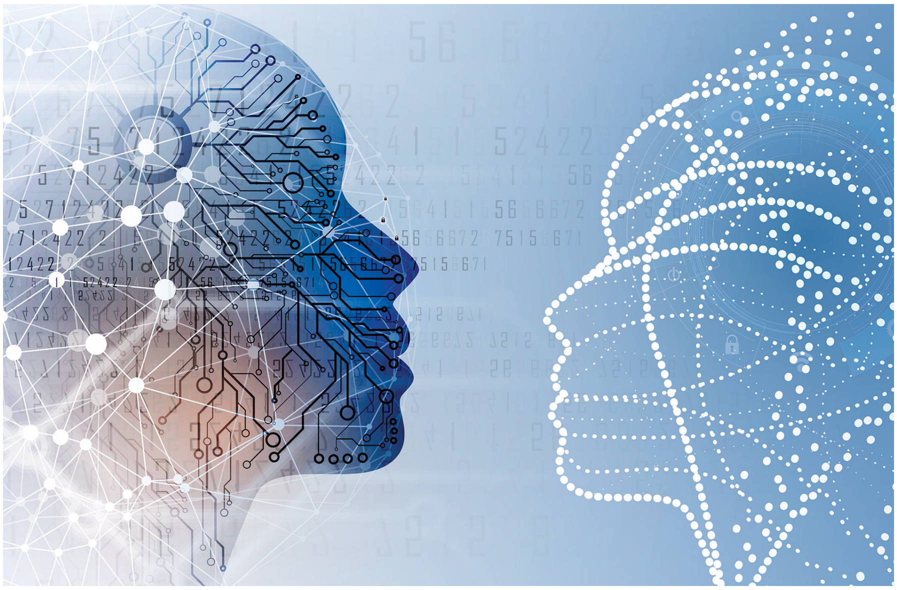
DOI:10.1145/3478516

Illah Reza Nourbakhsh

Viewpoint
AI Ethics: A Call to Faculty

Integrating ethics into artificial intelligence education and development.



IMAGERY BY VS148/SHUTTERSTOCK

THIS PAST YEAR has seen a significant blossoming of discussions on the ethics of AI. In working groups and meetings spanning IEEE, ACM, U.N. and the World Economic Forum as well as a handful of governmental advisory committees, more intimate breakout sessions afford an opportunity to observe how we, as robotics and AI researchers, communicate our own relationship to ethics within a field teeming with possibilities of both benefit and harm. Unfortunately, many of these opportunities fail to realize authentic forward progress during discussions that repeat similar memes. Three common myths pervade such discussions, frequently stifling any synthesis: education is not needed; external regulation is undesirable; and technological optimism provides justifiable hope.

Education

The underlying good news is that discourse and curricular experimentation are now occurring at scales that were unmatched in the recent past. World Economic Forum working groups, under the leadership of Kay Firth-Butterfield, have convened a series of expert-driven policy productions are topics including, for instance, the ethical use of chatbots in the medical field and in the financial sector. The IEEE Global Initiative on the Ethics of Autonomous and Intelligent Systems, led by John Havens, continues to make progress on international standards regarding the ethical application of robotics and AI. These are just two of dozens of ongoing international efforts. Curricular experiments have also garnered successful publication, from single-course pilots[2] to whole-curricular interventions across required course sequences.[3]

Yet despite international policy discourse and published curricular successes, the vast majority of faculty in robotics and AI report, in private discussion, that they do not feel empowered or prepared to integrate ethics into their course materials. The substance of this hesitation rests on the notion that 'teaching AI ethics' is like teaching ethics itself—lecturing on utilitarian and Kantian frameworks, for instance, which is best taught by ethics scholars. But AI ethics is not the science of ethics, but rather shorthand for the notion of applying ethical considerations to issues surfaced by AI technologies: surveillance, information ownership, privacy, emotional manipulation, agency, autonomous military operations, and so forth. As for integrating such reflection into an AI class, every case I am aware of does so, not with *sage on a stage* lecturing by the faculty member regarding Kant, but with case studies and small-group discussions on complex issues, lifting the students' eyes up from the technology to considering its possible social ramifications. No teacher can set the stage for such discussions better than an AI

expert, who can speak concretely about face recognition errors, and how such mistakes can be inequitably distributed across marginalized populations.

