The Social Implications of Artificial Intelligence Technologies in the Near-Term

AI NOW 2016 PRIMERS

These primers were distributed to attendees of the Experts Workshop as part of the AI Now 2016 Symposium. They were created to provide thorough overviews of the opportunities and issues posed by the introduction of AI across the event's four focus areas (Ethics, Healthcare, Labor and Inequality) and to spark discussion and solicit feedback during the workshop.

The AI Now 2016 Symposium was hosted by the White House and New York University's Information Law Institute on July 7th, 2016.

Co-Chairs

Kate Crawford (Microsoft Research & New York University)
Meredith Whittaker (Google Open Research)

Core Research Team

Madeleine Clare Elish (Columbia University)
Solon Barocas (Microsoft Research)
Aaron Plasek (Columbia University)
Kadija Ferryman (The New School)

Contents

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?

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

WORKSHOP PRIMER: HEALTHCARE & AI

Contents

Brief description

Artificial intelligence (AI) technologies represent a new frontier for healthcare.[1] AI in the health domain sits at the intersection of multiple fields (including computer science, the biological sciences, medicine and health policy) and implicates a diverse range of stakeholders (such as patients, clinicians, researchers, insurance providers, pharmaceutical manufacturers, biotech companies, and governments). Applications of machine learning, robotics, and related fields are generating the potential for significant medical insights and increasing the efficacy and efficiency of healthcare.

While there is much positive potential, the future impacts of AI in healthcare are subject to the same longstanding social and economic issues that have shaped both clinical care

[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 healthcare, 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.

and medical research.[2] These issues include increasing costs, differential health outcomes, and serious obstacles to patient access, among others. AI systems present opportunities to address many of these issues, but also have the potential to exacerbate old problems and to create new ones.

As AI systems change the norms and expectations of what constitutes care, the outcomes may be beneficial or may result in new harms. AI systems translate social and medical problems into technical solutions. This translation is far from neutral or without unintended consequences. What is lost, altered, or interposed in the process of translation? What feedback loops might be created? Who gets to decide which forms of care are most valuable or necessary?

In the US in particular, the introduction of AI systems is also subject to the substantial political and economic dynamics of health insurance, medical institutions, and for-profit healthcare companies like pharmaceutical or medical device manufacturers. How do these dynamics influence the design and regulation of new AI systems? How should professional ethics take these shifts into account? How can the complex range of incentives embedded into AI systems be made clear to patients and practitioners?

From expertise to data: AI histories & rapid advancements

The application of AI and "computer-aided diagnostics" more broadly has been an area of investigation in medicine since the 1960s. A motivating goal of these early applications was to use automated expert systems to assist physicians in making clinical decisions. These "expert systems" encoded decision trees and other forms of explicit knowledge, usually translated from interviews with domain experts.[3] The aim of integrating codified expertise into diagnostic processes was to standardize clinical decisions and minimize forms of clinical bias, including anchoring bias (relying on the first information received), availability bias (relating current information to a recent or memorable diagnosis), and premature closure (failing to consider alternatives).[4]

Fifty years hence, the need to mitigate clinical bias and ensure more accurate diagnoses remains a central goal of integrating AI into medicine and healthcare. In comparison to previous expert systems which relied on explicit clinical knowledge, the latest waves of AI

---

2 For example, significant scholarship has critically examined how dimensions of race are implicated in unequal health care, health outcomes, and medical research. See Dayna Bowen Matthew, *Just Medicine: A Cure for Racial Inequality in American Health Care,* New York: NYU Press, 2015; Brian Smedley, Adrienne Stith, and Alan Nelson, eds., *Unequal Treatment: Confronting Racial and Ethnic Disparities in Health Care,* Committee on Understanding and Eliminating Racial and Ethnic Disparities in Health Care, Board on Health Sciences Policy, Institute of Medicine, Washington, D.C.: National Academies Press, 2002; Alondra Nelson, *The Social Life of DNA: Race, Reparations, and Reconciliation After the Genome*, New York: Beacon Press, 2016.

3 For a critical perspective on the design and use of expert systems in the medical field, see Diana Forsythe, *Studying those who study us: An anthropologist in the world of artificial intelligence*, Stanford: Stanford University Press, 2001.

4 For a brief discussion of several kinds of biases, see Steven Dilsizian and Eliot Siegel, "Artificial Intelligence in Medicine and Cardiac Imaging: Harnessing Big Data and Advanced Computing to Provide Personalized Medical Diagnosis and Treatment," *Current Cardiology Reports* 16, no. 441 (2014): 1-8, https://www.ncbi.nlm.nih.gov/pubmed/24338557.

technologies take advantage of the affordances of big data sets, using complex statistical modeling and machine learning to generate insight from existing data. This is in part because the potential data sources have substantially increased with electronic health records (EHRs) and clinical and insurance databases, as well as patient-produced data from consumer devices and apps.[5] This wealth of data has positioned AI technologies as providing the potential to improve healthcare not only in the context of diagnostics but also to improve the production, organization, and communication of medical information.

Providers of medical care do not always have sufficient time to sift through, analyze, and apply this wealth of data, and may have little time to spend with patients to explain their decisions. AI is gaining traction in this new data-rich and often time-poor context of healthcare, particularly as a way to take advantage of the opportunities that accompany more fine-grained health data.[6]

AI systems rely on processing these large datasets to provide novel insights, and offer the opportunity to improve the accuracy of medical decision-making and reduce incorrect diagnoses. However, the primacy of data implicit in these paradigms may lead to unintended consequences. For instance, health providers may come under increasing pressure to prioritize statistical data over other forms of analysis and care. Or clinicians and care-workers may be expected to rely on AI-driven diagnoses and accept the underlying, often opaque models without being able to challenge the analysis or outcome.

In addition, as AI systems continue to be incorporated into healthcare, new technological norms are established, such as comprehensive tracking and surveillance of bodies and health data. The implications of such granular and ongoing data-tracking raise important questions regarding core health values such as confidentiality, autonomy, and informed consent. It also has serious implications for how we understand and value health labor, such as the difference between those who perform research, diagnostic work, and care work.

---

5 Electronic Health Records (EHRs) are quickly becoming a part of dynamic digital systems that hold multiple forms of patient information, from providers' notes and clinical assessments to lab test results and medication prescriptions. EHRs can also include non-patient specific information such as clinical guidelines, medical research literature, and scientific news. For an example of prediction models based on such data see N. Razavian, S. Blecker, A.M. Schmidt, A. Smith-McLallen, S. Nigam, D. Sontag, "Population-Level Prediction of Type 2 Diabetes using Claims Data and Analysis of Risk Factors" *Big Data: Data and Healthcare Special Issue* (January 2016), http://online.liebertpub.com/doi/pdf/10.1089/big.2015.0020.

6 See for instance Deep Mind's collaboration with the Royal Free Hospital London on Streams, an app to assist in the detection of acute kidney injury. Deep Mind Health, "Supporting direct patient care with clinical apps," DeepMind, 2016, https://deepmind.com/health/clinical-apps.

Five challenges for healthcare & AI

1. How will AI impact production of new health research?

The use of AI in medical research is in its infancy, but the insights drawn from AI and big data are already being heralded as a potential revolution in the field of medical research. One of the most high-profile successes has been the development of a new drug aimed at pancreatic cancer, BPM 31510. The drug, which has been described as the first drug to be developed through AI, was produced by BERG utilizing AI to process and analyze vast amounts of data, including biological data as well as clinical and patient data.[7] Based on these analyses, BERG was able to generate the first complete model of how pancreatic cancer functions, and in turn focused on developing a drug that could prohibit the cancerous cells from metabolizing energy and growing.[8] Phase I trials, which began in 2013, have shown positive results for the drug's safety and efficacy.[9]

The vast amounts of data that can be analyzed by AI systems are transforming not only how and how much biological or clinical data can be processed, but also how medical research itself can be conducted.[10] Projects such as the Allen Institute's "Semantic Scholar" allow researchers to search and locate relevant research through AI-driven insights.[11] One of the potentially transformative aspects of IBM Watson's integration into clinical decision-making is the capacity for Watson to access, process, and analyze thousands of new and existing studies that clinicians are unlikely to have read, rendering clinical decision-making more informed by the latest research.[12] However, the use of proprietary systems like Watson may have implications for the openness of medical research, and the ways in which new research can be evaluated and peer-reviewed.

The integration of AI into medical research holds out exciting prospects for developing new treatments and offers the possibility of tailoring medicine to specific individuals. Still, this research is subject to the existing limitations and biases that may affect research outcomes, including incomplete datasets that exclude certain minority populations and financial incentives that favor the development of certain drugs over others. Medical research data often presents itself as objective and universal, while in reality the findings may be partial, temporary, or specific to only some communities.[13] As AI systems learn

---

7 Liz Harley, "Taking on cancer with AI: Niven Narain, BERG," *Front Line Genomics*, April 22, 2016, http://www.frontlinegenomics.com/interview/3983/3983/.

8 Laura Lorenzetti, "This Company Wants to Cure Pancreatic Cancer Using AI," *Fortune*, April 22, 2016, http://fortune.com/2016/04/22/berg-pacreatic-cancer-artificial-intelligence/.

9 Ibid.

10 See for instance, Scott Spangler et al., "Automated Hypothesis Generation Based on Mining Scientific Literature," 20th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, August 24th - 27th, 2014, http://dl.acm.org/citation.cfm?id=2623667.

11 "Semantic Scholar," Allen Institute for Artificial Intelligence, http://allenai.org/semantic-scholar/.

12 Carl Zimmer, "Enlisting a Computer to Battle Cancers, One by One," *New York Times,* March 27, 2014, http://www.nytimes.com/2014/03/27/science/enlisting-a-computer-to-battle-cancers-one-by-one.html?_r=0

13 For example, research recently published in the journal *PNAS* has raised attention to issues with inferential statistical models and software bugs that likely invalidate many thousands of research papers within a particular field of fMRI-based

and contribute to medical research, awareness of these normative and structural limitations are needed to appropriately interpret the "new knowledge" that is generated.

