The Social & Economic Implications of Artificial Intelligence Technologies in the Near-Term
July 7th, 2016; New York, NY
http://artificialintelligencenow.com

WORKSHOP PRIMER: ETHICS & AI

Contents

Brief description

As artificial intelligence systems permeate more aspects of human life, complex questions arise about the ethics of their design and implementation.[1] The diverse range of contexts in which AI systems are already being used – from medical devices to insurance premiums to personalized ad delivery – have led some to ask if the deployment of these systems necessitates a revision of existing ethical frameworks.[2]

Classic ethical frameworks often examine situations and actions contextually, focusing on the relationships between human actors and the benefits and risks to different actors implied in these relations. For example, such frameworks might take into account how a specific doctor cares for her patients, how a specific researcher studies or experiments

[1] As Russell and Norvig point out, the history of artificial intelligence has not produced a clear definition of AI but rather can be seen as variously emphasizing four possible goals: “systems that think like humans, systems that act like humans, systems that think rationally, systems that act rationally.” In the context of this primer on ethics, we are relying on the emphasis proposed by Russell and Norvig, that of intelligence as rational action, and that “an intelligent agent takes the best possible action in a situation.” Stuart J. Russell and Peter Norvig, *Artificial Intelligence: A Modern Approach*, Englewood Cliffs, NJ: Prentice Hall, 1995: 27.

[2] See, for instance, Mike Ananny, “Towards an Ethics of Algorithms: Convening, Observation, Probability, and Timeliness,” *Science, Technology, & Human Values* 41, no. 1 (2016): 93-117. “Part of understanding the meaning and power of algorithms,” Ananny writes, “means asking what *new demands* they might make of ethical frameworks[.]” Ibid, 93, emphasis added.

on various subjects, or how the leadership of a company interacts with customers or workers.[3]

The increasingly use of AI has increased the prominence of emerging fields like “machine ethics,” “data ethics,” and “AI ethics.” These terms point to the underlying dynamics of socio-technical systems and machine intelligence, and have served to highlight the complexity of automated tasks and outputs where the original context is removed or difficult to define. This primer aims to surface foundational concerns, inquiries, and ethical questions arising from current implementations of AI in various areas of human life. However, we make no attempt to condense all approaches to exploring the ethics of technical systems into a comprehensive summary, nor advocate for a single definition of what AI ethics should be.

Given the profound implications that AI systems can produce on resource allocation, and their potential to concentrate power and information, key questions need to be asked to ensure these systems are not harmful, especially to already marginalized groups. These questions include: How are we to evaluate the ethical implications of AI systems in relation to the public good (and how are we to generally define “public good”)? What forms of disclosure, accountability, consent, and justice should we expect from and hold AI systems accountable to? If traditional ethical frameworks have served as an opportunity to interrogate underlying values of human action, how do we do the same with computer-augmented action in AI? How do we think about ethics in systems designed to ‘learn’ without human supervision? And how do we ensure that we do not merely entrench practices of discrimination and injustice, given that human history (and historical data) will often reflect these biases?

Machine ethics

Recent discussions of ethics and AI have tended to prioritize the challenges posed by hypothetical general AI systems in a distant future, such as the advent of the “singularity” or the development of a superintelligence that might become an existential threat to humanity.[4] Discussions of AI focusing on such speculative futures have tended to elide or ignore the immediate ethical implications of AI systems in the near- to medium-term, including the immediate challenges posed by the enormous number of task-specific AI systems currently in use.[5] Contemporary AI systems perform a diverse range of activities

[3] For a discussion of researcher ethics with particular emphasis on the ethical challenges posed by human subjects and data privacy, see: Michael Zimmer, “‘But the Data Is Already Public’: On the Ethics of Research in Facebook,” *Ethics and Information Technology* 12, no. 4 (2010): 313-325; Jacob Metcalf and Kate Crawford. "Where are Human Subjects in Big Data Research? The Emerging Ethics Divide." *Big Data and Society,* Spring (2016), http://bds.sagepub.com/content/3/1/2053951716650211

[4] See, for example, Nick Bostrom, *Superintelligence: Paths, Dangers, Strategies*, Oxford: Oxford University Press, 2014. For an early discussion of “ultraintelligent” machines, see Irving John Good, “Speculations Concerning the First Ultraintelligent Machine,” *Advances in Computers*, vol. 6, Eds. Franz Alt and Morris Rubinoff, New York / London: Academic Press, 1965, 31-88.

