A Zen Buddhist monk's approach to democratizing AI

May 29 2020, by Katharine Miller

Colin Garvey is a research fellow with Stanford's Institute for Human-centered Artificial Intelligence. Credit: Stanford University

Colin Garvey, a postdoctoral research fellow at Stanford University's Center for International Security and Cooperation (CISAC) and Institute

for Human-centered Artificial Intelligence (HAI), took an unusual path to his studies in the social science of technology. After graduating from college, he taught English in Japan for four years, during which time he also became a Zen Buddhist monk. In 2014, he returned to the U.S., where he entered a Ph.D. program in science and technology studies at Rensselaer Polytechnic Institute. That same year, Stephen Hawking co-authored an editorial in *The Guardian* warning that artificial intelligence could have catastrophic consequences if we don't learn how to avoid the risks it poses. In his graduate work, Garvey set out to understand what those risks are and ways to think about them productively.

As an HAI Fellow, Garvey is working on turning his Ph.D. thesis into a book titled "Terminated? How Societies Can Avert the Coming AI Catastrophe." He is also preparing a policy report on AI-risk governance for a Washington, D.C.-based think tank and guest editing "AI and Its Discontents," a special issue of *Interdisciplinary Science Reviews* featuring diverse contributions from sociologists to computer scientists, due out this December.

Here he discusses the need to change how we think and talk about AI and the importance of democratizing AI in a meaningful way.

How does the public's tendency to see AI in either utopian or dystopian terms affect our ability to understand AI?

The risk of accepting the utopian or dystopian narrative is that it reinforces a very common attitude toward the evolution of AI and technology more generally, which some scholars describe as technological determinism. Either the market forces are inescapable, or, as some AI advocates might even say, it's human destiny to develop a machine smarter than humans and that is the next step in evolution.

I think this narrative about inevitability is actually deployed politically to impair the public's ability to think clearly about this technology. If it seems inevitable, what else is there to say except "I'd better adapt"? When deliberation about AI is framed as how to live with the impact, that's very different from deliberating and applying public control over choosing what kind of impact people want. Narratives of inevitability ultimately help advance the agenda of beneficiaries of AI, while sidelining those at risk, leaving them very few options.

Another problem is that this all-good or all-bad way of framing the subject reduces AI to one thing, and that is not a good way to think about complex problems. I try to break that up by mapping risks in specific domains—political, military, economic, psychosocial, existential, etc. – to show that there are places where decision making can go differently. For example, within a domain, we can identify who is benefiting and who is at risk. This allows us to get away from this very powerful image of a Terminator robot killing everyone, which is deployed quite often in these types of conversations.

AI is not the first technology to inspire dystopian concerns. Can AI researchers learn from the ways society has dealt with the risks of other technologies, such as nuclear power and genetic engineering?

In the mid 20th century, social scientists who critiqued technology were very pessimistic about the possibility of humanity controlling these technologies, especially nuclear. There was great concern about the possibility of unleashing something beyond our control. But in the late 1980s, a second generation of critics in science and technology looked at the situation and said, here we are and we haven't blown up the world with nuclear weapons, we haven't released a synthetic plague that caused cancer in a majority of the population. It could have been much worse,

and why not? My advisor, Ned Woodhouse, looked into these examples and asked, when things went right, why? How was catastrophe averted? And he identified five strategies that form the Intelligent Trial and Error approach that I have written about in relation to AI.

One of the Intelligent Trial and Error strategies is public deliberation. Specifically, to avert disaster, deliberation should be deployed early in development; a broad diversity of concerns should be debated; participants should be well-informed; and the deliberations should be deep and recurring. How well do you think AI is doing on that score?

I would say the strategy of deliberation could be utilized more thoroughly in making decisions about risk in AI. AI has sparked a lot of conversations since about 2015. But AI had origins in the 1950s. One thing I've found is that the boom and bust cycle of AI hype leading to disillusionment and a crash, which has happened roughly twice in the history of AI, has been paralleled by quite widespread deliberation around AI. For example, in the '50s and '60s there were conversations around cybernetics and automation. And in the '80s there was a lot of deliberation about AI as well. For example, in the 1984 meeting of the ACM [Association for Computing Machinery], there were social scientific panels on the social impacts of AI in the main conference. So there has been a lot of deliberation about AI risk, but it's forgotten each time AI collapses and goes away in what's popularly known as an "AI winter." Whereas with nuclear technology, the concern has been more ongoing, and that influenced the trajectory of the nuclear industry.

