Todd Pittinsky

Algorithms and Ethical Diversity

Developing a More Holistic View of Technology and Society

Would you swerve off the road to avoid hitting a deer? Some of us would, some of us wouldn't. But these days, there's yet another angle: what would a self-driving car do if a deer jumped in front of it? Will different manufacturers' cars make different choices? Will your car do what *you* would do?

We each have our own "ethical algorithms" — who and what is worth how much risk or sacrifice to us. What should worry us about autonomous cars is not that they won't have an ethical algorithm — they certainly will — but that they will all have the same one, shaped only by a combination of government regulation and market research. And the safer we all become because of autonomous cars — which we certainly will — the greater cachet such ethical algorithms will have. There may not be any tragic loss of human variety and individualism in adopting — even inadvertently — a uniform ethical algorithm for driving. But there is plenty to be lost in the

Digital Object Identifier 10.1109/MTS.2020.3012317
Date of current version: 2 September 2020

ISTOCK/METAMORWORKS

many other realms, such as healthcare and education, that will inevitably be swept into the empire of AI and algorithms. And the cumulative effect, of ethical algorithms dictating most aspects of our lives, should be worrisome.

Clearly, autonomous cars should be programmed to obey traffic laws. They should steer *into* a skid on ice, something human beings too often do not do, deliberatively or instinctively. Lives and limbs will be saved. But let's look further. You would probably speed up — breaking the law — to avoid a collision or to rush someone to the hospital. But will your autonomous car do that? You might slow down to let someone pull ahead of you just because you're in a good mood and in no hurry. But would your autonomous car ever do that? People express their individuality and personalities when they drive. Sometimes they're expressing what jerks they are and sometimes how good-natured or compassionate they are. Are we sure we want to eliminate the latter in order to eliminate the former?

That is perhaps the most underappreciated challenge implicit in the algorithm-driven automation — or impending automation — of so many aspects of our society. Not whether we can come up with an ethical approach, but whether we want to live in a world with "an" approach rather than the diversity of ethical approaches, including some lousy ones, that come from being human. M.I.T.'s "Moral Machine" experiments illustrate the challenge beautifully. They have created a platform for "gathering a human perspective on moral decisions made by machine intelligence, such as self-driving cars" (1). The result? The platform shows tremendous diversity across individuals in such decision-making.

Who Is Really in the Driver's Seat?

Ethical diversity refers to the "diverse beliefs … as to what are the most ethically appropriate or inappropriate courses of actions" and takes into account the different values and beliefs people hold (2). This diversity is and has always been a source of confusion and conflict, from the personal to the international. The answer, however, is to have forums to debate and discuss the ethical choices embedded in everyday life, not algorithms that render the choice being made invisible. Not to mention leaving that invisible choice to for-profit corporations.

We need to recognize who is in the driver's seat, making the decisions about what new technology comes our way and what ethical decision-making (possibly but not necessarily a formal algorithm) is embedded in it. Self-driving cars are still largely in the future, but algorithms of very wide effect are already very much among us. More than 80 percent of viewing hours streamed on Netflix originate from automated recommendations. The vast majority of matches on dating apps — from casual "hook-ups" to lifelong partnerships — are initiated by algorithms.

Whose algorithms are these? Right now, it's overwhelmingly profit-seeking corporate interests and, to a far lesser extent, the whims of intellectual-thrill-seeking inventors and engineers. There's a sliver of input from academic ethicists, experimental psychologists, and legal scholars, but even these are all academic voices. When it comes to deciding how self-driving cars will "behave," how about gathering — *and using* — input from a heaping helping of ordinary people: you, your neighbor, your cousin, even your grandmother who shouldn't be driving anymore but still is. How about people who drive — or at least drive around — for a living: UPS and FedEx drivers, cab drivers, newspaper deliverers, EMTs, police, the Good Humor Man.