[5] “Task-specific” here does not imply any statement about complexity. The spell-checker in your word processor and the program that governs a “self-driving” car are both task-specific systems even though the complexity of the models used

and pose new challenges to traditional ethical frameworks due to the implicit and explicit assumptions made by these systems, and the potentially unpredictable interactions and outcomes that occur when these systems are deployed in human contexts.

In the 1950s, Norbert Wiener, in his book examining the implications of cybernetics for society, identified "*Liberté*, *Egalité*, [and] *Fraternité*" as "concepts...necessary for the existence of justice."[6] Wiener contended that individuals must have:

> [T]he liberty of each human being to develop in his [sic] freedom the full measure of the human possibilities embodied in him; the equality by which what is just for A and B remains just when the positions of A and B are interchanged; and a good will between man and man that knows no limits short of those of humanity itself.[7]

Wiener further contends that these ideals have no teeth unless "the law" is "unambiguous" and that no individual is subject to coercion.[8] Many ethical guidelines conform to some notion of these concepts, even though ethical frameworks are far from one-size-fits all.[9]

As an example, we can see the field of *research ethics* as a specific attempt to embody these values. Research ethics are broadly concerned with collecting data while maintaining a research subject's safety and anonymity. This includes destroying information about subjects when possible, limiting access to the data collected, and of course, not performing research that would harm the subject or others. This approach can be viewed through Wiener's framework as an attempt to promote *equality,* performed in the spirit of *fraternity,* with the goal of ensuring *liberty*. Likewise, the role of freely given consent in the U.S. Common Rule governing federally-funded human subjects research is similarly designed to facilitate the advancement of these values. The means to adhere to these ethical values when we move into the world of computer facilitated action and decision making is not simple, however.

In his influential essay on computer ethics, James Moor argues that the "transforming effect of computerization" is such that the "basic nature or purpose of an activity or

in these two tasks is very different. For context on superintelligences versus near-term ethical challenges see Kate Crawford, "Artificial Intelligence's White Guy Problem," *New York Times*, June 25, 2016.

[6] Norbert Wiener, *The Human Use of Human Beings: Cybernetics and Society*, Boston: Houghton Mifflin Co., 1954, 105.

[7] Ibid, 105-106, emphasis added.

[8] Ibid, 107.

[9] Mike Ananny provides a good summary of the three core "subareas" of ethics: (1) a "*deontological* approach" that relies on "duties, rules, and policies [that] define actions as ethical"; (2) a "*teleological* approach," where ethics is maximizing "good" for a particular group; and (3) a "*virtue* model of ethics" that examines "subjective, idiosyncratic and seemingly nonrational impulses that influence people in the absence of clear rules." Ananny, "Toward an Ethics of Algorithms: Convening, Observation, Probability, and Timeliness," 94. Each of these "subareas" can provide a slightly different interpretation of how ethical ideals might be implemented.

institution is changed," which also changes how we value it.[10] The shift to computerized systems in the social world produces what Moor calls "policy" and "conceptual vacuums."[11] Moor also notes that there are "invisibility factors" involved with the application of computers to specific problems, in that "one may be quite knowledgeable about the inputs and outputs of a computer," but have little understanding of its "internal processing."[12] This inscrutability, Moor notes, is by design. Moor observes that computers effectively hide those tasks we wish to automate, and that this elision produces at least three sites for potential ethical transgression: (1) "*invisible abuse*" where code is maliciously inserted, or the system otherwise does other than what is expected and intended by its "user," (2) "*invisible programming values*" in which non-trivial decisions made by a programmer result in important unintended mistakes (think of a software bug resulting in the movement of a decimal point), and (3) "*invisible complex calculation*" where the processes are too complex to be reviewed and understood by humans, making review, correction, or validation difficult, if not impossible.[13] We could think here of a complex, faster-than-human-thought system like high frequency trading (HFT) or the application of AI-driven predictive policing as spaces where policy and governance present ongoing challenges.[14]

