NOVEMBER 24, 2012

MORAL MACHINES

BY GARY MARCUS

Google's driver-less cars are already street-legal in three states, California, Florida, and Nevada, and some day similar devices may not just be possible but mandatory. Eventually (though not yet) automated vehicles will be able to drive better, and more safely than you can; no drinking, no distraction, better reflexes, and better awareness (via networking) of other vehicles. Within two or three decades the difference between automated driving and human driving will be so great you may not be legally allowed to drive your own car, and even if you are allowed, it would be immoral of you to drive, because the risk of you hurting yourself or another person will be far greater than if you allowed a machine to do the work.

That moment will be significant not just because it will signal the end of one more human niche, but because it will signal the beginning of another: the era in which it will no longer be optional for machines to have ethical systems. Your car is speeding along a bridge at fifty miles per hour when errant school bus carrying forty innocent children crosses its path. Should your car swerve, possibly risking the life of its owner (you), in order to save the children, or keep going, putting all forty kids at risk? If the decision must be made in milliseconds, the computer will have to make the call.

These issues may be even more pressing when it comes to military robots. When, if ever, might it be ethical to send robots in the place of soldiers? Robot soldiers might not only be faster, stronger, and more reliable than human beings, they would also be immune from panic and sleep-deprivation, and never be overcome with a desire for vengeance. Yet, as The Human Rights Watch noted in a widely-publicized report earlier this week (http://www.hrw.org/print/reports/2012/11/19/losing-humanity), robot soldiers would also be utterly devoid of human compassion, and could easily wreak unprecedented devastation in the hands of a Stalin or Pol Pot. Anyone who has seen the opening scenes of RoboCop knows why we have misgivings about robots being soldiers, or cops.

But what should we do about it? The solution proposed by Human Rights Watch—an outright ban on "the development, production, and use of fully autonomous weapons"—seems wildly unrealistic. The Pentagon is likely to be loath to give up its enormous investment in robotic soldiers (in the words of Peter W. Singer, "Predator [drones] are merely the first generation."), and few parents would prefer to send their own sons (or daughters) into combat if robots were an alternative.

****

With or without robotic soldiers, what we really need is a sound way to teach our machines to be ethical. The trouble is that we have almost no idea how to do that. Many discussions start with three famous laws from Isaac Asimov:

1. A robot may not injure a human being or, through inaction, allow a human being to come to harm.

2. A robot must obey the orders given to it by human beings, except where such orders would conflict with the first law.

3. A robot must protect its own existence as long as such protection does not conflict with the first or second laws.

The trouble with these seemingly sound laws is threefold. The first is technical: at least for now, we couldn't program a machine with Asimov's laws if we tried. As yet, we haven't figured out how to build a machine that fully comprehends the concept of "dinner", much less something as abstract as "harm" or "protection." Likewise, we are a long way from constructing a robot that can fully anticipate the consequences of any of its actions (or inactions). For now, a robot is lucky if it can predict would happen if it dropped a glass of water. A.I. has a long way to go before laws as abstract as Asimov's could realistically be encoded in software.

Second, even if we could figure out how to do the programming, the rules might be too restrictive. The first and second laws, for example, preclude robots from ever harming other humans, but most people would make exceptions for robots that could eliminate potential human targets that were a clear and present danger to others. Only a true ideologue would want to stop a robotic sniper from taking down a hostage-taker or Columbine killer.

Meanwhile, Asimov's laws themselves might not be fair—to robots. As the computer scientist Kevin Korb has pointed out, Asimov's laws effectively treat robots like slaves. Perhaps that is acceptable for now, but it could become morally questionable (and more difficult to enforce) as machines become smarter and possibly more self-aware.

The laws of Asimov are hardly the only approach to machine ethics, but many others are equally fraught. An all-powerful computer that was programmed to maximize human pleasure, for example, might consign us all to an intravenous dopamine drip; an automated car that aimed to minimize harm would never leave the driveway. Almost any easy solution that one might imagine leads to some variation or another on the Sorceror's Apprentice (http://singularity.org/files/SaME.pdf), a genie that's given us what we've asked for, rather than what we truly desire. A tiny cadre of brave-hearted souls at Oxford (http://www.fhi.ox.ac.uk/), Yale (http://www.yale.edu/bioethics/bioethicsscholars.shtml), and the Berkeley California Singularity Institute (http://singularity.org/research/) are working on these problems, but the annual amount of money being spent on developing machine morality is tiny.

****

The thought that haunts me the most is that that human ethics themselves are only a work-in-progress. We still confront situations for which we don't have well-developed codes (e.g., in the case of assisted suicide) and need not look far into the past to find cases where our own codes were dubious, or worse (e.g., laws that permitted slavery and segregation). What we really want are machines that can go a step further, endowed not only with the soundest codes of ethics that our best contemporary philosophers can devise, but also with the possibility of machines making their own moral progress, bringing them past our own limited early-twenty-first century idea of morality.

Building machines with a conscience is a big job, and one that will require the coordinated efforts of philosophers, computer scientists, legislators, and lawyers. And, as Colin Allen (http://www.indiana.edu/~hpscdept/people/allen.shtml), a pioneer in machine ethics put it, "We don't want to get to the point where we should have had this discussion twenty years ago." As machines become faster, more intelligent, and more powerful, the need to endow them with a sense of morality becomes more and more urgent. (http://singularity.org/files/SaME.pdf)

“Ethical subroutines” may sound like science fiction, but once upon a time, so did self-driving cars.

Gary Marcus, Professor of Psychology at N.Y.U., is author of “Guitar Zero: The Science of Becoming Musical at Any Age (http://www.amazon.com/Guitar-Zero-Science-Becoming-Musical/dp/0143122789/)” and “Kluge: The Haphazard Evolution of The Human Mind (http://www.amazon.com/Kluge-Haphazard-Evolution-Human-Mind/dp/B002ECETZY).”

Photograph by Justin Sullivan/Getty.

Gary Marcus is a professor of cognitive science at N.Y.U. and the author of “Guitar Zero.”