There is a systematic focus on the good that a new technology will provide, untempered by equal consideration of the harm it might do.

The Possibility of Algorithmic Self-Governance

We want to keep ourselves and others safe. But how much of our "infinite variety" are we willing to sacrifice for that? We can engineer a solution to practically anything. But what if we want to live in a world of ethical diversity and indeed, multiple solutions?

In technology, as in politics, we hand many important decisions over to someone we trust — or at least to someone we trust "enough." But we also have mechanisms, such as elections and investigative reporting, to monitor them. There are, however, no such mechanisms for monitoring the algorithms. One might say we are subject to "algorithm without representation."

And indeed, Jeremy Pitt and colleagues have called for *algorithmic self-governance*, by which they mean a greater role for self-organization in the community systems that will define our society as it becomes increasingly digital (4). Building on the work of the Nobel-Prize–winning

political economist Elinor Ostrom, they call for affirmative attention to a "participation principle" whereby individuals affected by collective choices — like what Google "finds," what Amazon suggests, and what a self-driving car will do in an emergency — should be able to help choose the "rules of the road" [4]. A successful system — in this case, a successful society in which algorithms play a tremendous role in governing everyday life — "must allow for such diversity and support a wide range of institutional types" [4]. The goal is that, rather than relying on top-down, uniform algorithms, we should instead empower individuals and self-organizing institutions to collaborate with policymakers in the shared governance of our increasingly "smart" communities" [5].

> **Our enthusiasm should be tempered.**

To put this plainly, one reason democracy works is that even when some of us do not get the president we wanted — which does happen every time — we do know roughly how that president got elected, by whom, and for what reasons and we know we will have a chance to change that decision at the next election. This is not true, however, of the many algorithms that exercise some control over our daily lives. We don't know how they were designed, by whom, and for what reasons. We will have no opportunity to alter them, although another company may be able to do so through competition. In many cases, we might not even know an algorithm is there at all. Many companies are trying to make it more difficult, not less, for individuals to discern when they are interacting with another human being versus algorithm-driven AI. In June 2014, a computer program (i.e., "Eugene Goostman") passed the "Turing Test" (i.e., a threshold whereby human beings are unable to distinguish computer-generated replies to questions from those of another human being) when it was mistaken for a human more than 30% of the time [6].

Alternatives to Losing Ethical Diversity

There are several practical alternatives to passively accepting the algorithms that corporations devise for their own purposes.

First, design the software to watch how an individual behaves — for example, how he or she drives — and then tune itself to that style. Speech recognition algorithms learn (and quite quickly) how you pronounce the basic sounds in your language; something like that could be done for driving, too.

Second, bring the algorithms out into the open. Generally, we have no idea what algorithms are at work, who created them, and who those people answer to. Since algorithms create moral consequences, and thus reinforce or undercut ethical principles, firms must be held responsible. Their responsibility must be not only for the values their algorithms embody but also for being transparent about it. If an algorithm is designed to preclude individuals from taking responsibility for making a decision, then it is the algorithm's creator who should be held responsible for the algorithm, including the ethical consequences of the algorithm initiated decisions.

Third, algorithms should be really smart *advisors* with access to more data than any individual person could process. In as many cases as feasible, they should suggest choices to enrich human decision making, but not make our choices for us.

Eyes Too Much on the Prize

The loss of ethical diversity to algorithmics is, in fact, a special case of a much larger set of social impacts that we tend to be unaware of until it's too late. The underlying reason for this is a systematic focus on the good that a new technology will provide, untempered by equal consideration of the harm it might do.

A particularly clear example is the merit review process for U.S. National Science Foundation funding applications. This process uses just two criteria: intellectual merit and broader impacts. The latter is defined as a project's "potential to benefit society and contribute to the achievement of specific, desired societal outcomes" [7]. Given the intense competition for NSF funding — some 40 000 proposals a year, out of which only approximately 11 000 are funded [8] — applicants respond to that criterion with a host of "good things" they can credibly foresee resulting from their research, ranging from greater diversity and inclusion in science and engineering to increased public scientific literacy to stronger national security.