Three challenges for AI ethical frameworks

1. Within reason, please: developing expectations of machine performance

Rearticulating a social problem as a technical problem to be solved by AI is not a neutral translation: framing a problem so as to be tractable to an AI system changes and constrains the assumptions regarding the scope of the problem, and the imaginable solutions.[15] Such social-to-technical translations provide no assurance that an AI system will produce fewer mistakes than the system it is intended to replace. As Ryan Calo points out, while it is usually believed that AI systems (such as self-driving cars) will commit fewer errors than humans – and, indeed, they may commit *fewer errors of the kind that humans do* – for any system of even modest complexity, the AI system will inevitably produce new kinds of errors, potentially of a type that humans do not make

[10] James Moor, "What is Computer Ethics?" *Metaphilosophy* 16, no. 4 (1985): 271. "As we consider different policies we discover something about what we value and what we don't." Ibid, 267.

[11] Ibid, 266.

[12] Ibid, 272.

[13] Ibid, 272-275.

[14] For example, in a study of the global financial system, Neil Johnson et al. indicated "an abrupt transition to a new all-machine phase characterized by large numbers of subsecond extreme events" that escalated in the build up to the 2008 financial crisis. The study concludes that there is an "emerging ecology of competitive machines featuring 'crowds' of predatory algorithms." Johnson et al. argue there is now a "need for a new scientific theory of subsecond financial phenomena." Neil Johnson, Guannan Zhao, Eric Hunsader, Hong Qi, Nicholas Johnson, Jing Meng, and Brian Tivnan, "Abrupt Rise of New Machine Ecology beyond Human Response Time," *Scientific Reports* 3, article #2627 (2013): 1.

[15] For a discussion of the "two basic problems for any overarching classification scheme" (which always come into play at some stage in the development of an AI), see Geoffrey Bowker and Susan Leigh Star, *Sorting Things Out: Classification and its Consequences.* Cambridge: MIT Press, 1999, 69-70.

(and thus, in many cases, that our current ethical frameworks may be ill-equipped to address).[16]

In one of the foundational papers in the field of AI, Alan Turing argued that for a "learning machine," "processes that are learnt do not produce certainty of result; if they did they could not be unlearnt."[17] This implies that whatever a machine is capable of learning, when it acts on this knowledge it is *guaranteed* a nonzero probability of making mistakes. In the case of Spotify picking a song we don't like, the cost of a mistake is minimal. But what should our ethical position be when our AI doctor is guaranteed to produce a nonzero number of mistakes, mistakes that may be very different from those human doctors would make? Traditional ethical frameworks assume errors and mistakes will occur and contribute to the ongoing development of professional and institutional standards in the relevant field. For AI, too, we must develop such standards as well as mechanisms for feedback and improvement. If, as Francesca Rossi argues, "hard-coding" ethics into AI systems is to be precluded precisely because "these machines should adapt over time," are we comfortable with the guarantee that an AI doctor will make lapses in ethical judgment, and that these lapses may be surprising, and potentially difficult to detect, but that they might improve for future patients?[18]

As with traditional ethical frameworks that focus on relationships, consideration of the ethics of AI must also involve a deep examination of power. The role of *power asymmetries* can be seen in instances where AI systems are used to make judgments that have a material impact on vulnerable populations, such as the case of an AI-generated "terrorist threat score" that is currently being used to judge the "fitness" of individual refugees for entry into a particular country, or in the use of machine learning techniques in an attempt to predict recidivism.[19] Clearly, the AI systems (and their designers and those who deploy them) have considerable power in these scenarios, while their subjects are relatively powerless. These new cases can alert us to instances of Moor's "conceptual vacuums" -- places where the system's inner workings and imbalances are unavailable for examination or contestation. The challenges posed by these types of invisibly-produced machine-made decisions are not often discussed in current ethics frameworks. To date, much of the attention has remained on more easily schematized problematics and

[16] Ryan Calo, "The Case for a Federal Robotics Commission," Brookings Institute, 2014, 6-8, http://www.brookings.edu/research/reports2/2014/09/case-for-federal-robotics-commission. Long-standing research in human factors engineering has demonstrated that automation should be understood to create new kinds of human error as much as it eliminates some kinds of human error. For example: Laine Bainbridge, "Ironies of Automation," *Automatica* 19 (1983): 775-779; Raja Parasuraman and Victor Riley, "Humans and Automation: Use, Misuse, Disuse, Abuse, " *Human Factors* 39 no.2 (June 1997): 230-253.