2. How will AI impact diagnostics and healthcare delivery?

Automated and artificial intelligence systems present an opportunity to mitigate or even prevent misdiagnoses. This would be a substantial contribution to the field of healthcare; the harm of incorrect diagnoses are well-known and significant, including patient morbidity and mortality, increased length of hospital stay, unnecessary testing, and increased health care and malpractice costs.[14] For instance, AI-based approaches have demonstrated great potential to reduce the harms and costs associated with hospital readmissions for heart failure.[15]

AI systems will also have limitations. When AI systems are presented as expert authorities understood to be free from error or bias, they may be subject to less scrutiny. [16] But opportunities to introduce new errors via AI are myriad, including data entry and other points of translation between physical and digital records. This is especially relevant for the transition to electronic health records (EHRs), upon which many AI health systems are based. While these systems may introduce or extend existing incorrect models, the use of big data sets also has the potential to mitigate the biases that may be inherent in the existing constitution of randomized control trials (RCTs) or other public health databases.[17]

Despite a desire to make healthcare accessible and affordable to all, substantial evidence shows that access to healthcare and health outcomes are unequally distributed, with poor, non-white, and female populations often systematically disadvantaged.[18] The introduction of AI will not automatically correct these systemic inequalities, and has the potential to amplify them.

The history of researching and treating heart disease is a case in point. Over the last decade, medical institutions such as the American Heart Association have acknowledged

---

research. Neuroskeptic blog, "False-Positive fMRI Hits The Mainstream," *Discover Magazine*, July 7, 2016, http://blogs.discovermagazine.com/neuroskeptic/2016/07/07/false-positive-fmri-mainstream/.

14 Bradford Winters et. al., "Diagnostic errors in the intensive care unit: a systematic review of autopsy studies," *BMJ Quality & Safety* vol 21 iss 11 (2012) http://qualitysafety.bmj.com/content/21/11/894.abstract?sid=647c745e-4b1f-45e4-8883-a700adfc0005.

15 Mohsen Bayati, Mark Braverman, Michael Gillam, Karen M. Mack, George Ruiz, Mark S. Smith and Eric Horvitz, Data-Driven "Decisions for Reducing Readmissions for Heart Failure: General Methodology and Case Study," *PloS One* vol 9 no. 10 (2014), http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0109264.

16 Nicholas Carr, "Automation Makes Us Dumb," *Wall Street Journal*, Nov 21, 2014 http://www.wsj.com/articles/automation-makes-us-dumb-1416589342.

17 For an overview of the potential limitations of data generated from RCTs see Peter M Rothwell, "Factors That Can Affect the External Validity of Randomised Controlled Trials," *PLoS Clin Trials* vol 1 iss 1 (May 2006), http://journals.plos.org/plosclinicaltrials/article?id=10.1371/journal.pctr.0010009.

18 Kaiser Family Foundation, "Health Coverage by Race and Ethnicity: The Potential Impact of the Affordable Care Act," Report, Washington, D.C., 2013, http://kaiserfamilyfoundation.files.wordpress.com/2013/04/8423.pdf; John Ayanian, "The Costs of Racial Disparities in Health Care," *Harvard Business Review*, October 1, 2015, https://hbr.org/2015/10/the-costs-of-racial-disparities-in-health-care.

that "most heart disease research is done in men, so how we categorize it is based on men. We need more science in women."[19] Numerous studies have provided evidence that standards around heart disease, from education to diagnosis to medication, have been based on male bodies to the detriment of the diagnosis and treatment of heart disease in women.[20] If "normal" or "standard" are equated with only a subset of people, then those who fall outside this subset may not benefit from treatments based on those standards. Again, AI systems have the potential to enable appropriately specialized care.[21] Still, as machine learning gains a more central role in establishing and identifying what illness is, additional attention must be paid to the faulty assumptions that may underlie constructions of "normal" or "average" health, constructions that are no doubt reflected in the data used to train AI systems.

Many AI systems currently being developed are intended to operate in the context of international health. Devices like Peek (Portable Eye Examination Kit), currently being tested in countries such as Kenya, create opportunities to access comprehensive eye exams without a doctor being physically present, expanding access to healthcare.[22] The potential benefits of these systems are immense, but this does not mean that their risks can be overlooked. For instance, in the case of Peek or other remote and AI-supported diagnostics, how can the integration of these apps into existing healthcare contexts reinforce rather than inhibit local capacity-building in medicine?

3. How will AI impact the relationship between patients and healthcare providers?

The existing and potential applications of AI technologies have profound implications for the constitution of care-work and what it means to care for sick or vulnerable bodies. AI systems may be placed as mediators of care work, or may be tasked with supplanting caregivers entirely. In this way, AI systems may change the relationships and norms between patients and doctors or other caregivers.

Common examples used to illustrate the potentials of AI technologies to replace or supplement human caregivers range from robotic surgery to virtual avatars to companion robots.[23] Such examples have catalyzed important debates about the social implications of delegating care and companionship to non-human agents.[24] What kinds of equivalences are being made when machines replace, not just augment, human

---

19 EurekaAlert, "American Heart Association makes first ever statement on female heart attacks," American Association for the Advancement of Science, Janurary 25, 2016, http://www.eurekalert.org/pub_releases/2016-01/m-aha012516.php.

20 Vidhi Doshi, "Why Doctors Still Misunderstand Heart Disease in Women," *The Atlantic Monthly*, http://www.theatlantic.com/health/archive/2015/10/heart-disease-women/412495/; Richard J McMurray et al. "Gender disparities in clinical decision making." *JAMA* 266 no 4 (1991): 559-562.

21 For example see Steven Dilsizian and Eliot Siegel, "Artificial Intelligence in Medicine and Cardiac Imaging: Harnessing Big Data and Advanced Computing to Provide Personalized Medical Diagnosis and Treatment," *Current Cardiology Reports* 16, no. 441 (2014): 1-8, https://www.ncbi.nlm.nih.gov/pubmed/24338557.

22 Peek, "Peek Vision: What It does," http://www.peekvision.org/what-it-does.

23 Laurel Riek, "Robotics Technology in Mental Healthcare," in D. Luxton (Ed.), *Artificial Intelligence in Behavioral Health and Mental Health Care*. New York: Elsevier, 2015: 185-203.

24 Nick Bilton, "Disruptions: Helper Robots Are Steered, Tentatively, to Care for the Aging," *New York Times* May 19, 2013. http://bits.blogs.nytimes.com/2013/05/19/disruptions-helper-robots-are-steered-tentatively-to-elder-care/

professionals? When we deem a machine capable of "care," what capacities are we assuming, and what definition of "care" are we using? Are these equivalences in the best interests of patients?

An example that highlights the potential risks and benefits of employing AI-based technologies to interface directly with patients is a system like AICure. AICure uses facial recognition, motion detection, and AI analytics to monitor patients for drug adherence. The app monitors participants in clinical trials and those in clinical care to ensure that they are taking their medication at the right times and in the appropriate amounts. Patients using AICure administer their pill or medication in front of the camera on a their smartphone.[25] The app records this information, collecting and disseminating the data to doctors, caregivers, and/or researchers. An interface available to those at the receiving end enables close patient tracking, and the ability to contact a patient if adherence is not being followed.

This technology has the potential to improve health outcomes by encouraging better medication adherence, to improve the accuracy of clinical trials, and to reduce pharmaceutical companies' revenue losses due to non-adherence. The attractiveness of AICure to large medical institutions and research hospitals is easy to see. Standardized data collection, and "proof" of adherence can be gathered much more concretely and cheaply, since there is no need for a human to witness and verify adherence. The time and skill it takes to develop trust between a patient and doctor or to monitor patients within the context of a human relationship is no longer required.

Still, the potential short-term and long-term downsides to this technology must be considered. What kinds of patients might not adapt well to app and smartphone-based systems? How might such an app exacerbate clinical stereotyping and the labeling of certain individuals and populations as "treatment resistant" or "non-compliant?" In turn, how might such technologies and the surveillance and tracking they enable facilitate unfair discrimination against particular populations? What would be the impact if patients receive adherence scores, similar to FICO scores, that would presume to reflect their worthiness to participate in research studies or receive medical care?[26] An app like AICure may inadvertently obscure the range of factors that contribute to a lack of adherence. In presenting a technological solution to one aspect of the treatment adherence problem, other social and cultural aspects, including environmental, economic, or psychological factors, may be overlooked.

AI systems are also being integrated into the modes of communication and courses of action available to patients. The field of health communication is shifting from a public health frame to a personalized frame, with information tailored, through the analysis of

---

[25] AiCure, "AiCure: Advanced Medication Adherence Solutions," https://www.aicure.com/.

[26] Tara Parker-Pope, "Keeping score on how you take your medicine," *New York Times* June 20, 2011, http://well.blogs.nytimes.com/2011/06/20/keeping-score-on-how-you-take-your-medicine/.

big data, to individuals using social media or devoted devices, including health coach avatars. For instance, ChronologyMD, a pilot application, allows patients with Crohn's disease to track and record their symptoms daily and also provides relevant reminders for sleep, medication, appointments and other forms of self-care.[27] Patients have reported better management of their disease resulting in improved quality of life, and providers were able to administer more comprehensive care because patients were more prepared with detailed information during appointments.

This application, and others like it, are marketed as tools of self-empowerment, and for some, they are.[28] However, additional implications need to be taken into account. Will the development of such applications effectively shift the responsibility for care and monitoring from healthcare professionals to patients themselves? What kinds of patients are favored in this new dynamic, and how do patients not well-equipped to manage and maintain their own data fall through the cracks? Moreover, how do such apps disrupt the locus of clinical expertise? What new roles and responsibilities do the designers and developers of such apps take on, and how do the ethical responsibilities at the heart of the medical profession get integrated into these differing design and engineering contexts?

4. How will AI interact with the economics of healthcare?

The economics of healthcare are already subject to increasing pressures from government actors, insurance companies, health institutions, pharmaceutical companies, medical device manufacturers, employers, and many others. With ongoing efforts to reduce costs and at the same time promote growth, many of these stakeholders have turned to large-scale data collection and now to AI to help sustain their economic models of research and care.[29] Many such efforts are focused on laudable goals, such as enabling more accurate clinical decisions or faster analysis of clinical data. AI systems also provide the possibility to improve preventative care and decrease costs. Still, the resources to develop and maintain these information technology systems within hospitals remain unstable and unequally distributed.[30] Moreover, because economic expectations for AI will depend heavily on making an increasing amount of patient data available for mining and learning across various devices, platforms, and networks, the implications for "surveillance capitalism" to transform fundamental structures within the health economy

---

27 Linda Neuhauser, Gary L. Kreps, Kathleen Morrison, Marcos Athanasoulis, Nikolai Kirienko, Deryk Van Brunt, "Using Design Science and Artificial Intelligence to Improve Health Communication: ChronologyMD Case Example," *Patient Education and Counseling* 92 (2013): 211-217, https://www.ncbi.nlm.nih.gov/pubmed/23726219.