In a five-year experiment, I have collaborated with a professor in the College of Social Sciences at Carnegie Mellon to design and deploy AI and Humanity as a freshman course that encourages technologies and humanities students alike to develop a grammar for considering and communicating about the interplay between AI and robotics technologies and power relationships in society. We build a new grammar on the backbone of keywords, thanks to McCabe and Yanacek's outstanding analysis of critical themes, including *surveillance*, *network*, *equality*, *humanity*, *technology*.[6,7] College freshman have shown an apt ability to conduct critical inquiry, evaluate the ethical ramifications of technology and even construct futuring visions that interrogate our possible trajectories as a society (see http://aiandhumanity.org).

Equally important is the question of how ethics can be integrated into extant, technical coursework throughout a department. This year, in collaboration with Victoria Dean at Carnegie Mellon, we revisit the question of curricular integration by deploying a graduate class with the capstone experience of students engaging with faculty in the Robotics Institute, studying each course's syllabus, and designing a complete ethics module for integration into each class. We believe this direct-intervention model, with case studies, futuring exercises and keywords at its heart, has the potential to affirmatively engage numerous courses and professors across our department with a low-barrier pathway to in-class ethics conversations. As Barbara Grosz and others have said, the ethics conversation should not be a one-time course, nor a one-time seminar. Thinking on societal consequences should happen regularly, so it becomes an enduring aspect of the design thinking around new AI and robotics technology research and development.

Regulation

Another common argument stems from a strong *anti-regulation* stance that embraces corporations as agents with the very best intentions. AI researchers in my workshops frequently point to ethics review programs implemented by top corporations to show that, being global hubs of innovation, these companies have already invented the best ways to self-regulate, eliminating the need for oversight. But the existence of a few corporate ethics programs does not upend the most basic observation of all: corporate technologic titans have incentives and reward structures that are *not* directly aligned with public good, equity, and justice. The misalignment of public-private values is a perpetual temptation for corporations to veer off the ethical course to privilege private interests over public concerns.

Examples abound. Google created an ethics board, and included the president of the Heritage Foundation. When employees noted the inclusion of an individual dedicated to denying climate change and fighting LGBTQ rights, Google dissolved the entire ethics board after one week.[10] In 2019, news organizations also reported that Amazon's Alexa and Google's Assistant both record audio, unbeknownst to home occupants, and that employees listen to home interactions that any reasonable user would presume to be private.[8] When the story first took hold, Apple touted its stronger privacy positioning, boasting that, in contrast, no Apple employees ever listen to Siri.

That was the end of the news cycle, until former Apple contractor Thomas le Bonniec became a whistleblower and described a vast program in which 200 contractors in County Cork, Ireland, were listening to Siri recordings that were very private.[4] Apple was strictly right, *employees* were not listening in, contractors were. The malintent of this fib is clear; but the larger lesson is key: corporations are beholden to their shareholders and to their own set of values and motives. We cannot expect their self-regulation to serve any purpose beyond their own value hierarchy.

We live in a world replete with examples of misaligned values that facilitate unjust outcomes; regulation of corporate technology innovation by corporations constructs a value misalignment between corporate mission and public good. As AI researchers, we derive legitimacy through our reasoned opinions regarding the arc of future technology innovation, including the use of guard rails that protect the public good. We can best serve both corporations and the public, not by arguing that regulation stifles innovation—we are all keenly aware that poorly designed regulation does that. Rather we can innovate by helping facilitate the creation of well-designed regulation, together with policymakers and industry, that encourages the *most just* AI futures and memorializes corporate transparency for the public.

Technological Optimism

It is one of the greatest ironies of these AI workshops when researchers argue that they do not feel equipped to opine on the ethics of AI in their classes at university, yet in the same breath announcing that their AI systems will be ethical because they will design autonomous technologies to have built-in ethical governors. This disconnect arises out of a natural bias we have as innovators: we have spent entire careers practicing how to be technology-optimistic—how to imagine a future with inspiring, new inventions that we can create. This is the attitude we need as salespeople, to convince funders to make bets on our future work; and yet this optimism does a disservice when we use it *within* our institution to imagine that shortcomings in present-day AI systems will be resolved simply through innovation.

In the 1990s, the AI field was far removed from social impact because it was as impractical as theoretical mathematics. Exciting progress, at the very best, resulted in publication. That world is ancient history now. To say that AI, today, is a technical discipline is entirely naïve: it is a social, worldwide experiment. Our tools have teeth that cut into the everyday lives of all, and this leaves a collection of engineers and scientists in the awkward position of having far more impact on the future than is their due.