[17] Alan Turing, "Computing Machinery and Intelligence," *Mind* 59, no. 236 (1950): 459.

[18] Francesca Rossi, "How do you teach a machine to be moral?" *The Washington Post,* November 5, 2015, https://www.washingtonpost.com/news/in-theory/wp/2015/11/05/how-do-you-teach-a-machine-to-be-moral/.

[19] See, for example: Kate Crawford, "Know your Terrorist Score," *Re:Publica 10*, Berlin, May 2, 2016, https://re-publica.de/en/file/republica-2016-kate-crawford-know-your-terrorist-credit-score; Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner, "Machine Bias," *Propublica,* May 23, 2016, https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

hypotheticals, such as the thought experiment of the trolley problem being applied to self-driving cars.[20]

The trolley problem, as it is frequently used, assumes that AI systems are a contained system, and have a complete mapping of all possible outcomes and relevant variables – something that can not be assumed even in the most "well-defined" tasks. Even for "mission critical" computer systems such as those used in aviation or space, it is not possible to comprehensively test a system for every potential input it may receive, and so researchers often rely on stopgap bug-detection schemes.[21] Moreover, it prevents further considerations of the values that shaped the creation of the system at hand - trolley problems focus more on how the risk will be spread. Instead, ethics should inform decisions throughout an AI system's development and deployment, continually assessing a system's impact situationally during the duration of its lifecycle. Based on this, dynamic ethical assessments are needed that go beyond the level of self-contained thought experiments, ones that contend with the real and complex impacts that AI systems are having on human populations.

2. Yes, and: Integrating AI into existing professional ethical frameworks

As AI systems become more integrated into professional environments, such as medicine, law and finance, new professional ethical dilemmas will arise. For instance, let's consider the case of a doctor caring for her patient. There are established ethical rules for conflicts of interest governing the conduct of human doctors.[22] These rules, for example, govern the prescription of drugs to patients in cases where a doctor has a vested interest in the drug manufacturer's success. These rules acknowledge that there are recognized risks that a doctor incentivized to do so may prescribe the drug when it is not in the best interests of the patient, and that it may cause her to underestimate certain risks associated with using the drug. Because ethical practice is a significant component of medical professionalization, the doctor is expected to recognize and avoid any conflict of interest, both because it intersects with her professional culture (including the Hippocratic Oath) and because it carries a particular legal liability.

In the case of a hypothetical AI medical advice system, which could easily draw on the data from a corpus of previous providers and various statistical models of biology and epidemiology, the AI system will necessarily reflect the bias of doctors, researchers, and other institutions from which the data was generated. If some doctors were biased concerning certain drugs or procedures where they had conflicts of interest, the ability of

[20] For an introduction to the trolley problem, see Wendall Wallach, "Moral Values and Constraints" panel for Moral Algorithms: the Ethics of Autonomous Vehicles conference, Ohio State University, 2016, http://livestream.com/WOSU/MoralAlgorithms/videos/120075247.

[21] See Donald MacKenzie, *Mechanizing Proof: Computing, Risk, and Trust,* Cambridge: MIT Press, 2004, 41-46.

[22] See, for example, the AMA Code of Medical Ethics: http://www.ama-assn.org/ama/pub/physician-resources/medical-ethics/code-medical-ethics.page

the AI system to rely on this data would also need to be subject to ethical considerations. For example, if this data was collected by a drug manufacturer, the ways in which the company categorizes the uses of the drug might implicitly devalue those risks that would prevent it from getting approved by the FDA. For the designers and patients of the AI provider, the situation is further complicated by the fact that the corpora of training data are in many cases privately held by the actors involved, meaning that the basis for a particular AI system's "judgment" would be extremely difficult, if not pragmatically impossible, to reconstruct.