28 For scholarly work on the range of self-quantification and health tracking systems, see Dawn Nafus (ed.), *Quantified: Biosensing Technologies in Everyday Life*, Cambridge, MA: MIT Press, 2016.

29 Claudia Grossmann et al., *Clinical Data as the Basic Staple of Health Learning: Creating and Protecting a Public Good*, Workshop Summary, Institute of Medicine of the National Academies, Washington DC: National Academies Press, 2010, http://www.ncbi.nlm.nih.gov/books/NBK54296/.

30 For instance, see American Hospital Association, *Adopting Technological Innovation in Hospitals: Who Pays and Who Benefits*, Report. American Hospital Association, Washington, DC., 2006, http://www.aha.org/content/2006/pdf/061031-adoptinghit.pdf.

are profound.[31] The economic emphasis on surveillance and consumption of sensitive health data will only increase with the recent moves to promote evidence-based medicine and the Affordable Care Act's (ACA) shift from a fee-for-service to a pay-for-performance model.[32]

Insurers may also feel increased pressure to justify cross-subsidization models in the context of AI systems. Despite prohibitions in the Genetic Information Nondiscrimination Act of 2008, there is already growing interest in using genetic risk information for insurance stratification.[33] In fact, differential pricing has become one of the standard practices for data analytics vendors, introducing new avenues to perpetuate inequality.[34]

Such efforts raise significant ethical questions. For example, how should AI systems that utilize this data take up questions of ethics, privacy, and potential discrimination in health care outcomes? What happens when one's insurer not only seeks to review your EHRs but also to query your FitBit and health agent AI about your status before setting the price of your premium? How will these dynamics influence the design and regulation of new AI systems?

5. How will AI impact professional ethics for health providers?

The use of AI in health contexts can also present challenges to core values upheld within the ethical codes of medical professionals, such as confidentiality, dignity, continuity of care, avoiding conflicts of interest, and informed consent.[35] Patient privacy and confidentiality of care has emerged as a primary concern in the adoption of new medical technologies.[36] As noted above, the shift to ubiquitous tracking and self-surveillance through "smart" devices and apps may often push the limits of existing privacy protections, such as the Health Insurance Portability and Accountability Act (HIPAA).[37] For example, computer scientist Latanya Sweeney has demonstrated that the vast majority of Americans can be identified using only three pieces of perceived "anonymized data:" ZIP code, birthdate, and sex.[38] Risks that patients will be reidentified from granular-level

---

31 Shoshana Zuboff, 'Big other: surveillance capitalism and the prospects of an information civilization', *Journal of Information Technology*, Vol. 30, Issue 1 (March 2015): 75-89, http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2594754.

32 David Blumenthal, Melinda Abrams, and Rachel Nuzum, "The Affordable Care Act at 5 Years." *New England Journal of Medicine* Vol. 372, Issue 25, (2015), 2453, http://www.nejm.org/doi/full/10.1056/NEJMhpr1503614#t=article.

33 Yann Joly et al., "Life Insurance: Genomic Stratification and Risk Classification." *European Journal of Human Genetics* 22 No. 5 (May 2014): 575–79, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3992580/.

34 Jason Furman and Tim Simcoe, "The Economics of Big Data and Differential Pricing," The White House Blog, February 6, 2015, https://www.whitehouse.gov/blog/2015/02/06/economics-big-data-and-differential-pricing.

35 Amercian Medical Association, Code of Medical Ethics, "Opinion 10.01 - Fundamental Elements of the Patient-Physician Relationship", http://www.ama-assn.org/ama/pub/physician-resources/medical-ethics/code-medical-ethics/opinion1001.page.

36 Helen Nissenbaum and Heather Patterson, "Biosensing in Context: Health Privacy in a Connected World," in Dawn Nafus (ed.), *Quantified: Biosensing Technologies in Everyday Life*, Cambridge, MA: MIT Press, 2016.

37 Paul Ohm, "Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization," *UCLA Law Review* 57 (2010): 1701-1778, http://ssrn.com/abstract=1450006.

38 Latanya Sweeney, "Simple Demographics Often Identify People Uniquely," Carnegie Mellon University, Data Privacy Working Paper 3, Pittsburgh: PA, 2000. See also, Bradley Malin, Latanya Sweeney and Elaine Newton, "Trail

data or have their identities, illnesses, or other health information predicted through proxy data will only increase as AI systems become integrated into health and consumer products.[39]

Moreover, conflict of interest issues may arise in new and unanticipated ways if health AI providers, developers, and vendors fail to disclose all the relevant risks to patients posed by systems in which they have a financial stake. Even continuity of care might become an issue if the AI systems behind decision-support tools are kept proprietary to AI companies, and health providers are unable to share or transfer patient information or predictive analytics in the same way that Electronic Health Records (EHRs) are guaranteed to be portable with the patient.

Particularly thorny ethical issues may arise as AI emerges in "smart" medical devices. Currently, there are efforts to incorporate AI systems across multiple devices, from pacemakers to blood pressure and glucose monitors to activity trackers to multimodal bed sensors for the elderly to "smart" insulin pumps.[40] These AI applications promise many benefits including better remote patient monitoring.

However, the software that powers these devices is often proprietary, rather than open source (open to external scrutiny and auditing).[41] A recently granted exemption to the Digital Millennium Copyright Act (DMCA) provides the opportunity to examine code for external medical devices; internal medical devices, however, are equally if not more important to examine.[42] Experts have warned of grave security issues in the deployment of networked technologies across Internet of Things (IoT) devices, many focusing on medical devices as specifically problematic. A team at the University of South Alabama demonstrated these issues in 2015, successfully hacking the iStan brand networked pacemaker and "killing" a mannequin that had the pacemaker installed.[43]

Re-identification: Learning Who You are From Where You Have Been," LIDAP-WP12. Carnegie Mellon University, Laboratory for International Data Privacy, Pittsburgh, PA, March 2003, http://dataprivacylab.org/dataprivacy/projects/trails/index3.html. Latanya Sweeney, "Matching Known Patients to Health Records in Washington State Data," Harvard University. Data Privacy Lab, White Paper 1089-1, Cambridge, MA, June 2013, http://dataprivacylab.org/projects/wa/index.html.

39 Nicolas Terry, "Big Data Proxies and Health Privacy Exceptionalism," *Health Matrix* 24 (2014): 65-108, http://ssrn.com/abstract=2320088. For how search data can predict health issues, see: Ryen W. White and Eric Horvitz, "Predicting escalations of medical queries based on web page structure and content," in *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, ACM, 2010, 769-770. http://research.microsoft.com/en-us/um/people/horvitz/SIGIR_2010_page_structure.pdf

40 G. P Moustris., S. C. Hiridis, K. M. Deliparaschos, and K. M. Konstantinidis, "Evolution of Autonomous and Semi-Autonomous Robotic Surgical Systems: A Review of the Literature." *The International Journal of Medical Robotics and Computer Assisted Surgery* 7, no. 4 (2011): 375–92, http://onlinelibrary.wiley.com/doi/10.1002/rcs.408/abstract.

41 For a discussion of the implications of such proprietary systems, see Frank Pasquale, *The Black Box Society: The secret algorithms that control money and information*, Cambridge: Harvard University Press, 2015.

42 Karen Sandler, Lysandra Ohrstrom, Laura Moy and Robert McVay, "Killed by Code: Software Transparency in Implantable Medical Devices," Software Freedom Law Center, New York, 2010, http://www.softwarefreedom.org/resources/2010/transparent-medical-devices.html.

43 William Bradley Glisson, Todd Andel, Todd McDonald, Mike Jacobs, Matt Campbell, and Johnny Mayr, "Compromising a Medical Mannequin," *arXiv preprint, arXiv:1509.00065,* 2015, https://arxiv.org/pdf/1509.00065v1.pdf

Compounding the issue, regulatory agencies such as the FDA struggle to maintain sufficient expertise and resources to audit such devices for biases or inaccuracies.[44] With the addition of AI systems, such challenges will multiply in both complexity and cost. In particular, because AI tends to generate many of its decisions dynamically and in emergent patterns, even access to source code may not be sufficient to assess its impact on any particular medical device. Thus, the right of patients to be informed about the benefits and risks of these devices becomes a significant challenge in a world of AI-enabled medical technologies. Determining the dimensions of transparency and accountability required for effective regulation will be a crucial area of debate. How could patients and practitioners audit or contest AI-based clinical decisions? What responsibility do those who provide AI health products and services have to be accountable for health outcomes?

Questions to consider

- What are the goals of AI-based systems in healthcare? Given the range of customers for such systems (healthcare providers, governments, and insurers), what different outcomes are they optimizing for, and are these outcomes in the best interest of patients?
- What does the relative paucity of data for certain demographics (women, people of color) mean for the accuracy of AI diagnostic and decision-support technologies trained on available data? How will this impact existing health disparities?
- Data sharing and standardization is key for the expansion of AI in health. In June, 2016, the US Secretary for Health and Human Services announced a federal initiative aimed at facilitating health data sharing among industry leaders. What challenges are raised by these new sharing agreements? Are current data privacy and ethics frameworks built to handle sharing for purposes of AI?
- Who gets to decide which forms of care are prioritized in a world of AI diagnostics? What rights might doctors, patients, and other caregivers have to contest and alter decisions made by AI systems? How might such decisions be audited and understood by patients and doctors?
- How can patient confidentiality, autonomy, and informed consent be maintained in the face of the increasingly granular tracking required for precision medicine and its attendant AI systems?
- How does AI impact the relationship between patients, doctors, and other caregivers? What does it mean for a machine to “care”, and what impact will mechanized care work have on patients? What is the role of affective labor, empathy, and human connection in a world in which “care” is increasingly automated?

[44] Gregory Curfman and Rita Redberg, “Medical Devices — Balancing Regulation and Innovation,” *New England Journal of Medicine* 365, no. 11 (September 15, 2011): 975–77, http://www.nejm.org/doi/full/10.1056/NEJMp1109094#t=article.

