An Academic Vision for AI Ethics

By: Bryant Walker Smith

In early 2022 I was asked to write a brief statement on an academic vision for AI ethics. As this topic has become only more important in the year since, I thought I'd share my statement here:

Ethics and AI

Ethics, in my view, refers to normative inputs that the sciences cannot independently provide. Engineers, for example, can supply terms, values, and methodologies for cost-benefit analysis, but without recourse to ethics they cannot say whether such an exercise is appropriate or whether its methods for valuing human life or discounting uncertainty are externally proper.

For this statement, ethics encompasses at least two dimensions. The first contemplates ethical *issues:* normative questions of right and wrong that will likely remain contested. The second contemplates ethical *imperatives:* consequences of the positions taken on ethical issues. Exploration of ethical issues is akin to the natural sciences, which continually test their foundations. Implementation of ethical imperatives is akin to the applied sciences, which employ workable principles from the natural sciences to achieve tangible human goals.

I emphasize this distinction because I understand this ethics project to implicate public policy without presuming to subsume it. The primacy of humanity, for example, is an ethical issue. The adoption of laws protecting any such primacy may be an ethical imperative, but ethics must share that policy debate with many other disciplines that have an equal (or equally limited) claim to policy expertise.

I adopt a broad view of AI as the automated resolution of uncertainty through experience, which encompasses applications of machine learning that already exist and extends to conceptions of artificial general intelligence that do not.

The Role of Ethics of AI in a University Setting

A university committed to AI ethics can powerfully:

Identify, frame, and inform key issues. AI discussions are often dominated by the shiny over substantive, urgent over important, and financially lucrative over socially beneficial. A university can instead highlight the questions, answers, and—critically—voices that might otherwise be overlooked or undervalued in the design, deployment, regulation, and evaluation of AI. It can foster exploration by connecting diverse actors through common language, shared knowledge, and credible structures. Today's insights and interventions could have profound effects tomorrow—akin to nudging an asteroid while it is still billions of miles from Earth.

Develop and communicate an affirmative vision for ethical AI. A university can describe both a future in which AI is an ethical good and a path to reach this future. This substantive and procedural vision embraces opportunities while mitigating risks. It enlists numerous disciplines by situating ethical imperatives as motivations as well as limitations. Interdisciplinary collaboration is especially important on contested issues where arguments from other domains may be as persuasive as those that come from ethics. For example, whether to address a policy issue through an evolutionary or revolutionary approach may depend as much on social science as on ethics.

Model that visions throughout the work of the university. Even today, AI is used far beyond engineering. A university can fully embrace the responsible use of AI in its teaching,

research, service, and even administration across its disciplines. A sandbox for ethical AI can highlight rather than obscure ethical challenges through stories of success and failure. By transparently showing its work on AI ethics, A university can model the trustworthiness that will be essential to this field.

Major Challenges and Opportunities for the Field

Clarify AI. A contemporary chatbot is categorically different from artificial general intelligence. Five years is different from fifty. Common AI fears—mass murder, control, displacement, disempowerment, disruption, and discomfort—are radically different from each other.

Map our relationship to AI. Human affection, fallibility, and judgment will remain central to both ethics and AI. Conceptually, humans can be designers, users, subjects, or elements of AI systems, and AI applications can be understood as products, services, agents, instruments, or conceivably persons—each with distinct ethical and legal ramifications. Reality may circumscribe human authority far more than intended or desired. Vexing boundary problems could arise as technologies mature: In a connected future, what is a single robot or even a single human?

Be explicit and inclusive. System designers, training data, and intended users do not reflect the whole of the human experience—not in this country and certainly not in the world. Design and regulatory decisions often hide rather than highlight meaningful ethical issues. Debates about whether systems should be open or closed, centralized or distributed, simple or complex, and certain or flexible are about competing philosophies as much as conflicting evidence.

Contextualize AI. Tomorrow's AI will exist in tomorrow's world—alongside changes to norms, laws, conditions, and other technologies. Not every change is unprecedented. Prior technologies offer limited lessons about speed/connectivity, disruption/distortion, adaptation/exploitation, identity/culture, systemization/centralization, risk/uncertainty, and trust/trustworthiness. Human bodies and human societies provide examples of systems potentially as complex, dynamic, and stochastic as AI.

Recognize the novel. Understanding how AI could be truly unprecedented, by degree or by kind, can target ethics discussions. Exponential technological improvement, breakneck social change, massive power concentration, and human- or god-like perceptions of AI could fundamentally challenge conventional ethics.

Understand AI as an instrument of power. AI ethics should focus on who could be intentionally or unintentionally empowered or disempowered—governments, companies, individuals, collectives, even animals. In this way, the differences between centralized and decentralized systems could be more consequential than the differences between humans and machines. I have long argued that the popular question of whether a technology is trusted should give way to the question of whether the companies behind that technology are trustworthy. Privacy is part of this story. Changing power dynamics also implicate notions of discrimination, default rules of society, and tensions between autonomy and community.

Manage social change. AI promises profound changes. Human and technological lifecycles could become increasingly incompatible. Macroscopic vibrancy could obscure microscopic despondency. Equilibria could become explosions. An ethical approach to AI may

accordingly demand human-focused pressure releases and safety nets that have little to do with AI itself.

Think *around* AI. The field of AI ethics must credibly engage with AI and with *everything else*. Visionaries often turn their attention—eventually—from technology toward humanity. A key challenge and opportunity for AI ethics is to emphasize education, equity, and justice far sooner—to ensure that AI ultimately serves humanity by reflecting and amplifying our better self.

This article is cross-posted on the blog for Stanford Law School's Center for Internet and Society