As we've seen, AI has the potential to reenact and amplify existing power asymmetries. While most patients understand on some level the power asymmetry between themselves and their doctor, and many human research subjects might understand the asymmetry between themselves and the laboratory researcher experimenting on them, understanding the asymmetry between AI systems and those with whom these systems interact presents more complex challenges. The new risk is that AI systems will not only aggregate power by reducing the ability of the weakest to contest their treatment, but also redefine the grounds of what counts as 'ethical behavior', privileging the most powerful interests. Such power can take very subtle forms. For instance, we see that various kinds of AI, loosely defined, are used to influence or 'nudge' individual agents in particular directions largely dictated by those who design and deploy AI systems, occasionally putting individuals at risk.[23] As Illah Reza Nourbakhsh notes, "further empowerment of corporations can cause disempowerment in communities as new technologies asymmetrically and opaquely confer the power to shape information and manufacture desire."[24] Even the particular manner of consenting to a terms and conditions agreement form to use software is a 'nudge' that relies, in part, on what is known as the "subjective utility" of the software at a particular moment.[25]

3. The need for accountability: delegating decision-making to AI systems

Ethical frameworks often require the production of a record, for example, a medical chart, a lawyer's case file, or a researcher's IRB submission. They also provide mechanisms for redress by patients, clients, or subjects when these people feel they have been treated unethically. Contemporary AI systems often fall short of providing such records or mechanisms for redress. As Helen Nissenbaum has argued, there are four main barriers to the establishment of accountability, or answerability, in the development and use of computational technologies: "the problem of many hands," "the

[23] To offer a specific example: a 2009 change in Facebook's "privacy defaults" reversed the decline of "personal disclosure." See Ryan Calo, "Digital Market Manipulation," *The George Washington Law Review* 82, no. 4 (2014): 1013, footnote 103; see also Calo's discussion on "disclosure ratcheting" on 1012-1015. For an introduction to the ethics of "nudges," see Jason Borenstein and Ron Arkin, "Robotic Nudges: The Ethics of Engineering a More Socially Just Human Being," *Science and Engineering Ethics* 22 (2016): 31-46.

[24] Illah Reza Nourbakhsh, *Robot Futures*, Cambridge: MIT Press, 2013, 110.

[25] For a discussion of "subjective utility," the discomfort that "cognitive dissonance" can produce for an individual making privacy decisions, and the resulting decisions that can accompany such "dissonance," see Ian Kerr, Jennifer Barrigar, Jacquelyn Burkell, and Katie Black, "Soft Surveillance, Hard Consent," *Personally Yours* 6 (2006): 1-14.

problem of bugs," "blaming the computer" and "software ownership without liability."[26] Each of these barriers has implications not only for accountability during development and maintenance of AI systems, but also for the agency of the subjects of these systems. Data systems (including current AI technologies) used to make claims about particular groups or individuals are often unreviewable by those affected because they are considered the proprietary property of private companies or are under the purview of national security.[27]

AI systems tend to accentuate *information asymmetry*: individuals are not able to see into or understand the workings of an automated system and tend to have fewer opportunities to appeal algorithmic decisions performed by bureaucratic institutions.[28] As Zeynep Tufekci observes, "while algorithmic gatekeeping performs some traditional gatekeeping functions, it reverses or significantly modifies other key features of traditional gatekeeping with regard to visibility, information asymmetry, and the ability of the public to perceive the results of editorial work."[29]

Since these systems must be trained on existing data sets that are often kept private for reasons of commerce or security, it may be impossible to impose systems of accountability, such as the ability to cross-examine training data for implicit or explicit bias or to appeal these decisions. Often when an AI system is being developed to perform a particular task for which no algorithmically tractable data set readily exists, the data must be acquired by repurposing data originally collected for a different purpose (the translation inherent in this repurposing being itself something that should be subject to ethical standards and frameworks). Alternatively, if no data can be repurposed to train the AI system for a particular task, then that data must be generated, often at great expense to the system engineers. Among other things, the primacy of data in the construction of AI means that leading AI providers -- those that have the data and compute resources -- may have strategic advantages over new or alternative AI developers, advantages that may increase power asymmetries over time. It also means that there are few incentives to make such data open for use and scrutiny, at least from the perspective of market competition.

[26] Helen Nissenbaum, "Accountability in a Computerized Society," *Science and Engineering Ethics* 2, no. 1 (1996): 25-42.

