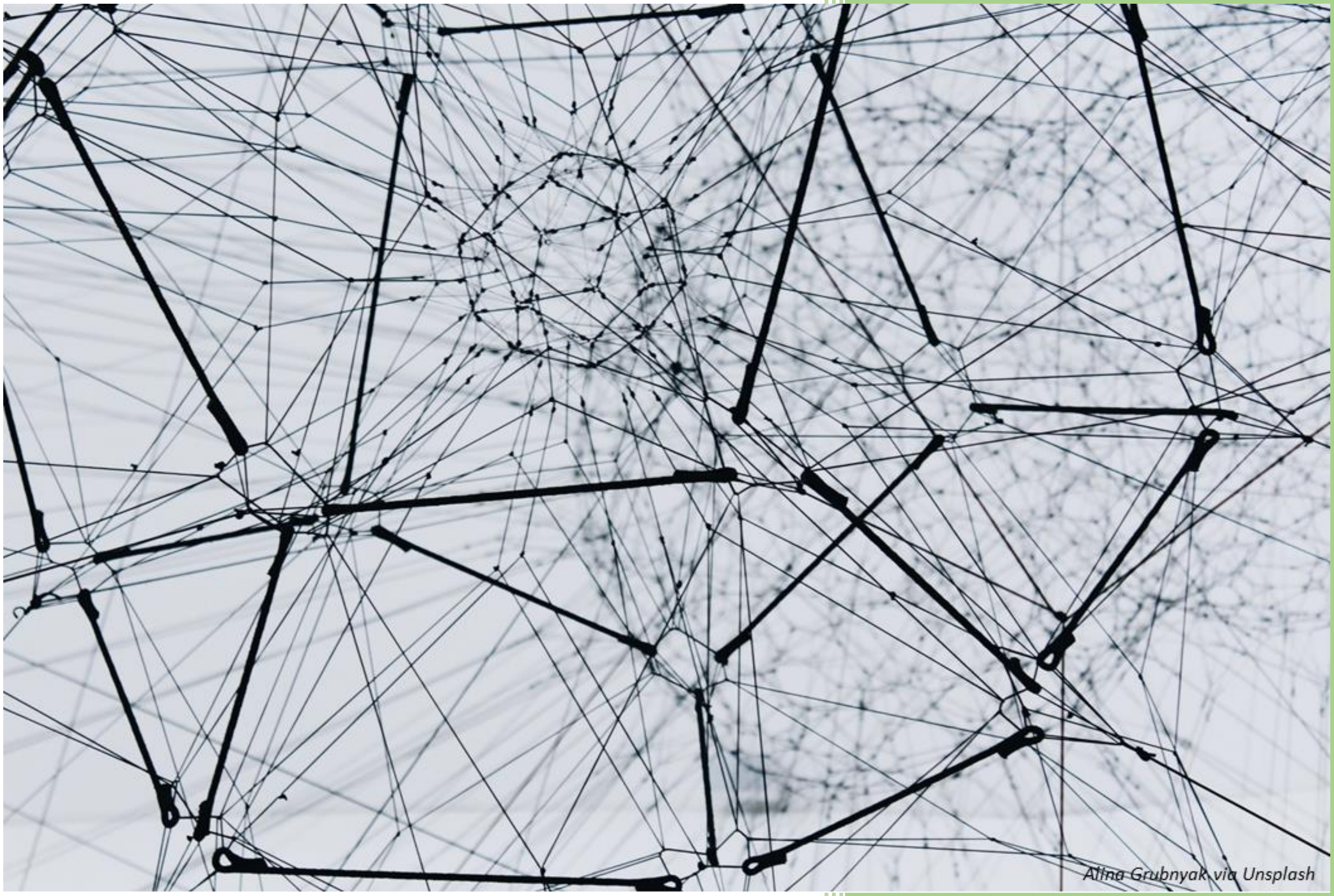
Algorithmic Bias in Machine Learning



Alina Grubnyak via Unsplash

September 19-20, 2019
Duke University, Durham, NC

Hosted by Duke Forge

Sponsored by the Gordon and Betty Moore Foundation

Table of Contents

Executive Summary

Algorithms, essentially computer programs either instructed or trained to perform tasks, are increasingly being used in healthcare. In many cases, they are being used to help clinicians assimilate the high volumes of data now seen in healthcare in support of clinical decision making. Though it would seem a computer program would not exhibit bias, it is increasingly clear that algorithms often incorporate the conscious and unconscious biases of their creators or the data on which they are trained. This introduces the possibility that by using them, algorithms will cause clinicians to care for subpopulations of patients inequitably. With funding from the Moore Foundation, Duke Forge hosted a conference of experts to discuss algorithmic bias and its implications in healthcare and regulation.

