Normalized and Normative

Ethical Implications of Machine Learning Decision Making

Spring 2019

William Yao

Professor Jasanoff

All models are wrong, but some are useful.

Coined nearly seventy years ago by pioneering computer scientist John McCarthy, the phrase *artificial intelligence* still lacks a single agreed upon definition. Since then, the terms *machine learning*, *deep learning*, and *neural network* have permeated the now burgeoning field of study. However, even in science and technology policy scholarship, the distinctions between the terms are often not articulated accurately nor clearly. As a result, research communities, legal actors, policymakers, and society at large have ascribed their own understandings of them, informed by various contextual, social, and ethical circumstances. Moreover, the term *artificial intelligence* is dynamic, updating as a function of technological progress, adding a dimension of complexity. Consequently, a stark knowledge and understanding asymmetry has arisen that must be addressed in various situations where these parties interact, via translational work. Two major symptoms include overly optimistic projections of AI capabilities and dystopian formulations of the demise of our society (Atkinson). A product of the general public's skewed perception of AI technologies and the media's tendency to overstate certain achievements has resulted in this inflated perspective. In a similar vein, some have fashioned a narrative that omnipotent forms of artificial intelligence will lead to the imminent demise of humans. Nonetheless, researchers, policymakers and legal actors have begun to recognize the potential for machine learning tools (that are used in various industrial and legal applications) to exacerbate systemic discrimination and perpetuate social inequities. Notably, Joy Buolamwini, MIT researcher and the founder of the *Algorithmic Justice League*, exposes some of the shortcomings of commercial facial recognition and analysis technologies and illuminates the reasons for the resultant bias.

In a letter to the Massachusetts Legislature regarding a new Senate Bill, scholars of the Berkman Klein Center for Internet and Society warn against oversimplified conceptions of the issues raised by algorithmic solutions in a criminal justice context, characterizing the terms of the bill as "cursory" and failing to "address the complex and nuanced issues implicated by actuarial risk assessments" (Barabas). In the context of these concerns, this paper first surveys the existing social perceptions of machine learning on a meta level, followed by a careful examination of the pertinent ideas of transparency, privacy, accountability and fairness in a machine learning context. Lastly, I propose the establishment of a governing body to support the ethical application of machine learning decision making tools and methodologies.

* * *

In environmental science, *global warming* and *climate change* are scientifically distinct concepts, but are used interchangeably in the media and by the general public. This terminology entanglement blurs key scientific distinctions and contributes to a loss of understanding. Moreover, *global warming* and *climate change* have contrasting connotations, and this nuance is ignored when the terms are used improperly. In a similar fashion, in the machine learning domain, distinctions between terms mentioned above are frequently blurred[1]. Encouraging an

[1] In simple terms, a neural network is a mechanism where a set of weights is determined over a series of sequential linear and nonlinear functions. Deep learning is a type of machine learning where a multilayer neural network is used.

increased emphasis on the correct usage of terminology will be productive for the body of discourse around the subject. If the machine learning community acknowledges the long-term importance of clear terminology, it can augment our society's understanding and consequently benefit the research community (through additional support via funding or public interest).

Computer scientist and tablet technology pioneer Jerry Kaplan remarks that the field "has a long history of exploiting our natural tendency to anthropomorphize objects" (Aktinson). Although originally inspired by the chemistry of the human brain, neural networks do not resemble the structure of the brain of the brain beyond a very preliminary degree (Laskar). However, the term *neural network* itself implies a brain-like quality, and thus has contributed to the slew of faulty, and sometimes entirely erroneous, depictions of the capabilities of the technique. This phenomenon is a product of the lack of understanding of the technical details by the general public combined with the absence of proper transitional work via an intermediary. Unfortunately, imprecise terminology can have a snowballing impact on the initial perception and subsequent understanding of a novel scientific or technical phenomenon for the general public. As such, I propose a shift towards using a different term to describe the same methodology. It will be critical for computer scientists to consciously present the new term in a manner that will add clarity, not confusion.