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

WORKSHOP PRIMER: INEQUALITY & AI

Contents

Brief Topic Description

Artificial Intelligence (AI) promises to reshape the contours of inequality in society in at least five significant ways.[1]

1. AI is already creating new ways to generate economic value while also affecting how this value is distributed. *If the ability to develop and implement AI systems is not available equitably, certain parts of the population may claim the lion's share of the economic surplus.* In this scenario, those who have data and the computational power to use AI will be able to gain insight and market advantage, creating a "rich-get-richer" feedback loop in which the successful are rewarded with more success.[2]

2. AI may help to expose and counteract the prejudices and biases that undergird persistent social inequalities along the lines of race, gender and other characteristics.

[1] As Russell and Norvig point out, the history of artificial intelligence has not produced a clear definition of AI but 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." Here we rely 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] Donella H. Meadows and Diana Wright, *Thinking in Systems: A Primer*, White River Junction, VT: Chelsea Green Publishing, 2008.

Alternatively, AI systems may inherit, even amplify, many of these same prejudices and biases. AI depends on the data it is given and may reflect the characteristics of such data, including any biases.

3. There are two ways in which data can be biased: one is because the data available is not an accurate reflection of reality; the other is because the underlying process itself exhibits long-standing structural inequality. The former type of bias can sometimes be addressed by 'cleaning the data' or improving the data collection process; the latter requires interventions with complex political ramifications.

4. With more data, analytics, and automation, the capacity to draw more fine-grained distinctions between people will increase, and this may substantially increase profits for industries like insurance. This has the potential to affect how risk is pooled in society. AI may endanger the solidarity upon which insurance and other collective risk mitigations strategies rely.

5. Finally, women and minorities continue to be under-represented in the field of AI. A lack of inclusion and "like me" bias may limit the degree to which practitioners choose to recognize or prioritize the concerns of other communities.

AI redistributes wealth, but to whom?

AI has the potential to generate new economic value by creating and expanding products and services, as well as by reducing costs through automation and other approaches. How this economic value will be distributed? This is a key concern for economic inequality.

If the vast majority of economic value is distributed to those already among the wealthiest today, then the expansion of AI technologies could widen existing wealth and income disparities. In such a scenario, AI would exacerbate inequality.

At the same time, if automation reduces the costs of creating certain goods and services (and these savings are passed on to consumers), AI could narrow the gap between the haves and have-nots (at least in terms of access to these goods and services).[3] In this case, AI would increase the standard of living overall, and could even have a progressive redistributive effect.

AI could also give rise to entirely new ways of making a living, either by allowing people whose jobs have been eliminated to seek other ways of obtaining resources, or by creating new jobs that ensure affected workers an easy transition to other paid labor. Rather than simply replacing workers or reducing workers' share of the profits from productive activity, AI could free people to pursue new forms of living and work, increasing overall welfare in society.

[3] Dean Baker, "Can Productivity Growth Be Used to Reduce Working Time and Improve the Standard of Living of the 99 Percent? The Future of Work in the 21st Century," Economic Analysis and Research Network, 2014, http://www.earncentral.org/Future_of_work/Baker%20Shorter%20Work%20Time%20Final.pd.

Of course, as some commentators have observed, AI could also render certain workers' skills redundant, leaving those who have been replaced by automation with few options for alternative paid employment.[4] Even if workers can find new employment, these jobs may be of lower quality, may pay less, and may provide less stability and security than the jobs eliminated by AI.[5] This is why understanding AI's potential impact on human employment is an important aspect of understanding its impact on economic equality.

Further, if learning new skills is prohibitively expensive, workers may find it impossible to pivot to a new profession. Under these circumstances, not only would AI increase inequality, it might push certain parts of the workforce into permanent unemployment and poverty.

Such concerns have driven growing interest in social safety net programs, including universal basic income (UBI). UBI is a redistribution framework that would provide all members of society with the minimum amount of money to cover life's necessities.[6] UBI's supporters see it as a way to address whatever ongoing unemployment AI might cause. Unlike need-based welfare programs, UBI would divert the same amount of money to all members of society. This eliminates, according to proponents, the complex and costly bureaucracy necessary to administer programs targeted at the poor.[7] Crucially, UBI would aim to provide an income floor beyond which no one would fall. But it would not aim to ensure a more equitable distribution of income in society overall. This means that inequality could still increase, even in a society that has adopted UBI. Some critics discount UBI as a way of simply making growing inequality more palatable by keeping the unemployed out of poverty, and potentially providing corporations with an effective subsidy that allows them to compensate human workers less, and reap more of the profits.[8]

The uneven availability of AI technologies may increase the developers or adopters' bargaining power, providing them a "crystal ball" of analytical insight unavailable to others.

For example, if AI techniques allow one party in an exchange to effectively infer the other party's sensitivity to price (say, the knowledge that the other party is unlikely to bargain for a lower price, or, will require a 20% drop in price before being willing to sign a

---

4 Erik Brynjolfsson and Andrew McAfee,*The Second Machine Age: Work, Progress and Prosperity in a Time of Brilliant Technologies*, New York: W.W. Norton & Company, 2014. For more context, see the Labor & AI primer from *AI Now*, https://artificialintelligencenow.com/

5 Henry Siu and Nir Jaimovich, "Jobless Recoveries," Third Way (April 2015), https://s3.amazonaws.com/content.thirdway.org/publishing/attachments/files/000/000/862/NEXT_-_Jobless_Recoveries.pdf?1428093868; Roosevelt Institute, "Technology and the Future of Work: The State of the Debate," Open Society Foundations Future of Work Project (April 2015), https://www.opensocietyfoundations.org/publications/technology-and-future-work-state-debate.

6 See, for example, Nick Srnicek and Alex Williams, *Inventing the future: Postcapitalism and a World without Work*, New York: Verso Books, 2015; but also Sam Altman, "Basic Income," *Y Combinator Posthaven*, January 27, 2016, https://blog.ycombinator.com/basic-income.

7 Eduardo Porter, "A Universal Basic Income Is a Poor Tool to Fight Poverty," *The New York Times*, May 31, 2016, http://www.nytimes.com/2016/06/01/business/economy/universal-basic-income-poverty.html.

8 Jathan Sadowski, "Why Silicon Valley is embracing universal basic income," *The Guardian*, June 22, 2016, https://www.theguardian.com/technology/2016/jun/22/silicon-valley-universal-basic-income-y-combinator.

contract), the party with access to AI will likely be at an advantage.[9] Likewise, benefits may go to the party that can rely on AI to determine the other party's susceptibility to different modes of persuasion (say, the knowledge that the other party will likely buy a product if they are 'primed' with a story about puppies).[10] At the extreme, these techniques may lead to predatory practices.[11]

And yet, if deployed more equitably, AI could also address power dynamics in markets that already exhibit such information asymmetries, and it could enable people to make more considered choices and counteract tendencies to make poor, irrational, or impulsive decisions.[12] The means of providing such access, however, remains an open question.

AI: reproducing or correcting human bias?

Automated decisions will increasingly shape people's life chances.

As AI takes on a more important role in high-stakes decision-making, it will begin to affect who gets offered crucial opportunities, and who is left behind—from offers of credit and insurance, to the availability of job opportunities and parole. Some hope that AI will help to overcome the biases that plague human decision-making;[13] others fear that AI will amplify such biases, denying opportunities to the deserving and subjecting the deprived to further disadvantage.[14]

Either way, the underlying data will play a crucial role and deserves keen attention. This is because recent advances in AI have happened almost entirely in the subfield of Machine Learning (ML), where computers learn how to perform a task by finding patterns in a large dataset of examples—examples often taken from human activities in similar domains.

There is the hope that by giving AI "all of the data," and allowing it to detect complex and subtle patterns that escape human recognition, it could illuminate and help to tear down artificial barriers that have contributed to social inequality along the lines of race, gender, and age, among other characteristics.

At the same time, there is the risk that if the data used reflects these biases, AI trained on this data will replicate and magnify those biases. In such cases, AI would exacerbate discriminatory dynamics that create social inequality, and would likely do so in ways that

---

9 Council of Economic Advisers, *Big Data and Differential Pricing*, February 2015, https://www.whitehouse.gov/sites/default/files/docs/Big_Data_Report_Nonembargo_v2.pdf.

10 Ryan Calo, "Digital Market Manipulation," *George Washington Law Review* 82, no. 4 (October 3, 2014): 995-1051.

11 Upturn, *Led Astray: Online Lead Generation and Payday Loans*, October 2015, https://www.teamupturn.com/reports/2015/led-astray.

12 Richard T Ford, "Save the Robots: Cyber Profiling and Your So-Called Life," *Stanford Law Review* 52, no. 5 (2000): 1578-1579.

13 Future of Privacy Forum and Anti-Defamation League, *Big Data: A Tool for Fighting Discrimination and Empowering Groups*, September 11, 2014, https://fpf.org/wp-content/uploads/Big-Data-A-Tool-for-Fighting-Discrimination-and-Empowering-Groups-Report1.pdf.

14 Cathy O'Neil, *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*, New York: Crown, 2016.

would be less obvious than human prejudice and implicit bias (and that could be justified as the work of "intelligent technology," and thus potentially perceived as more neutral, and more difficult to contest).[15] AI presents the possibility of both removing and replicating the bias in decisions that materially shape people's lives.

Such concerns were previously recognized in the 2014 White House reports on big data[16] and have become a major issue for the civil rights advocacy community.[17] These awareness-raising efforts have pushed the issue of discrimination to the center of current debates about big data, and the concerns are amplified in the context of AI.[18] The White House recently issued a follow-on report specifically on the topic of "Algorithmic Systems, Opportunity, and Civil Rights," examining the risks, benefits, and challenges of using data in automated decision-making for credit, employment, education, and criminal justice.[19]

While these reports emphasize that AI may help to advance civil rights and diversity, they also acknowledge that simply reconfiguring decision-making to rely more on AI and less on ad hoc human judgment may not eliminate human bias from the decision-making process or radically transform the composition of key institutions. AI may, potentially, even amplify such bias or undermine efforts to increase diversity. There are many reasons for this caution. AI requires data and, as Solon Barocas and Andrew Selbst point out, "data is frequently imperfect in ways that allow machines trained on it to inherit the prejudices of prior decision-makers or reflect the widespread biases that persist in society at large."[20] Other researchers have documented cases of bias in the wild in such varied applications as online advertising for both criminal background checks[21] and employment services,[22] web search,[23] pre-trial detention,[24] and facial recognition.[25]

---

15 Tal Zarsky, "The Trouble with Algorithmic Decisions: An Analytic Road Map to Examine Efficiency and Fairness in Automated and Opaque Decision Making," *Science, Technology & Human Values* 41, no. 1 (2016): 118-132.