In earlier times, our computational peers forged Computer Professionals for Social Responsibility (CPSR), largely in response to the threat of thermonuclear destruction and other existential threats arising from the Strategic Defense Initiative. Because nuclear destruction was palpable, the arc from technology to personal responsibility was short and well-founded. But today our AI technology is not as obviously threatening. When misused, AI's reinforcement of bias and power configurations in society can be insidious and sub-lethal, like petrochemical industry toxins that hurt entire communities, not as quickly as bullets, but across vastly greater scope and timescales. And unintended side effects are not limited in potential scope; when AI-led political micro-marketing directs the outcome of an election, ensuring undemocratic policy decisions *can* have existential impact on the population.

The AI research community cannot sit this out.

Yet publicly consumed literature ranges dramatically on the issue of technology optimism and technology realism. The singularity, espoused by Kurzweil, suggests a postmodern evolutionary pathway for a new humanity[4] or a pathway to greater equity through low-cost robotic production.[9] At the same time, counter-narratives explain the role of ritual surveillance in the very creation of the Internet[5] as well as the ethical ramifications of war-fighting robots.[1] We, as public outreach specialists need to reference the existing literature on *both* sides and add to the body of counter-narratives, creating depth and sharp focus along each critical issue where society and AI technology meet, from surveillance and information ownership to authenticity and democracy.

If you are not concerned about the effects of fielded AI systems on democracy, on stakeholder capitalism, on power and bias in society, then you are operating on an unfounded level of optimism that goes against your own scientific nature.

Conclusion

The AI research community cannot sit this out. We are a critical expert group with sufficient know-how to separate authentic issues from hyperbole, to distinguish plans of action that can actually make a difference from hot air. If we do not become part of the solution, we will lose our legitimacy as well-intentioned visionaries.

Education for all stakeholders is imperative for awareness. AI is the very definition of a boundary technology that is sufficiently alien that *everyone* needs scaffolding to make informed decisions; and we cannot pass off the duty of care to create broad educational interventions to anyone else. Rule-making and regulation is equally essential. Nothing about historical corporate and governmental behavior can rationalize a *laissez-faire* approach when the consequences of inaction are so clearly inequitable. Finally, the hyperbole of techno-optimism needs to end. The public invests our opinions with significant credence, and when we state that our algorithms will be ethical innately, they actually imagine autonomous systems with human meta-cognition. There is no room for us to promulgate such a gap between computational reality and blue-sky wishes, particularly when AI is already so consequential to our lived experience. Let's embrace strong education, clear-headed regulation, and let's tone down the hyperbole of technological optimism. C

References

1. Chamayou, G. *A Theory of the Drone.* New Press, 2015.
2. Furey, H. and Martin, F. Introducing ethical thinking about autonomous vehicles into an AI course. *Thirty-Second AAAI Conference on Artificial Intelligence.* 2018.
3. Grosz, B.J. et al. Embedded EthiCS: Integrating ethics across CS education. *Commun. ACM 62*, 8 (Aug. 2019), 54–61.
4. Hern, A. Apple Whistleblower goes Public over 'lack of action.' *The Guardian*, (May 20, 2020).
5. Kurzweil, R. *The Singularity Is Near: When Humans Transcend Biology.* Penguin, 2005.
6. Levine, Y. *Surveillance Valley: The Secret Military History of the Internet.* Public Affairs; Illustrated edition (Feb. 6, 2018).
7. MacCabe, C. and Yanacek, H. Eds. *Keywords for Today: A 21st Century Vocabulary.* Oxford University Press, 2018.
8. Nourbakhsh, I.R. and Keating, J. *AI and Humanity.* MIT Press, 2020.
9. O'Flaherty, K. Amazon staff are listening to Alexa conversations—Here's what to do. *Forbes*, (Apr. 12, 2019).
10. Rifkin, J. T*he Zero Marginal Cost Society: The Internet of Things, the Collaborative Commons, and the Eclipse of Capitalism.* St. Martin's Press, 2014.
11. Wakefield, J. Google's ethics board shut down. BBC News (Apr. 5, 2019).

Illah Reza Nourbakhsh (illah@cs.cmu.edu) is Executive Director, Center for Shared Prosperity Director, CREATE Lab K&L Gates Professor of Ethics and Computational Technologies Carnegie Mellon University, Pittsburgh, PA, USA.

Watch the authors discuss this work in the exclusive *Communications* video. https://cacm.acm.org/videos/ai-ethics