One way of looking at how little deliberation is going on is to look at examples of privacy violations where our data is used by an AI company

to train a model without our consent. We could say that's an ethical problem, but that doesn't tell you how to solve it. I would reframe it as a problem that arose because decisions were made without representatives of the public in the room to defend the citizens' right of privacy. This puts a clear sociological frame around the problem and suggests a potential strategy to address the problem in an institutional decision-making setting.

Google and Microsoft and other large companies have said that they want to democratize AI, but they seem to focus on making software open source and sharing data and code. What do you think it should mean for AI to be democratized?

In contrast to economic democratization, which means providing access to a product or technology, I'm talking about political democratization, which means something more like popular control. This isn't mob rule; prudence is a key part of the framework. The fundamental claim is that the political system of democratic decision making is a way to achieve more intelligent outcomes overall compared to alternatives. The wisdom of crowds is a higher order effect that can arise when groups of people interact.

I think AI presents us with this challenge for institutional and social decision making, in that as you get more intelligent machines, you'll need more intelligent democracies to govern. My book, based on my dissertation, offers some strategies for improving the intelligence of decision making.

What's an example of how democratizing AI might make a difference today?

One area I'm watching closely and working on is the AI arms race with China. It's painted as a picture of authoritarian China on the one hand and democracy on the other. And the current administration is funding what they call "AI with American values." I would say that's great, but where is democracy among those values? Because if they only refer to the values of the market, those are Chinese values now. There's nothing distinct about market values in a world of global capitalism. So if democracy is America's distinguishing feature, I would like to see the big tech companies build on that strength rather than, as I see happening now, convincing policy makers and government officials to spend more on military AI. If we've learned anything from the last cold war arms race, it's that there really aren't winners. I think a long-term multi-decade cold war with China over AI would be a race to the bottom. A lot of AI scientists would probably agree, but the same narrative framed in terms of inevitability and technological determinism is often used here in the security space to say, "We have no choice, we have to defeat China." It will be interesting to see what AI R&D gets justified by that narrative.

Is there a connection between your Buddhism and your interest in AI?

When people hear that I'm a Zen Buddhist monk, they often say, you must want to tell programmers to meditate. But my concern has more to do with reducing suffering in the world. I see a huge risk for a profound kind of spiritual suffering that we are already getting some evidence of. Deaths of despair are an epidemic in the United States; and there's a steep rise of suicide and depression among teenagers, even in the middle class. So there are some surprising places where material abundance isn't translating into happiness or meaning. People are often able to withstand serious suffering if they know it's meaningful. But I know a lot of young people see a pretty bleak future for humanity and aren't sure where the meaning is in it all. And so I would love to see AI play a more positive

role in solving these serious social problems. But I also see a potential for increased risk and suffering, in a physical way, maybe with killer robots and driverless cars, but potentially also psychological and personal suffering. Anything I can do to reduce that gives my scholarship an orientation and meaning.

In a world where much AI R&D is privatized and driven by capitalist profit motives at corporations around the globe, is it possible for thought leaders at a place like Stanford to make a difference in the trajectory of AI research overall?

Stanford certainly has the institutional capital and cultural cachet to influence the AI industry; the question is how it will use that power. The major problems of the 21st century are problems of distribution, not production. There's already enough to go around; the problem is that a small fraction of humanity monopolizes the resources. In this context, making AI more "human-centered" requires focusing on the problems facing the majority of humanity, rather than Silicon Valley.

To pioneer a human-centered AI R&D agenda, thought leaders at Stanford's HAI and elsewhere will have to resist the powerful incentives of global capitalism and promote things like funding AI research that addresses poor people's problems; encouraging public participation in decision making about what AI is needed and where; advancing AI for the public good, even when it cuts into private profits; educating the public honestly about AI risks; and devising policy that slows the pace of innovation to allow social institutions to better cope with technological change.

Stanford has a chance to lead the world with innovative approaches to solving big problems with AI, but what problems will it choose?