16 The White House Office of Science and Technology Policy, *Big Data: Seizing Opportunities, Preserving Values*, May 2014, http://www.whitehouse.gov/sites/default/files/docs/big_data_privacy_report_5.1.14_final_print.pdf; President's Council of Advisors on Science and Technology, *Big Data and Privacy: A Technical Perspective*, May 2014, http://www.whitehouse.gov/sites/default/files/microsites/ostp/PCAST/pcast_big_data_and_privacy_-_may_2014.pdf

17 See The Leadership Conference, *Civil Rights Principles for the Era of Big Data*, 2014, http://www.civilrights.org/press/2014/civil-rights-principles-big-data.html, and the recurring conference on Data & Civil Rights: http://www.datacivilrights.org/

18 The Federal Trade Commission, *Big Data: A Tool for Inclusion or Exclusion? Understanding the Issues*, January 2016, https://www.ftc.gov/system/files/documents/reports/big-data-tool-inclusion-or-exclusion-understanding-issues/160106big-data-rpt.pdf.

19 The White House. *Big Data: A Report on Algorithmic Systems, Opportunity, and Civil Rights*, May 2016, https://www.whitehouse.gov/sites/default/files/microsites/ostp/2016_0504_data_discrimination.pdf

20 Solon Barocas and Andrew Selbst, "Big Data's Disparate Impact," *California Law Review* 104, no. 3 (June 2016): 671.

21 Latanya Sweeney, "Discrimination in Online Ad Delivery," *Communications of the ACM* 56, no. 5 (2013): 44-54.

22 Amit Datta, Michael Carl Tschantz, and Anupam Datta, "Automated Experiments on Ad Privacy Settings," *Proceedings on Privacy Enhancing Technologies* (2015): 92-112.

23 Safiya Umoja Noble, "'Just Google It': Algorithms of Oppression," University of British Columbia, December 8, 2015, https://youtu.be/omko_7CqVTA.

24 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.

25 Clare Garvie And Jonathan Frankle, "Facial-Recognition Software Might Have a Racial Bias Problem," *The Atlantic*, April 7, 2016, http://www.theatlantic.com/technology/archive/2016/04/the-underlying-bias-of-facial-recognition-systems/476991/.

AI paved with best intentions

When AI leads to discrimination, it is most often unintentional. The people who create the algorithms, the people who input the training data, and the people who maintain AI systems rarely *want* to create a biased outcome. And yet, properties of the data can lead to biased results, best intentions notwithstanding.

Given these risks, there have been calls for greater transparency and due process for automated decisions.[26] Understanding how bias creeps into AI systems is a difficult but critical challenge. Even being able to *recognize* that it is happening can be technically challenging, including for experts.[27]

Two complimentary communities of researchers have formed over the past few years to address these challenges. The first focuses on auditing algorithmic systems through "black box testing," carefully varying the information presented to an AI system and observing any apparent disparity in output according to race, gender, age, or other characteristics.[28] Rather than aiming to understand the inner-working on an AI system, these techniques involve creating a range of profiles for a website with personalized content, say, and seeing how the online experience differs for each.

The second focuses on correcting for bias in the training data, ensuring that models developed through machine learning meet some formally specified standard of fairness, prioritizing interpretability in the process of generating models, or introducing accountability mechanisms.[29] Both of these initiatives can play a crucial part in detecting and attempting to mitigate bias in AI, but they face a number of practical challenges, including uncertainty regarding the legality of some of the technical auditing methods.[30]

---

[26] Danielle Keats Citron, "Technological Due Process," *Washington University Law Review* 85 (2007): 1249-1313; Kate Crawford and Jason Schultz, "Big Data and Due Process: Toward a Framework to Redress Predictive Privacy Harms," *Boston College Law Review* 55, no. 1 (2014): 93-128.

[27] Sweeney, "Discrimination in Online Ad Delivery"; Datta, Tschantz, and Datta, "Automated Experiments on Ad Privacy Settings."

[28] See, for example, the Web Privacy and Transparency Conference, https://citp.princeton.edu/event/web/, and Auditing Algorithms From the Outside: Methods and Implications, https://auditingalgorithms.wordpress.com/.

[29] See Fairness, Accountability, and Transparency in Machine Learning, http://www.fatml.org/.

[30] Esha Bhandari and Rachel Goodman, "ACLU Challenges Computer Crimes Law That is Thwarting Research on Discrimination Online," Free Future, June 29, 2016, https://www.aclu.org/blog/free-future/aclu-challenges-computer-crimes-law-thwarting-research-discrimination-online; Ifeoma Ajunwa, Sorelle Friedler, Carlos E. Scheidegger, and Suresh Venkatasubramanian, "Hiring by Algorithm: Predicting and Preventing Disparate Impact," March 10, 2106, http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2746078.

Making inequality while seeking profit

The lasting historical effects of discrimination mean that AI systems may do the opposite of what their designers intend.

The reason for this, again, goes back to data, and what computer scientists working on discrimination in data mining call "redundant encodings."[31] In especially rich datasets (those with many variables tracking all sorts of things—the kinds of datasets that large social media companies or credit reporting agencies may collect), characteristics like race, gender, or age are almost certainly reflected in other, seemingly benign data points. Take race as an example. In the most obvious cases, redundant encoding can involve fairly obvious proxies for race like zip codes (such as areas historically subject to redlining and segregation). In subtler instances, other variables can act as unintentional proxies. For instance, race may be redundantly encoded in the list of websites visited by a particular user, sites especially popular among the Latino population in the United States, or African American teens on the East Coast, or white men above the age of 45, etc.

However, if the same data that redundantly encodes race also serves as a reliable predictor of customer value (people who visit a given combination of websites are more likely to click on payday loan ads, say), a company could systematically disadvantage members of a certain race in an attempt to maximize its profits.

Because inequality is not random, this is not very surprising. In a society suffering from bias along lines of gender, race, and other characteristics, these groups are more likely to be economically disadvantaged. And simply reproducing existing economic inequality, as in the examples above, will have a disproportionately negative effect on these groups.

Rendering judgment on the fairness of such rational market optimization is complex and difficult, especially because these practices are part of a wider set of forces and dynamics in capitalist societies. At the extreme, efforts to mitigate against these will necessarily begin to blur the line between a policy designed to ensure procedural fairness and one aimed at distributive justice.[32]

Certain computer scientists have put a fine point on this problem by demonstrating that there can be "a quantitative trade-off between fairness and utility."[33] The only way to ensure that consequential decisions do not amplify social biases—even unintentionally—is to ignore certain *relevant* details that also happen to act as redundant encodings. The resulting decisions would be less "accurate" (have less "utility") because they would grant opportunities to some people who—by the model's original estimate—might not "deserve" them. And it would do this at others' expense, effectively re-allocating the opportunities from one group to another. In some ways, this issue

---

31 Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel, "Fairness through Awareness," *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, 2012, 215.

32 Barocas and Selbst, "Big Data's Disparate Impact," 723–728.

33 Dwork et al., "Fairness through Awareness," 215.

echoes debates about affirmative action versus the idea of a level playing field, and whether long-standing social inequalities should or should not be addressed by direct intervention in individual decision-making.

AI ultimately raises a basic question about the line that divides objectionable discrimination from the reproduction of inequality. Should normative principles enshrined in discrimination law address these practices? Even setting aside the likely reception of a policy to address economic disparities with discrimination law, such regulations might not be the most efficient, effective, or fair way to remedy what is ultimately a problem of economic inequality. Nevertheless, in societies where economic redistribution is limited and progressive policies might raise political objections, discrimination law might be one important vehicle for ensuring that AI does not contribute to a more unequal society.

The lifted veil

By offering more accurate predictions, AI promises to reduce economic uncertainties that currently result in cross-subsidies between different groups in society. For example, the inability to predict with perfect accuracy whether someone will fall sick has tended to ensure that people engage in risk-pooling through health insurance, with the result that the financial contributions of the healthy help to cover the costs of the ill.

To the extent that AI can reduce the uncertainty around the onset of some medical conditions, it could help manage these risks more effectively. But it could also reduce the willingness of people to participate in such programs if the cross-subsidization becomes more obvious, and people realize they personally won't reap the benefits. In other words, those who know they're healthy might balk at the idea of paying for the sick.[34]

As the President's Council of Economic Advisors notes, this will very likely apply in other economic arenas with risk-based price discrimination,[35] and competition will likely push insurers down this path. Policyholders who possess traits or exhibit behaviors that AI recognizes as reliable predictors of the future onset of certain diseases or disorders will be charged higher premiums. In such cases, those in already less favorable circumstances may end up bearing more of the economic brunt of poor health. This is why critics often charge that even if these predictions are accurate and the insurer's behavior is rational, the effects are ultimately perverse.[36]

The capacity to precisely know who within a population will be more or less costly may endanger the solidarity upon which insurance and other such social safety nets rely. AI will allow companies to engage in far more effective "adverse selection," identifying the particular populations that they would do better to avoid or under-serve. Companies will be better placed to focus their attention on those populations that will prove most

---

[34] Alistar Croll, "New Ethics for a New World," *O'Reilly Radar*, October 17, 2012, http://radar.oreilly.com/2012/10/new-ethics-for-a-new-world.html.

[35] Council of Economic Advisers, *Big Data and Differential Pricing*.

[36] Upturn, "As Insurers Embrace Big Data, Fewer Risks Are Shared," *Civil Rights, Big Data, and Our Algorithmic Future*, September 2014, https://bigdata.fairness.io/insurance/.

profitable.[37] We may be undoing an important feature of insurance by giving organizations extraordinarily precise information that could allow these companies to make inferences that will deny care or limit coverage to people in need.