[27] See the work of Danielle Citron on the ways that algorithmic and automated systems essentially undermine the mechanisms of due process that have been developed in administrative law over the course of the 20th century. Danielle Keats Citron, "Technological Due Process," *Washington University Law Review*, vol. 85 (2007): 1249-1313.

[28] See Virginia Eubanks' discussion of algorithms becoming "decision makers" in cases of policing practices and public assistance. Virginia Eubanks, "The Policy Machine," *Slate*, April 30, 2015, http://www.slate.com/articles/technology/future_tense/2015/04/the_dangers_of_letting_algorithms_enforce_policy.html. In the former case, individuals were subjected to heightened police scrutiny for being placed on the Chicago Police Department's "heat list". In the latter case, an individual with a disability was unable to appeal an algorithmic decision to deny her public assistance despite her attempts to place an in-person interview to advocate on her own behalf. See also Citron, "Technological Due Process;" Danielle Citron and Frank Pasquale, "The Scored Society: Due Process for Automated Predictions," *Washington Law Review* 89 (2014): 1-33.

[29] Zeynep Tufekci, "Algorithmic Harms Beyond Facebook and Google: Emergent Challenges of Computational Agency." *Journal on Telecommunications and High Technology Law* 13 (2015): 209.

Further, it is a difficult and risky process to reverse engineer AI systems to learn of any biases, let alone challenge their fairness. Even attempting this in the US can be a legal offense.[30] In contrast, Article 15 of the EU Data Protection Directive grants people the ability to opt out of "decisions" that are "based solely on automated processing of data intended to evaluate certain personal aspects."[31] In the US, there is no current mechanism to alert people that they have been assessed by a particular algorithm or AI system, and outside the realm of credit scoring, few formal grievance procedures to dispute the correctness of such judgments. AI systems that classify individuals and subject them to predictive assessments have commonly undergone no formal evaluation or verification beyond that of the original AI system engineers, who have little incentive to publicize its flaws. Furthermore, the accuracy of these judgments can rarely be disputed precisely because the systems are usually considered to be proprietary. Many of these issues arose in the context of "big data", and may apply in even greater force with AI systems that are applied to social institutions like education, employment, housing, and criminal justice.

As we've seen, the term ethics is used across a range of contexts, from professional responsibility and compliance, to research ethics, to thought experiments and hypotheticals, and perhaps most importantly, to assessing the everyday implications of AI systems for different human communities. It is this last area where we can see the greatest potential in analyzing and addressing the near-term impacts of AI systems: how do they affect marginalized populations? How do they shift or concentrate power? How do they interact with underlying structural inequality and injustice? These are the kinds of ethical questions that AI systems invoke, and a nuanced ethics of AI will need to be able to contend with them, and locate accountability mechanisms, as part of ensuring that AI serves the public interest.

Questions to consider

- What frameworks of disclosure, accountability, consent, and justice should we expect or impose on AI systems?
- Is there an ethical imperative to prevent AI systems from contributing to social inequality?
- Is merely removing biases from datasets enough (assuming such is possible), or should there be an active intervention into forms of long-standing structural inequality?

[30] See, for example, the restrictions to auditing systems and reverse engineering imposed by the Computer Fraud and Abuse Act. These provisions are currently being contested by the ACLU: https://www.aclu.org/blog/free-future/aclu-challenges-computer-crimes-law-thwarting-research-discrimination-online

[31] European Union, Directive 95/46/EC of European Parliament and of the Council on Protection of Individuals with Regard to the Processing of Personal Data on the Free Movement of Such Data, 1995. chap II, art 15, sec. 1., https://www.dataprotection.ie/docs/EU-Directive-95-46-EC-Chapter-2/93.htm.

- How should we evaluate if an AI system is serving the broader public good? In what ways might such evaluations be similar to or different from other social systems?
- How can we best meet the ethical challenges of training or supervising AI systems?
- What principles should guide an applied AI ethics framework? How will users be alerted to risks and benefits?
- What kinds of procedural due process rights should users have when they are subject to assessments by AI systems?
- How can we review algorithms and training data while also maintaining the proprietary information of private companies?
- Is it possible to create a code of ethics for AI systems? How should this code be incorporated and how can it account for long-term impacts?
- How should the professional codes for engineers and computer scientists reflect the new challenges of artificial intelligence?