods. The cars function more like trains than like true autonomous vehicles, or A.V.s; they don't deviate from set paths and make almost no decisions on their own. But, in 2012, when A.V.s were almost entirely theoretical, whirring around in a car with no steering wheel and no brakes felt electrifying for Shariff. As he travelled through the city with Rahwan, he held his phone out in front of him, filming the entire ride.

Today, cars with semi-autonomous features are already on the road. Automatic parallel parking has been commercially available since 2003. Cadillac allows drivers to go hands-free on pre-approved routes. Some B.M.W. S.U.V.s can be equipped with, for an additional seventeen-hundred dollars, a system that takes over during "monotonous traffic situations"—more colloquially known as traffic jams. But a mass-produced driverless car remains elusive. In 2013, the U.S. Department of Transportation's National Highway Traffic Safety Administration published a sliding scale that ranked cars on their level of autonomy. The vast majority of vehicles are still at level zero. A car at level four would be highly autonomous in basic situations, like highways, but would need a human operator. Cars at level five would drive as well as or better than humans, smoothly adapting to rapid changes in their environments, like swerving cars or stray pedestrians. This would require the vehicles to make value judgments, including in versions of a classic philosophy thought experiment called the trolley problem: if a car detects a sudden obstacle—say, a jackknifed truck—should it hit the truck and kill its own driver, or should it swerve onto a crowded sidewalk and kill pedestrians? A human driver might react randomly (if she has time to react at all), but the response of an autonomous vehicle would have to be programmed ahead of time. What should we tell the car to do?

Shariff moved to the U.S. at the end of 2012, and then to Canada. Rahwan eventually got a job at the M.I.T. Media Lab, in Cambridge, but the pair kept in touch. They reached out to a third collaborator with whom they previously worked, a French professor of cognitive science named Jean-François Bonnefon, and the scientists had periodic phone calls. "The three of us just jelled really well together," Shariff said. "We would kind of geek out." One of their most frequent topics of conversation was the ethics of self-driving cars. In 2015, Rahwan invited Shariff and Bonnefon to visit him in Boston. In the course of a week, they met regularly, either in the Media Lab, which, with its grid of steel and glass, looks like it's made of graph paper, or in one of the cafés on Kendall Square. On a drizzly day in November, they came up with an

idea. Before determining the ethical decisions that A.V.s should make, they had to understand what decisions human drivers would make if they had the time to react. By crowdsourcing this question, they could figure out what people's values were. Rahwan still has his notes from that day, which include scribbled phrases like "# of people on the road," "# of people in the car," and "odds of dying."

In June of 2016, the Media Lab launched a Web site that invited people from all over the world to play a game called Moral Machine. In the game, players are presented with a version of the trolley problem: a driverless car can either stay its course and hit what is in its path, or swerve and hit something else. Each round features a new version of the problem, with different obstacles and different groups of people to be killed or spared. In the next two years, more than two million people—from some two hundred countries and territories—participated in the study, logging more than forty million decisions. It is the largest study on moral preferences for machine intelligence ever conducted.

The paper on the project was published in *Nature*, in October, 2018, and the results offer an unlikely window into people's values around the globe. On the whole, players showed little preference between action and inaction, which the scientists found surprising. "From the philosophical . . . and legal perspective . . . this question is very important," Shariff explained. But the players showed strong preferences for what kinds of people they hit. Those preferences were determined, in part, by where the players were from. Edmond Awad, a research fellow, and Sohan Dsouza, a graduate student working with Rahwan, noticed that the responses could be grouped into three large geographic "clusters": the Western cluster, including North America and Western Europe; the Eastern cluster, which was a mix of East Asian and Islamic countries; and the Southern cluster, which was composed of Latin-American countries and a smattering of Francophone countries.

We should be wary of drawing broad conclusions from the geographical differences, particularly because about seventy per cent of the

respondents were male college graduates. Still, the cultural differences were stark. Players in Eastern-cluster countries were more likely than those in the Western and Southern countries to kill a young person and spare an old person (represented, in the game, by a stooped figure holding a cane). Players in Southern countries were more likely to kill a fat person (a figure with a large stomach) and spare an athletic person (a figure that appeared mid-jog, wearing shorts and a sweatband). Players in countries with high economic inequality (for example, in Venezuela and Colombia) were more likely to spare a business executive (a figure walking briskly, holding a briefcase) than a homeless person (a hunched figure with a hat, a beard, and patches on his clothes). In countries where the rule of law is particularly strong—like Japan or Germany—people were more likely to kill jaywalkers than lawful pedestrians. But, even with these differences, universal patterns revealed themselves. Most players sacrificed individuals to save larger groups. Most players spared women over men. Dog-lovers will be happy to learn that dogs were more likely to be spared than cats. Human-lovers will be disturbed to learn that dogs were more likely to be spared than criminals.

In its discussion, the paper skims over the uglier aspects of the study to identify "three strong preferences" that might provide a starting point for developing a standardized machine-ethics framework: sparing human lives, sparing more lives, and sparing young lives. The paper concludes with a soaring look into the future, and recasts machine ethics as a "unique opportunity to decide, as a community, what we believe to be right or wrong; and to make sure that machines, unlike humans, unerringly follow these moral preferences." But, when I asked Shariff what he thought of the human prejudice shown in the data, he laughed and said, "That suggests to us that we shouldn't leave decisions completely in the hands of the demos."

The U.S. government has clear guidelines for autonomous weapons—they can't be programmed to make "kill decisions" on their own—but no formal opinion on the ethics of driverless cars. Germany is the only country that has devised such a framework; in 2017, a German

government commission—headed by Udo Di Fabio, a former judge on the country's highest constitutional court—released a report that suggested a number of guidelines for driverless vehicles. Among the report's twenty propositions, one stands out: "In the event of unavoidable accident situations, any distinction based on personal features (age, gender, physical or mental constitution) is strictly prohibited." When I sent Di Fabio the Moral Machine data, he was unsurprised by the respondent's prejudices. Philosophers and lawyers, he noted, often have very different understandings of ethical dilemmas than ordinary people do. This difference may irritate the specialists, he said, but "it should always make them think." Still, Di Fabio believes that we shouldn't capitulate to human biases when it comes to life-and-death decisions. "In Germany, people are very sensitive to such discussions," he told me, by e-mail. "This has to do with a dark past that has divided people up and sorted them out."

The decisions made by Germany will reverberate beyond its borders. Volkswagen sells more automobiles than any other company in the world. But that manufacturing power comes with a complicated moral responsibility. What should a company do if another country wants its vehicles to reflect different moral calculations? Should a Western car de-prioritize the young in an Eastern country? Shariff leans toward adjusting each model for the country where it's meant to operate. Car manufacturers, he thinks, "should be sensitive to the cultural differences in the places they're instituting these ethical decisions." Otherwise, the algorithms they export might start looking like a form of moral colonialism. But Di Fabio worries about letting autocratic governments tinker with the code. He imagines a future in which China wants the cars to favor people who rank higher in its new social-credit system, which scores citizens based on their civic behavior.

Both Di Fabio and Shariff agree that the advent of autonomous vehicles will force us to make our underlying moral calculations explicit. In twenty to fifty years, the majority of cars on the road will likely be driverless. If billions of machines are all programmed to make the same

judgement call, it may be a lot more dangerous to cross the street as, say, an overweight man than as a fit woman. And, if companies decide to tweak the software to prioritize their customers over pedestrians, it may be more dangerous to be beside the road than on it. In a future dominated by driverless cars, moral texture will erode away in favor of a rigid ethical framework. Let’s hope we’re on the right side of the algorithm.