What this reveals is that *the absence of error or bias does not make the treatment of individuals appropriate or desirable*. Here we see the well-recognized danger of "a computer-generated class system" that "could unfairly stratify consumers, covertly offering better pricing to certain people while relegating others to inferior treatment."[38]

This is what Oscar Gandy calls a 'cumulative disadvantage': even accurate inferences have their costs if they further entrench already disadvantaged populations into less favorable circumstances.[39] These are not worries about prejudice, bias, or error; they are broader concerns about the fairness of a system that, in rationally apportioning opportunities and costs, compounds existing structural inequities.

Ultimately, the structure of the market may determine how discerning we want AI to be. In a society with universal health care, for example, there are no opportunities to engage in adverse selection because one insurer covers all costs. As Paul Krugman remarks, health insurers in the United States essentially compete by trying to deny coverage to those who are most likely to need it;[40] drawing especially fine distinctions between people simply exacerbates these tendencies. In a society that does not rely on markets for the provision of health insurance, however, the same tactics could benefit both a sole insurer and patients because it would allow the sole insurer to more effectively manage risk and thus reduce the overall costs of health insurance.

AI needs women and people of color

As a field, computer science suffers from a lack of diversity. Women, in particular, are heavily underrepresented. The situation is even more severe in the subfield of AI. For example, while some AI academic labs are being run by women, only 13.7 percent of attendees were women this past year at NIPS, one of the most important annual AI conferences.[41]

A community that lacks diversity is less likely to consider the needs and concerns of those not among its membership. As Jane Margolis and Allan Fisher point out, the underrepresentation of women in AI can have very serious consequences:

> In an example from computer science, some early voice recognition systems were calibrated to typical male voices. As a result, women's voices were literally

---

37 Bart Custers, *The Power of Knowledge: Ethical, Legal and Technological Aspects of Data Mining and Group Profiling in Epidemiology*, Nijmegen, Netherlands: Wolf Legal Publishers, 2004.

38 Natasha Singer, "Your Online Attention, Bought in an Instant by Advertisers," *The New York Times*, November 17, 2012, http://www.nytimes.com/2012/11/18/technology/your-online-attention-bought-in-an-instant-by-advertisers.html.

39 Oscar Gandy, *Coming to Terms with Chance: Engaging Rational Discrimination and Cumulative Disadvantage*, New York: Routledge, 2016.

40 Paul Krugman, "Why Markets Can't Cure Healthcare," *The New York Times*, July 25, 2009, http://krugman.blogs.nytimes.com/2009/07/25/why-markets-cant-cure-healthcare/.

41 Jack Clark, "Artificial Intelligence Has a 'Sea of Dudes' Problem," *Bloomberg*, June 23, 2016, http://www.bloomberg.com/news/articles/2016-06-23/artificial-intelligence-has-a-sea-of-dudes-problem.

> unheard...Similar cases are found in many other industries. For instance, a predominantly male group of engineers tailored the first generation of automotive airbags to adult male bodies, resulting in avoidable deaths for women and children.[42]

Amid growing recognition that heterogeneous groups outperform those that are more homogeneous,[43] there is also good reason to believe that increased diversity in the field of AI will help AI technologies to serve the interests of a diverse population.

Researchers note that the field's struggle to identify and confront issues of bias and inequality has been hampered by the current lack of diversity.[44] The concerns that shaped the debates about risk and AI have been driven by the interests and anxieties of wealthy, white men, rather than members of historically disadvantaged communities.[45] To address pressing issues of bias, discrimination, and inequality, the AI community will need to draw on a broader range of perspectives.[46] Like all technologies before it, AI reflects the values of its creators, and increased diversity may assist in a future for AI technologies that promotes greater equality.[47]

Questions to consider

- Will AI merely redistribute economic value in ways that favor the already wealthy, or grow the economy in new ways?
- How might we ensure that the economic benefits of AI technologies are widely distributed throughout society? Can market forces alone ensure that AI has a progressive redistributive effect, or are regulations or programs necessary to achieve these goals? Which mechanism is likely to be most efficient, effective, or most widely supported?
- How should we address the likely information and power asymmetries produced by AI technologies? Is making AI capacities available in non-proprietary modes enough to redress imbalances and produce a level playing field?
- How should we guard against bias in AI? Should discrimination law play a role? How might we make use of technical solutions in combination with institutional procedures and professional ethics?
- Is liability a sufficient incentive to guard against data errors and avoidable bias or is there a need for new laws?
- Are most consequential cases of bias in AI likely to come to light on their own or remain unnoticed? Who is best positioned to determine whether AI suffers from biases? How can that work be supported?
- What practical, institutional, and legal challenges do those working on auditing AI systems face and how should these be addressed?

---

42 Jane Margolis and Allan Fisher, *Unlocking the Clubhouse: Women in Computing*, Cambridge, MA: MIT Press, 2003, 2-3.

43 Scott Page, *The Difference: How the Power of Diversity Creates better Groups, Firms, Schools, and Societies*, Princeton, NJ: Princeton University Press, 2008.

44 Kate Crawford, "Artificial Intelligence's White Guy Problem," *The New York Times*, June 25, 2016, http://www.nytimes.com/2016/06/26/opinion/sunday/artificial-intelligences-white-guy-problem.html

45 Ibid.

46 For example, Women in Machine Learning, http://wimlworkshop.org/.

47 Crawford, "Artificial Intelligence's White Guy Problem."

- To whom should the burden of correcting for bias fall? Are the designers/developers of AI responsible for purging bias from their systems, even if these biases reflect widely held social beliefs expressed in the historical data from which AI systems learn?
- How should we differentiate between cases where AI discriminates and cases where AI replicates economic inequality along, for instance, racial lines? Should designers of AI systems incur the costs of addressing inequality in society?
- Ultimately, what will “fair” AI aim to achieve? And how will such notions of fairness gain political legitimacy and practical buy-in?
- Will AI unravel insurance markets or improve the management of risk? What structures are necessary to ensure that AI can improve how institutions manage risk while not contributing to further inequality?
- How will greater diversity in the professional field of AI shift the values and norms embedded in these systems? What kinds of diversity should the community attempt to cultivate and how should it do this?
- How do we encourage people to consider the risks of AI while not discouraging people from adopting the technology in ways that might benefit disadvantaged communities?

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

WORKSHOP PRIMER: LABOR & AI

Contents

Brief overview

Artificial intelligence (AI) systems are set to transform labor in the 21st century.[1] Although AI technology is in its early stages, powerful systems are already at work across an array of sectors, allowing researchers ample opportunities to investigate current and expected impact. Too often, discussions of labor and AI have focused only on the fear of a jobless future. Current research demonstrates that more complex and more immediate issues are at stake, affecting not only labor markets in the abstract, but also employer-employee relations, power dynamics, liability and responsibility, and the role of work in human life.

[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 labor, 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.

Economists disagree

The role of automation[2] in the economy is far from a new subject of inquiry, and considerations of the impact of AI emerge from within long-standing debates. We do not aim to summarize the full scope of economic arguments on this topic here, but simply to review key points and areas of contention within a much broader discourse. A key topic of contention is whether automation decreases the demand for human labor as it increases productivity, or not. While intuitively it may seem the demand for labor would decline as automation increases because there would be only a finite amount of work to be done – and some economists argue this – others don't see it this way, referring to this assertion as the "lump of labor" fallacy.[3] These dissenting economists suggest that as productivity rises in one industry (due to automation or other factors), new industries emerge along with new demand for labor. For example, in 1900 agriculture comprised 41 percent of the United States workforce. By 2000, agriculture comprised only 2 percent. Labor economists David Autor and David Dorn point out that even with this dramatic change, unemployment has not increased over the long-term and the employment-to-population ratio has in fact grown.[4] Nonetheless, Autor and Dorn, among others, point out the rise of "job polarization," in which middle-skill jobs decrease and leave some high-skill and more low-skill jobs. Other economists, such as James Huntington and Carl Frey, are more dire in their prediction that AI will dramatically reduce the number of jobs available.[5]

There are also economists debate whether the transformations and fluctuations in labor markets are related to technology at all, or instead caused by economic policy. Such arguments focus on what how and when institutions and regulatory mechanisms should be brought to bear on AI technologies. Robert Gordon, for example, argues that the current waves of innovation are not as transformative as they seem.[6] But many economists are beginning to concur that labor markets are undergoing a consequential transformation due to technological change. These include Joseph Stiglitz and Larry Mishel, who argue that a keen focus must be maintained on regulation and other policy

---

2 Automation is defined here as "a device or system that accomplishes (partially or fully) a function that was previously, or conceivably could be, carried out (partially or fully) by a human operator." At this broad level, technologies of automation and artificial intelligence are interconnected. Raja Parasuraman et al., "A Model for Types and Levels of Human Interaction with Automation," *IEEE Transactions on Systems, Man and Cybernetics* 30(3) (2000).

3 Tom Standage, "Artificial Intelligence: The return of the machinery question," *The Economist*, June 25 2016. http://www.economist.com/news/special-report/21700761-after-many-false-starts-artificial-intelligence-has-taken-will-it-cause-mass.

4 David Autor and David Dorn, "How Technology Wrecks the Middle Class." *New York Times* April 24 2013, http://opinionator.blogs.nytimes.com/2013/08/24/how-technology-wrecks-the-middle-class/?_r=0.

5 Carl B. Frey and Michael Osbourne, "The Future of Employment: How Susceptible Are Jobs to Computerization" Oxford Martin School Programme on the Impacts of Future Technology Working Paper, (September 17, 2013).

6 Robert J. Gordon,"The Demise of U.S. Economic Growth: Restatement, Rebuttal and Reflections," National Bureau of Economic Research Working Paper, (February 2014), http://www.nber.org/papers/w19895.

changes concerning AI and automation in order to protect workers.[7] Economic researchers are closely tracking national labor markets and institutions as a means to examine the social and economic impact of AI in the near term.[8]

Beyond a purely economic lens, one could also approach these issues by examining local communities and individual lives. In recent years, researchers have begun to examine how automated systems relying on big data (from Uber to automated scheduling software used in large retailers) are destabilizing traditional dynamics between employers and employees.[9] Findings suggest that while these systems could be designed to empower workers, there are substantial ways in which these technologies as currently designed disempower workers, entrench discrimination, and generate unfair labor practices.[10]

AI as management

We now turn to examine the increasing use of AI technologies to replace and augment employee management. Shift scheduling is one of the ways in which AI technologies are used in such a way. While scheduling has always been a part of managing labor, AI technologies change the scale and precision of this task. Kronos,[11] one of the most widely used "workforce management" platforms with customers like Starbucks and Target, offers a suite of services including just-in-time scheduling. This kind of scheduling relies on the analyses of large datasets (from weather patterns to past sales) in order to predict peak hours of consumer demand and schedule employees accordingly. The fewer employees on the clock during slow periods, the more the employer saves on labor.