* * *

The stark information asymmetry that the field of AI has presented for society at large renders it essential to clearly and rigorously articulate the pertinent ethical implications of using machine learning decision-making tools and address them precisely. In her book *Hello World*, British mathematician Hannah Fry expresses that we often think about algorithms in the wrong way. Rather than overemphasizing the pursuit of algorithmic perfection, we should shift our focus to addressing “the very human habit of over-trusting machines” (qtd. in Segal). The sociological forces that drive the design, implementation, and interpretation of automated decision-making methodologies are deeply entrenched and should not be viewed as exogenous. Harvard Computer Science Professor Cynthia Dwork and UC Berkeley Information Policy Professor Deirdre Mulligan describe this phenomenon as a “sociotechnical system,” where the discussions of values must precede the mathematical ones. Moreover, Fry remarks that algorithms are malleable and adjustable, and can be improved as people continue to recognize their shortcomings. On the other hand, it is impossible calibrate a human’s thought process. Fry says, “humans can’t tell you accurately how their arriving at decisions. Humans are sloppy and messy and irrational” (Segal). Dwork and Mulligan concur that algorithmic solutions provide a valuable opportunity to leverage technology to elucidate and execute on the values that humans aim to exhibit. For machines, “policies amenable to formal description can be built in and tested for,” whereas the same cannot be done for the brain (Dwork).

There is a reasonable consensus that some weighted combination of the ideas of accountability, transparency, privacy and fairness constitute most machine learning ethics discussions. Dartmouth Computer Science Professor Hany Farid advocates for “thoroughly understanding how [AI tools] work,” “thoroughly understanding their accuracy” and “thoroughly understanding their fairness” (Farid). Harvard Law School Professor Urs Gasser values the notions of “transparency, accountability, explainability and fairness of AI systems.” MIT researcher Dr. Micah Altman et al. emphasize the importance of “privacy, equity, fairness and autonomy.” Professor Dwork and Professor Mulligan feel that privacy and transparency are important paradigms in these discussions. While these notions are clearly central in establishing a basic framework under which to deliberate about machine learning ethics, they are not sufficient. In the following sections, I will closely examine these virtues and the relationships between them in a machine learning context. Then, I will pose some of the nuanced ethical and normative questions that arise.

There are myriad technologies that consumers place blind faith in. For example, most medication that we self-administer is effectively a black box algorithm in the sense that we have a minimal understanding of how it works but nonetheless have a certain degree of trust in its efficacy and benefits. What forms of socialization must occur for society to trust an entity that it does not understand? Individuals are socialized to trust certain technologies because they have undergone requisite validation procedures or are subject to some form of oversight. For example, the FDA mandates a "safe and effective" standard for drugs and outlines a rigorous multiphase testing process to achieve that standard (FDA). While the general public may not understand the chemical technicalities of an antibacterial agent, a governing body has served as an intermediary between the researchers and the public to demonstrate that the drug is safe. As such, the information asymmetry between the medicine developers and consumers has been leveled via a form of translation by the FDA. Such a framework is absent in field the machine learning, and the stark information asymmetry between the algorithm designer and user is potentially prohibitive as a result. In light of this lack of comprehension by the consumer, how can we assess accountability?

Accountability speaks to our intuitive desire seek recourse when an incorrect judgment is made, or an operation goes awry. What course of action can we take when an individual is harmed? Can a machine learning model take responsibility for its inaccuracies and flaws? Cathy O'Neil, mathematician and author of *Weapons of Math Destruction*, explains the phenomenon of hiding behind algorithms as a way to perpetuate bias without the burden of responsibility.

Revering algorithmic tools with an aura of objectivity and interpreting their results with this mindset will inevitably exacerbate biases, as ProPublica's report on COMPAS software elucidates (Angwin). However, if and when we move away from this mindset, it will become necessary to develop a framework and methodology under which accountability can be established. A crucial aspect of that process involves model transparency.