This overly simplistic focus in the "Broader Impacts" criteria — that is, its focus on directly foreseeable, specific, *positive* outcomes of the proposed research — is misguided. It is unrealistic and it can have serious consequences. What's wrong with encouraging positive thinking about technology and society? The problem is that it borders on hubris to think that one can clearly see how technologies will evolve and how they will be positively used. Why not ask researchers to instead consider and present a wide range of possible outcomes — positive and negative,

easily foreseen and farther afield but conceivable?

The combination of overly deterministic and overly positive thinking steers both the applicants and the funders away from considering the harm that a given project might make possible. We read every day about destructive uses of the Internet in society — and specifically about huge platforms such as Facebook and Amazon — that are surely not what their inventors and early developers had in mind. But that's likely in part because no one asked. After all, science fiction writers were able to imagine some of the current craziness even in the 1940s.

In this way, the National Science Foundation selection process — and many similar decision-making processes in the world of technology — is the opposite of "algorithmic self-governance." Only the voices of those who will benefit — or who *maybe* will benefit — are clearly heard. The voices of those who will or might be harmed or who will be left out are … left out.

It's long past time to ask researchers and technology developers to think in advance about the social harm their work might cause down the road and not simply to enumerate good things they directly anticipate from their research. That would not, of course, prevent all possible harm — the only way to do that would be to put a permanent stop to all scientific and technological progress. But it might give society a head start in controlling or ameliorating some of that harm.

Some humility about how little *can* be predicted and some willingness to predict what *can* be predicted even when it doesn't make for good selling points for a proposal — these could make our headlong scientific and technological progress at least somewhat less dangerous. The National Science Foundation is in a particularly important, indeed unique and critical, position to encourage such thinking about technology and society.

A More Holistic View of Technology and Society

This is not a call to swing from technophilia and technothusiasm to technophobia. Exciting opportunities are afoot. But as public bodies, corporations, and individual consumers, our enthusiasm should be tempered. To any sufficiently important innovation, there is bound to be a dark side. There is a dark side to pain-reducing and life-saving medications; that's why they are tested and why they have side-effect warnings and why some require a doctor's prescription. A more complete view of technology's social impacts would replace the search for positive impacts with a search for *all* conceivable impacts — positive and negative. The main players in today's advance of technology — both the private companies and the governments — need to lead by example.

In the specific case of autonomous cars, a dangerous barrier to such analysis may be, perhaps surprisingly, the ubiquity of levels-of-driving-automation frameworks. While levels-of-driving-automation frameworks can surely be useful to inform discourse and policy, they are at their core, a set of engineering specifications. As such they do not necessitate consideration of broader goals, including social and humanistic ones. For example, as researchers have noted, by their very nature, levels-of-driving-automation frameworks may rule out more creative forms of cooperation between vehicles and their drivers, i.e., humans and their machines (9), (10).

Governments, for example, must mandate balanced and nuanced analysis. But so too must every stakeholder. We should not let either too much optimism or too much pessimism cloud our vision: the goal of social impact analysis should be scientific realism, so that we as a society get more of the best and less of the worst of influential new technologies.

The goal of social impact analysis should be scientific realism, so that we as a society get more of the best and less of the worst of influential new technologies.

Here are two strategies that can help:

1) Aggressively seek diverse perspectives from different stakeholders. The most challenging circumstances are often completely unexpected because we never even made an effort to look for them. These are what Donald Rumsfeld, the former U.S. Secretary of Defense, memorably called "the unknown unknowns." One reason an array of diverse perspectives is useful is that it allows for a more thorough set of social impacts (positive and negative, easily foreseen and less discernable) to be surfaced. In every field, there are implicit assumptions about "how things work." Typically, those assumptions hold true — that is why they have been adopted and internalized. But they can still backfire under new conditions. One way

to break out of the rut and question implicit assumptions is to solicit advice from unlikely sources who may see the problem differently.