With many companies in the retail and service industry using just-in-time scheduling, it is now normal to demand that workers be available (proximate to the workplace, with the ability to receive a call or text) during all "open" hours, but only to assign short, four-hour shifts to workers on the clock. Often, these assignments are made at the last minute. Those not able to accommodate this kind of uncertainty (those with children, second

---

7 Lawrence Mishel, John Schmitt and Heidi Shierholz, "Assessing the Job Polarization Explanation of Growing Wage Inequality," EPI Working Paper Paper (January 2013). http://www.epi.org/publication/wp295-assessing-jobpolarization-explanation-wage-inequality/. See also Joseph Stiglitz, *The Price of Inequality: How Today's Divided Society Endangers Our Future*, New York: W.W. Norton and Co, 2012.

8 Roosevelt Institute. "Technology and the Future of Work: The State of the Debate," White paper. Open Society Foundations Future of Work Project (April 2015), https://www.opensocietyfoundations.org/publications/technology-and-future-work-state-debate.

9 For a review of recent work, see Min Kyung Lee et al., "Working with Machines: The Impact of Algorithmic and Data-Driven Management on Human Workers," *CHI 2015 Proceedings* (April 2015), https://www.cs.cmu.edu/~mklee/materials/Publication/2015-CHI_algorithmic_management.pdf. For a clear case study, see Alex Rosenblat and Luke Stark, "Algorithmic Labor and Information Asymmetries: A Case Study of Uber's Drivers, " *International Journal of Communication* 10 (2016): 3758-3784.

10 See, for instance: Julia R. Henly, H. Luke Shaefer, and Elaine Waxman, "Nonstandard Work Schedules: Employer- and Employee-Driven Flexibility in Retail Jobs," *The Social Service Review* 80(4): 609-634 (2006); Leila Morsy and Richard Rothstein, "Parents' Non-Standard Work Schedules Make Adequate Childrearing Difficult," Economic Policy Institute Issue Brief (Aug. 6, 2015), http://www.epi.org/files/pdf/88777.pdf. Additional research and examples are discussed below.

11 http://www.kronos.com/

jobs, or other duties of care), are less employable. And, of course, uncertainty about hours necessarily means uncertainty about finances. According to a 2014 national survey of workers aged 26-32, three-quarters of hourly workers reported fluctuations in the hours they worked during the previous month, with hours fluctuating on average by 49 percent. For part-time workers, hours fluctuated by 87 percent. Moreover, 41 percent of hourly workers reported that they learn what their work schedule will be a week or *less* in advance.[12]

Though a few researchers have found that just-in-time scheduling provides valuable flexibility, far more researchers have documented the strain and precarity experienced by workers subject to these systems.[13] The adverse experiences of workers managed by such systems include chronic underemployment, financial instability, insufficient benefits and protections that are traditionally granted full-time employees, as well as the structural inability to plan for family or self-care (or even search for another job given the round-the-clock availability that such jobs demand from workers). Moreover, the workers most likely to be affected by these practices are disproportionately women and minorities.[14]

Companies using automated scheduling systems argue that efficiency gains achieved through such systems allow them to offer lower consumer prices. However, management researchers like Zeynep Ton have argued that the perceived trade-off between low prices and investment in labor is a false dichotomy. Her research shows that companies like Costco and Trader Joe's have been able to implement programs that invest in their employees while delivering low-cost goods and increased customer satisfaction.[15]

It is important to emphasize that any AI technology requires both the use of big datasets (such as past sales, email or call pattern data, etc.) and algorithms to optimize for specific goals, be they profitability or other discrete quantitative metrics, such as fuel consumption or job retention. In the examples above, minimizing labor costs appear to be the goal, outweighing all other considerations. Across the board, these goals are being set solely by employers.

---

12 Susan J. Lambert, Peter J. Fugiel, and Julia R. Henly, "Precarious work schedules among early career employees in the US: A national snapshot," Employment Instability, Family Well-being, and Social Policy Network, University of Chicago (2014), https://ssascholars.uchicago.edu/sites/default/files/work-scheduling-study/files/lambert.fugiel.henly_.precarious_work_schedules.august2014_0.pdf.

13 Farhad Manjoo, "Uber's Business Model Could Change Your Work." *New York Times* 28 June, 2015, http://www.nytimes.com/2015/01/29/technology/personaltech/uber-a-rising-business-model.html?_r=0.

14 Carrie Gleason and Susan Lambert, "Uncertainty by the Hour," Position paper. Open Society Foundation Future of Work Project (2014), http://static.opensocietyfoundations.org/misc/future-of-work/just-in-time-workforce-technologies-and-low-wage-workers.pdf.

15 Zeynep Ton, "Why Good Jobs Are Good for Retailers," *Harvard Business Review*, (January 2012), https://hbr.org/2012/01/why-good-jobs-are-good-for-retailers.

Logistics by AI

Instant delivery is already augmented by automated and intelligent systems, and in the near future, possibly drones as well. Amazon's Prime Delivery, for example, has conditioned consumers to expect that anything they order can arrive at their doorstep in a matter of days, even hours. Fulfilling these expectations requires immense socio-technical systems that coordinate, sort, transport and track vast numbers of packages from one physical location to another. At every step of this process, from warehouses to distribution centers to trucks on local streets, this logistics work is often structured by AI systems.

For instance, UPS uses a proprietary analytics system, called a "telematics" system, to track, measure and manage UPS delivery truck drivers.[16] This system wirelessly transmits data from remote sensors, including RFID chips and GPS devices, to centralized computers, where the data is analyzed. In addition to the handheld devices drivers carry, there are over 200 sensors on any given truck, tracking and recording specific information about the movement, location, and speed of the truck, as well as monitoring specific actions, like fastening a seat belt. All of this information is transmitted, often in real time, to the computer screens of supervisors remote from the driver and truck. In this case, workforce management is centralized and the capabilities are vastly expanded. Meanwhile, the autonomy of frontline workers and managers is greatly reduced.

These systems are celebrated by those who sell and implement them as promoting safety and increasing efficiency regarding fuel, maintenance, and labor. Investigative journalist Esther Kaplan, who wrote about the conditions of UPS workers in 2014, pointed out that UPS corporate filings during the time the telematics system was introduced showed daily domestic package deliveries growing by 1.4 million between 2009 and 2013, while the total number of employees was reduced by 22,000.[17] Fewer workers doing more work means higher profit margins and happy shareholders. However, Kaplan found that driver safety and overall working conditions were reduced when these systems were implemented. This finding is supported by academic research from Karen Levy, who found that increased digital surveillance of long-haul truck drivers did not necessarily promote safety, but rather created new conditions of economic hardship for truck drivers who in turn developed workarounds that often made their work more dangerous.[18]

---

16 Esther Kaplan, "The Spy Who Fired Me," *Harper's Weekly* (March 2015): 31-40.

17 Ibid, 32.

18 Karen Levy, "To fight trucker fatigue, focus on economics not electronics." *The Los Angeles Times*, June 28 2015, http://www.latimes.com/opinion/op-ed/la-oe-levy-trucker-fatigue-20140716-story.html.

A boss without a face

Over the last decade, monitoring and tracking of employees has become common in a variety of industries from retail to logistics (as in the UPS example) to office administration. This monitoring provides a primary source of "big data" used by complex algorithms and AI systems to make just-in-time decisions. These decisions are not just about scheduling, but about resource allocation and perceived job performance, among others. Remote and disembodied management can effectively shift power from frontline managers (people who can be approached face-to-face and who may understand nuanced situational context) to executives whose view is shaped by more impersonal aggregate employee data. In addition, power and control are also significantly shifted to engineers and those who design the systems, in which possible actions and reactions are structured in code months or even years in advance, potentially without knowledge of local context.

Remote management via AI system can make it harder to hold employers accountable for decisions made by "the system" that immediately and materially impact employees. For example, platforms like Uber (powered by big data and basic AI) serve to remotely control routes, pricing, compensation, and even interpersonal norms -- determinations and decisions that would traditionally be in the hands of human management.[19] Beyond simply obscuring the face and reasoning behind a given decision, this type of remote management is very often not recognized as "employee management" at all. Because these new technologies do not fit neatly into existing regulatory models, companies like Uber see themselves as technology companies, not managers of employees. Following this, such companies view their role as providing a marketplace only to facilitate connection, not an employer of workers, and not responsible to workers as a traditional employer would be. In this configuration, workers end up assuming the risks of employment without guaranteed benefits (such as decreased tax burdens, healthcare, and other workplace protections) or potential modes of redress.

AI as the door to opportunity

Hiring is an important domain where AI-driven processes are already being deployed in the workforce. This is an area for special scrutiny. As scholars have demonstrated, there are good reasons to assume that discrimination in hiring practices is a widespread occurrence, and this discrimination disproportionately affects women and minority applicants.[20] AI hiring systems have the troubling potential to encode, and perhaps

---

[19] Min Kyung Lee et al., "Working with Machines: The Impact of Algorithmic and Data-Driven Management on Human Workers." *CHI 2015 Proceedings* (April 2015), https://www.cs.cmu.edu/~mklee/materials/Publication/2015-CHI_algorithmic_management.pdf.

[20] Solon Barocas and Andrew Selbst, "Big Data's Disparate Impact," *California Law Review* 104 (2016), http://ssrn.com/abstract=2477899.

amplify existing biases against historically disadvantaged groups. For instance, Cornerstone (formerly Evolv),[21] a "talent management" company, found a correlation between job retention and the distance an applicant lived from her workplace. However, the firm also realized that including this correlation in making hiring assessments might unfairly advantage people who were able to live near work (which might be in a high-rent neighborhood, or a neighborhood far from a given ethnic community), disparately impacting disadvantaged socio-economic groups. Considering the potential for discrimination, they did not include the metric in their system.[22] This example underscores that correcting bias and protecting workers does not happen without keen attention, and sometimes a metric needs to be discarded altogether. Such systems must be closely examined and audited to assess disparate impact, analyzing who is getting hired, who is not, and why.