Transparency vis-à-vis machine learning algorithms may not be as straightforward compared to other domains, simply because most policymakers and legal actors would struggle to evaluate a machine learning model even if granted complete access to it. Harvard Computer Science Professor Finale Doshi-Velez and Harvard Law School Clinical Fellow Mason Kortz investigate the role of explanation for establishing accountability and transparency in AI applications. Presently, AI tools are *able* to, and should be expected to, produce explanations for their decisions via "human-understandable" factors. Consequently, machine learning models should be held to the same standards of explanation as humans are in legal proceedings (Doshi-Velez). Under this framework, we are given the opportunity, as a research community, policymaking body, or society as a whole, to discuss and formalize what it means to be a "human-understandable" and what it means to be a valid explanation. If algorithm designers are mandated to create tools that can produce such explanations, they are obliged to establish standards of interpretability and thus augment their own understanding of what aspects of an explanation are important. In "The Nature of Patents in Biotechnology" I outline a framework for distilling complex ideas into simpler terms. That framework may be useful vis a vis machine learning. This proposal makes an important step in developing a meaningful form of transparency for AI tools. Doshi-Velez and Kortz acknowledge that, as these two forms of intelligence diverge, a new framework for explainability may need to be created. As the field strives towards more rigorous standards of transparency, it is important to balance them again societal expectations of privacy.

Privacy concerns present their own set of conundrums in a machine learning context. Machine learning models are trained via data, so intentionally removing certain features in the interest of data privacy may compromise the accuracy of the model. Cathy O'Neil contends that certain data should be removed from training datasets even if it reduces the accuracy of the model (Kehl). However, in some cases, attributes such as race, for example, can be inferred by the model via proxies, or other data values that are correlated with another feature (Farid). As such, removing such data will actually have little impact on the model. What if sensitive data and all its potential proxies were removed from the dataset? Altman explains that such a technique may leave little to no data to train the model, which would render the predictive power very poor. Additionally, it may be necessary that in order to treat certain groups of individuals fairly, taking "protected features" into consideration may be necessary (Altman). Furthermore, enabling the study and subsequent rectification of discriminatory practices in this context may require the use of this data.

Fairness is a virtue that is simple to describe in the abstract but often much more difficult to define and execute in practice. The current gerrymandering debate is demonstrative of this idea. There is a great deal of controversy over how certain district voting lines are drawn and their potential impact on particular election results. Lines can be drawn to favor a certain political party or marginalize a minority group. In the abstract, we certainly should not allow such an inequitable tactic to pervade our political system. However, the Supreme Court has struggled to draw clear lines on this issue. Last year, the Court punted the issue down, but this year *Benisek v. Lamone* and *Rucho v. Common Cause* brought the issue back up. Yet, the Court has still not definitively and actionably addressed the important questions of constitutionality and potential First Amendment violation that arise. What degree of partisan districting constitutes an infringement on the Constitution? How can districts be drawn in a fully equitable manner? Myriad solutions have been proposed for the latter question, but none have achieved any degree of widespread acceptance or implementation. Methods based on isoperimetric quotients, convex polygon ratios, efficiency gap calculations and other mathematical formulations have been developed to tackle this issue. Each model aims to alleviate a particular perceived manifestation of bias, such as jagged borders, oddly shaped districts, or unusual voter proportions; however, each model also comes with a set of inherent tradeoffs. In light of these novel approaches, the same fundamental questions persist. What is our standard of fairness? How can we measure it? Fairness is a virtue to strive for, but in order to do so, we must first formalize our understanding of it and subsequently methodize our approach to move towards it.