2) Conduct premortems [11]. Often students of the social impacts of new technologies on society find themselves lamenting, with hindsight, negative social impacts of a particular technology. The idea of a premortem is to imagine that a proposed technology has negative impacts, to identify what those would be, and then deduce backwards to understand the reasons why. This approach helps to correct against a very natural bias humans have to assume that actions will have only intended consequences. It forces one to become the devil's advocate: If we have to assume — for the sake of argument — that our lovely technology is actually a social disaster waiting to happen, what might account for that?

People are not just one thing. Neither is any given technology. We know full well now that when we introduce new technologies, we are altering our societies and our own lives within them. We may be creating something that both bullies and defends us. It should not be done without thoughtful examination from as many angles as possible — positive and negative. That, of course, calls upon the very same ethical diversity that is under threat in this Age of Algorithms.

Author Information

Todd L. Pittinsky is a Professor of Technology and Society in the College of Engineering and Applied Sciences at Stony Brook University (SUNY). He is the editor, most recently, of *Science, Technology, and Society: New Perspectives and Directions* (Cambridge University Press).

References

[1] "Moral Machine," M.I.T.; http://moralmachine.mit.edu [Accessed June 30, 2020].

[2] S.D. Hunt and J.M. Hansen, "Understanding ethical diversity in organizations," *Organizational Dynamics*, vol. 36, no. 2, pp. 202–216, 2007; http://sdh.ba.ttu.edu/OrgDyn07%20Ethics.pdf [Accessed Mar. 1, 2020].

[3] L. Plummer, "This is how Netflix's top-secret recommendation system works," *Wired UK*, Sept. 22, 2017. https://www.wired.co.uk/article/how-do-netflixs-algorithms-work-machine-learning-helps-to-predict-what-viewers-will-like [Accessed Mar. 1, 2020].

[4] J. Pitt, and A. Diaconescu, "Structure and governance of communities for the digital society," in *Proc. 2015 IEEE Int. Conf. Autonomic Computing*, pp. 279-284, 2015. https://ieeexplore.ieee.org/document/7266980 [Accessed Apr. 2, 2020].

[5] J. Pitt, D. Busquets, A. Diaconescu, A. Nowak, A. Rychwalska, and M. Roszczynska-Kurasinska, "Algorithmic self-governance and the design of socio-technical systems," in *CEUR Workshop Proc.*, vol. 1283, pp. 262–273, Jan. 2014. http://ceur-ws.org/Vol-1283/paper_33.pdf [Accessed March 15, 2020].

[6] BBC, "Computer AI passes Turing test in 'world first'," June 9, 2014. https://www.bbc.com/news/technology-27762088 [Accessed April 5, 2020].

[7] National Science Foundation, "Perspectives on broader impacts," Jan. 1, 2015. https://www.nsf.gov/od/oia/publications/Broader_Impacts.pdf [Accessed Feb. 20, 2020].

[8] https://www.nsf.gov/funding/aboutfunding.jsp [Accessed Feb. 20, 2020].

[9] M. Johnson, J. M. Bradshaw, and P. J. Feltovich, "Tomorrow's human–machine design tools: From levels of automation to interdependencies," *J. Cognitive Engineering and Decision Making*, vol. 12, no. 1, pp. 77-82, 2018.

[10] E. Stayton and J. Stilgoe, "It's time to rethink levels of automation for self-driving vehicles," *IEEE Technology and Society Mag.*, this issue.

[11] G. Klein, "Performing a project premortem," *IEEE Engineering Management Rev.*, vol. 36, no. 2, pp. 103-104, May 30, 2008. https://ieeexplore.ieee.org/document/4534313 [Accessed Apr. 1, 2020].