AI's human caretakers

AI systems require more than computer code and human creativity. They require physical infrastructures to store, process, and transmit data. They require vast libraries of ancillary code (on which they along with most Internet technologies are dependent). They also require humans to maintain these infrastructures and tend to the "health" of the system. This labor is often invisible, at least in the context of the popular stories and ideas of what AI is and what it does. It includes custodial staff that clean data centers, maintenance or repair workers who fix broken servers, and what one reporter termed "data janitors," those who "clean" the data and prime it for analysis.[23] While most media discussions elide this kind of work or frame it as an obstacle that will soon be performed by computers, scholar Lilly Irani has pointed out that this kind of work is integral to AI systems, describing it as "the hidden labor that enables companies like Google to develop products around AI, machine learning, and big data."[24]

Taking a closer look at one of the roles that make up this class of "hidden labor," a 2014 *Wired* article by Adrian Chen described a day in the life of content moderators. In this article Chen followed workers in the Philippines, where much content moderation is done, as well as workers in the United States.[25] Sitting in front of screens, staring at a rapid stream of images or videos (think of a modern factory assembly line), a content moderator must quickly decide whether the content is objectionable given a particular set of policies. Much of this work is outsourced at low wages. However, because

---

21 https://www.cornerstoneondemand.com/

22 Dustin Volz, "Silicon Valley Thinks It Has the Answer to Its Diversity Problem," *National Journal*, Sep 26 2014, http://www.theatlantic.com/politics/archive/2014/09/silicon-valley-thinks-it-has-the-answer-to-its-diversity-problem/431334/.

23 Steve Lohr,"For Big-Data Scientists, 'Janitor Work' is Key Hurdle to Insights." *New York Times* August 17 2014, http://www.nytimes.com/2014/08/18/technology/for-big-data-scientists-hurdle-to-insights-is-janitor-work.html.

24 Lilly Irani, "Justice for 'Data Janitors, " *Public Books,* (January 2015), http://www.publicbooks.org/nonfiction/justice-for-data-janitors.

25 Adrian Chen, "The Laborers Who Keep Dick Pics and Beheadings Out of Your Facebook Feed," *Wired*, October 23 2014, http://www.wired.com/2014/10/content-moderation/.

companies want moderators to have a deep cultural context informing many of these assessments (such as discerning pornography from art in a museum), there are also many workers within the United States who perform these jobs, which are also low-wage contract work. Beyond the paid labor, everyday users of platforms are encouraged to 'flag' and label inappropriate content, which helps to train underlying AI models.[26] This work is valuable but completely unpaid. The lesson here is that even the most sophisticated machine learning object recognition algorithms are still far away from automatically detecting the appropriate cultural meaning of such images, i.e. from appropriately classifying them for use in AI systems. These algorithms still need humans, checking and approving the data before it is entered into the system, even when they are unaware that this is what they are doing. This constitutive need for "artificial artificial intelligence" (the tagline for what is provided by Amazon's Mechanical Turk, a microtasking platform) is an example of the kinds of work that is newly necessary as an accompaniment to AI systems.

AI across the labor spectrum

Discussions about the future of AI and labor tend to focus on what are traditionally conceived of as low-wage, working class jobs such as manufacturing, trucking, and retail or service work. With justification, researchers have focused on these jobs because the vulnerability of workers in these industries appears to be most acute: with the decline of unions and collective bargaining power in conjunction with wage stagnation and increasing income inequality, the working class and working poor are increasingly finding themselves in difficult positions.[27]

However, research on the rapidly expanding deployment of AI systems demonstrates that the impact of this deployment will be felt across sectors, from so-termed "low skilled" or "semi-skilled" work that may not require specialized training or an advanced degree, to professional and expert work that requires advanced degrees and highly specialized experience.[28] Generalized repetitive work that is composed of basic tasks that can, in theory, adhere to a specific ruleset (e.g. accounting, law, and radiology) is the quality that invites the possibility of automation, irrespective of the education or expertise required of a human performing the work.

---

[26] Kate Crawford and Tarleton Gillespie, "What is a flag for? Social media reporting tools and the vocabulary of complaint," *New Media & Society* 18(3): 410-428 (2016), http://nms.sagepub.com/content/18/3/410.abstract

[27] For example, see the speech by Jason Furman, Chairman of the Council of Economic Advisers, at *AI Now* on July 7, 2016: "Is This Time Different? The Opportunities and Challenges of AI," https://www.whitehouse.gov/sites/default/files/page/files/20160707_cea_ai_furman.pdf

[28] For instance, see Henry Siu and Nir Jaimovich, "Jobless Recoveries," Third Way, April 8 2015, http://faculty.arts.ubc.ca/hsiu/pubs/NEXT_-_Jobless_Recoveries.pdf. See also Tom Standage,"Artificial Intelligence: The return of the machinery question," *The Economist*, 25 Jun 2016, http://www.economist.com/news/special-report/21700761-after-many-false-starts-artificial-intelligence-has-taken-will-it-cause-mass.

Consider the legal profession. According to a variety of pundits, lawyers are in danger of being made obsolete by AI. Like many headlines, this claim is overblown, at least in the near-term. However, like the examples above, the implications of AI for lawyers are not so much about job *replacement* as they are about how AI restructures the job. AI is currently and will continue to change what lawyers do and what is perceived as "legal expertise." In a report released in early 2016, legal scholar Dana Remus and economist Frank Levy analyzed the actual billed hours by lawyers at two law firms.[29] What the researchers found was that only a modest portion of the activities billed would be able to be automated in the near future. AI technology would not be feasible for all the tasks, including reading and analyzing documents, counseling, appearing in court, and arguing in front of juries. In fact, while software programs that are meant to review and surface significant content in the context of a case document review exist, these software programs require a great deal of complex analysis and time-consuming work by humans, involving planning and structuring the program for each specific case. Moreover, many of the tasks that involved automated review or the automation of filling out forms (performed by companies like LegalZoom or Rocket Lawyer) are tasks that junior lawyers or paralegals perform, not senior level partners. This more fine-grained analysis of AI's impact on lawyers demonstrates that on one hand, claims of job replacement are often overstated, but that, on the other hand, professional work that requires advanced education will also be impacted *alongside* low-wage work.

Expert AI will impact expert humans

Even as AI systems are called upon to perform rote tasks and immense calculations, they are also increasingly designed to augment or even replace human expertise. The example of law, above, is one such potential case. Expert-augmenting AI is also being developed in the medical field and has already shaped the current field of aviation. For instance, today, a modern aircraft spends most of its time in the air under the control of a set of technologies, including an autopilot, GPS, and flight management system, which govern almost everything it does, relatively autonomously. This has allowed aviation safety to advance tremendously in the past decades. Nonetheless, researchers have found that the ways in which these advances have developed have come at a certain cost. While automation is often assumed to relieve humans of menial tasks, freeing them to think about more important decisions, the field of human factors engineering research has proven this not to be the case.[30] Specifically, in aviation, pilot awareness has been documented to generally decrease with increased automation, leading to skill atrophy or

[29] Dana Remus and Frank Levy, "Can Robots Be Lawyers? Computers, Lawyers, and the Practice of Law," Working paper, December 30, 2015, http://ssrn.com/abstract=2701092.

[30] Laine Bainbridge, "Ironies of Automation," *Automatica* 19 (1983): 775-779; Raja Parasuraman and V. Riley, "Humans and Automation: Use, Misuse, Disuse, Abuse," *Human Factors* June 39(2) (1997): 230-253.

even deskilling.[31] A recent article also observed that a rising trend of physician burnout was linked to the use of electronic health records and computer-entry systems.[32]

Moreover, automation raises issues of professional responsibility. For example, the pilot and co-pilot may only be actively in control of the aircraft for a few minutes on any given flight; yet pilots are viewed as being in control and responsible, both by the general public and according to existing laws and regulations.[33] This distributed control raises questions about accountability in automated and AI systems as well professional identity and expertise. Beyond aviation, we can imagine how these kinds of augmented systems raise important questions. For example, a police officer might find herself in a similar situation when working with a predictive policing system: is she, or the system, ultimately responsible for determining an arrest? Alternatively, how should a healthcare professional negotiate when working with a diagnostic system? Which expertise is "most expert," in what contexts is "overriding the system" allowed, and how would this impact liability?

Questions to consider

Some questions to consider regarding the rapid introduction of AI systems into labor sectors:

- Does the development of AI mean a future with rapidly declining employment? If so, what would such a future look like and who will be hardest hit? How can we ensure equitable distribution of resources in the event of such a future?
- What impact does AI have on the power dynamics between employers and employees, and in which contexts?
- How might AI provide new mechanisms of control over workers and leave them more vulnerable to exploitation? Are AI systems already creating these dynamics, and if so, how can we best assess their impact?
- Alternatively, how might AI systems empower disadvantaged workers and effectively augment skill, expertise, and agency?
- What would AI systems used for worker management look like if their goals were set by workers or other stakeholders, as opposed to (or in collaboration with) employers? How might AI systems used for worker management introduce a more balanced way to set and evaluate goals between employers and workers?
- When AI systems are deployed, whose labor is required for computer software to appear 'intelligent' (e.g. when humans are used to 'perform' the last mile in AI systems)? How is this work valued if it is kept hidden?

---

[31] Nadine B. Sarter, David D. Woods, Charles E. Billings, "Automation Surprises," in *Handbook of Human Factors & Ergonomics* 2nd ed., G. Salvendy, ed. New York: Wiley, 1997.

[32] Tait D. Shanafelt et al., "Relationship Between Clerical Burden and Characteristics of the Electronic Environment With Physician Burnout and Professional Satisfaction," *Mayo Clinic Proc* XXX (2016): 1-13.

[33] M.C. Elish and Tim Hwang, "Praise the Machine! Punish the Human! The Contradictory History of Accountability in Automated Aviation," Data & Society Intelligence & Autonomy Working Paper, 2015, http://ssrn.com/abstract=2720477.

- How should professional discretion - such as challenging a decision - be exercised by experts whose work is partially automated, or augmented by AI? What social, economic and institutional pressures might come to bear?
- What will the shift to increasingly automated decision making mean for perceptions of expertise and the value of human capabilities? How will this change the imperatives of primary, secondary and postsecondary education? Further in the future, what role will work play in society and our constitution of identity and self-worth?