In the context of machine learning software and decision making, our present conception of fairness is particularly hazy and ill-defined. Lauren Smith, policy counsel at the Future of Privacy Forum, a DC Think Tank, believes, as a consequence of the constantly updating nature of algorithmic models and general lack of understanding by consumers and policymakers, “it would be difficult to decide how to measure unfairness” and “may leave even companies who care deeply about avoiding discrimination unsure as to what best practices really are” (Atkinson). It is also important to acknowledge that there are inherent tradeoffs between different metrics for fairness. Cornell computer science professor Jon Kleinberg et al. prove that three of the most widely discussed measures of fairness for risk assessment tools are virtually impossible to satisfy simultaneously[2]. In simple terms, these measures of fairness pertain to balancing the error rates

[2] Complete mathematical proof is provided in *Inherent Trade-Offs in the Fair Determination of Risk Scores* by Jon Kleinberg, Sendhil Mullainathan and Manish Raghavan.

between different groups of individuals (Altman). MIT researcher Chelsea Barabas and other scholars of the Berkman Klein Center acknowledge that there are “intrinsic tradeoffs between fairness and accuracy that are mathematically impossible for any RA took to overcome” (Barabas). Moreover, beyond the already complex and incongruous mathematical formulations of fairness, other factors influence our conception of what decisions are fair and unfair. For example, Dr. Micah Altman and colleagues explored life course analysis[3] as a means to measure the long-term effects an algorithmic sentencing decision can have on an individual’s life trajectory. How can we meaningfully adjudicate between the life course analysis approach and other methodologies for measuring the impact of a decision on an individual? Consequently, how do these judgments inform our conception of what algorithmic decisions are fair versus unfair? As such, it is critical that we work towards a consensus of what metrics for fairness are applicable and appropriate in particular contexts where machine learning decision making is used[4].

* * *

I propose the establishment of a federal agency tasked with establishing standards of accountability, transparency, privacy and fairness. As discussed above, this will prove to be a monumental task requiring expertise from many domains. Hannah Fry remarks that establishing a governing body to regulate algorithmic technologies, the relevant intellectual property rights can be preserved while ensuring “that the benefits to society outweigh the harms” (Segal).

Presently, machine learning approaches are widespread enough such that we are beginning to recognize their potency, but not widespread enough such that there is a robust

[3] A methodology for measuring the impact of events on the life trajectory of an individual.

[4] The studies of fairness and reciprocity in economics may provide a useful framework to think about these questions.

framework to ensure their efficacy and benefit. In this burgeoning field, we are presented with the opportunity to discuss the ethical values that are important to us and, quite literally, codify them. The issues raised above are sophisticated and multipronged, but by recognizing and addressing the nuances that arise from a simple framework based on accountability, transparency, privacy, and fairness, we can further elucidate and formalize our society's recognition and comprehension of the most salient ethical and normative values.

Works Referenced

Altman, Micah, Alexandra Wood, and Effy Vayena. 2018. “A Harm Reduction Framework for Algorithmic Fairness.” IEEE Security & Privacy 16 (3) (May): 34–45. doi:10.1109/msp.2018.2701149.

J. Angwin, “Make Algorithms Accountable,” New York Times, August 1, 2016, https://www.nytimes.com/2016/08/01/opinion/make-algorithms-accountable.html.

J. Angwin, et al. “Machine Bias.” *ProPublica*, 6 Mar. 2019, www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Atkinson, Robert, D. “‘It’s Going to Kill Us!’ and Other Myths About the Future of Artificial Intelligence,” Information Technology & Innovation Foundation (June 2016) http://www2.itif.org/2016-myths-machine-learning.pdf

Barabas, Chelsea, Christopher T. Bavitz, Ryan H. Budish, Karthik Dinakar, Cynthia, Dwork, et al. 2017. An Open Letter to the Members of the Massachusetts Legislature Regarding the Adoption of Actuarial Risk Assessment Tools in the Criminal Justice System. Berkman Klein Center for Internet & Society.

Benisek v. *Lamone,* 585 U. S. ___ (2018) *(per curiam)*

J. Buolamwini, “When the Robot Doesn’t See Dark Skin,” New York Times, June 21, 2018, https://www.nytimes.com/2018/06/21/opinion/facial-analysis-technology-bias.html.

Buttar, Shahid. “Cambridge, MA Joins Growing Ranks of Cities Requiring Civilian Control of Police Surveillance Tech.” *Electronic Frontier Foundation*, 20 Dec. 2018, www.eff.org/deeplinks/2018/12/cambridge-ma-joins-growing-ranks-cities-requiring-civilian-control-police.

Doshi-Velez, Finale, and Mason Kortz. 2017. Accountability of AI Under the Law: The Role of Explanation. Berkman Klein Center Working Group on Explanation and the Law, Berkman Klein Center for Internet & Society working paper.

Dwork, Cynthia, and Deirdre K. Mulligan. "It's Not Privacy, and It's Not Fair." *Stanford Law Review*, 1 May 2019, www.stanfordlawreview.org/online/privacy-and-big-data-its-not-privacy-and-its-not-fair/.

Fehr, Ernst and Schmidt, Klaus M., Theories of Fairness and Reciprocity - Evidence and Economic Applications (December 23, 2000). CESifo Working Paper Series No. 403; University of Zurich, IEER Working Paper No. 75. Available at SSRN: https://ssrn.com/abstract=255223

"Global Warming vs. Climate Change." *Global Warming vs. Climate Change | North Carolina Climate Office*, climate.ncsu.edu/edu/DefineCC.

Hessekiel, Kira, Eliot Kim, James Tierney, Jonathan Yang, and Christopher T. Bavitz. 2018. AGTech Forum Briefing Book: State Attorneys General and Artificial Intelligence, May 8-9, 2018, Harvard Law School. Berkman Klein Center for Internet & Society.

E. T. Israni, "When an Algorithm Helps Send You to Prison," New York Times, October 26, 2017, https://www.nytimes.com/2017/10/26/opinion/algorithm-compas-sentencing-bias.html.

D. Kehl, P. Guo, and S. Kessler, "Algorithms in the Criminal Justice System: Assessing the Use of Risk Assessments in Sentencing," Responsive Communities Initiative, Berkman Klein Center for Internet & Society (2017), https://dash.harvard.edu/bitstream/handle/1/33746041/2017-07_responsivecommunities_2.pdf.

J. Kleinberg, S. Mullainathan, and M. Raghavan, “Inherent trade-offs in the fair determination of risk scores,” Proceedings of Innovations in Theoretical Computer Science (ITCS), 2017.

Laskar, M. N. U., Giraldo, L. G. S. & Schwartz, O. Correspondence of Deep Neural Networks and the Brain for Visual Textures. 1–17 (2018).

S. T. Levin, “Imprisoned by algorithms: the dark side of California ending cash bail,” Guardian, September 7, 2018, https://www.theguardian.com/us-news/2018/sep/07/imprisoned-by-algorithms-the-dark-side-of-california-ending-cash-bail.

Rucho v. *Common Cause*, ___ U. S. ___ (2019)

Segal, Michael. “We Need an FDA For Algorithms - Issue 66: Clockwork .” *Nautilus*, 1 Nov. 2018, nautil.us/issue/66/clockwork/we-need-an-fda-for-algorithms.

Shepard, Steven, et al. “Supreme Court Weighs Crackdown on Gerrymandering.” *POLITICO*, 26 Mar. 2019, www.politico.com/story/2019/03/26/supreme-court-gerrymandering-1233319.

TEDxTalks. (2018, October 2). *Hany Farid: The danger of predictive algorithms in criminal justice* [Video file]. Retrieved from https://www.youtube.com/watch?v=p-82YeUPQh0

Waldman, Paul, and Greg Sargent. “A Court Just Dealt a Blow to Rigged Elections. It Probably Won't Last.” *The Washington Post*, WP Company, 3 May 2019, www.washingtonpost.com/opinions/2019/05/03/court-just-dealt-blow-rigged-elections-it-probably-wont-last/?utm_term=.f15a41c1cc1f.