Broad Themes

Over the course of the conference, the following major themes, as well as specific points for further exploration, emerged from the discussion:

Identifying Motivations, or the "Objective Function"

- When discussing bias in algorithms it is important to consider the motivation for using an algorithm. For example: if the goal is solely profit maximization, the users may not be concerned with mitigating ethnic bias.
- Therefore, the objective of an algorithm should not only include increasing the efficiency of healthcare delivery, but also normative considerations, such as treating populations equitably.

Regulatory Implications

- With regulators focused on safety and efficacy, and with algorithmic bias affecting both of these concerns, regulators must have insight into how an algorithm is "formulated," analogous to how a device or drug is manufactured and tested.
- There are not yet any consensus standards for "Good Algorithmic Practice" equivalent to FDA-mandated Good Manufacturing Practice or Good Laboratory Practice. It is likely that this will be necessary for regulators, and that these standards should incorporate identifying bias in algorithms as an element of good practice.

The Computer Science of Fairness

- Increasingly, the computer science community is confronting the issue of bias in algorithms.
- There are many metrics and much debate on fairness in algorithms. Some researchers have demonstrated that it is impossible for an algorithm to simultaneously satisfy multiple fairness metrics. Therefore, it is imperative that the community's focus include not only the development of algorithms, but how they are applied when faced with such constraints.
- A fertile area of research is incorporating normativity—desired social goals—into algorithms.

Legal Implications

- There is a tension between the desire for more representative data for algorithms to learn from and privacy.
- Legal frameworks for balancing the benefits and risks of more representative data are currently immature.

- Another source of bias emerges from the fact that well-resourced regions and health systems are more likely to benefit from the beneficial effects of algorithms than under-resourced regions or hospitals, both because of the data available to train algorithms and the technology infrastructure to implement them.

A Societal Perspective

- Corporations that develop algorithms can feign "strategic ignorance," side-stepping the implications of their algorithms treating people inequitably.
- While a community of computer scientists has made forays into the moral and ethical implications of technology (e.g., the Association for Computing Machinery's Conference on Fairness, Accountability, and Transparency [FAccT]), computer science and the technology sector have not had their "day of reckoning" regarding the potential negative social consequences of algorithms.
- There is a need for public education on this topic, but the segments of society best suited to provide this education—whether the media, academia, and/or other institutions—have yet to be determined.

Next Steps

Conference attendees identified the following key points as priorities for ongoing work to build on this and other efforts:

- Developing a consensus over Good Algorithmic Practices and the infrastructure and data science culture that supports such practices;
- Developing a regulatory workforce that is facile with healthcare and machine learning, and considering whether independent third party organizations can supplement this workforce; and
- Because algorithms will inevitably behave differently in diverse, real-world environments, consider creating a model analogous to the FDA's Sentinel System, in which regulatory clearance or approval also requires that real-world data be collected in a central regulatory repository.

Background

The "Algorithmic Bias in Machine Learning" conference, hosted by Duke Forge (Duke University's Center for Health Data Science), was held on September 19-20, 2019 at the J.B. Duke Hotel on Duke University campus in Durham, North Carolina. This symposium represented an effort to extend work previously funded by the Gordon and Betty Moore Foundation, namely the "Human Intelligence and Artificial Intelligence Symposium" conducted at Stanford University (April 2018) and the "Regulatory Oversight of Artificial Intelligence & Machine Learning" meeting sponsored by Duke's Robert J. Margolis, MD, Center for Health Policy. The overarching purpose of the symposium was to move concretely toward a practical framework for evaluating artificial intelligence and machine learning applications in the context of use in health and healthcare.

Artificial intelligence and machine learning have undisputed potential for distilling large bodies of data into clinical action. There are compelling justifications for the use of these technologies in healthcare: data sources such as genomics (and other –omics), social data, socioeconomic variables, and streaming data from wearable devices all can yield data directly relevant to human health. Currently, however, many clinicians are overwhelmed even by the volume of "conventional" health data encountered in typical electronic health records (EHRs). Machine learning (ML) has the potential to bridge this gap by condensing large, complex, and multilayered datasets into actionable insights that will free clinicians to maximize the utility of their time and increase the quantity of high-quality data suitable for research and for informing decision-making about health and healthcare by patients, clinicians, administrators, policymakers, and the public.

However, algorithmic applications have a substantial and demonstrated capacity for encoding and propagating biases, whether inadvertently or intentionally. The social cost of bias incorporated into machine learning applications in healthcare in particular can be clearly seen in the case of a widely used medical algorithm that consistently misclassified the severity of illness in Black patients, leading to systematic undertreatment.[1] Algorithm developers, regulators, and ultimately clinicians, patients, and the public would all benefit from a structured approach to identifying, evaluating, and countering bias in algorithmic products with clinical or health-related applications.

To this end, we convened a conference of experts that included a former U.S. Food and Drug Administration (FDA) Commissioner, representatives from the FDA, a journalist, computer scientists, experts in the law and ethics of algorithmic applications, quantitative experts, and clinicians to engage in exploratory work that would support the development of a reference architecture for evaluating bias in algorithms—one that could potentially be used by the scientific community and regulatory bodies for vetting algorithms used in healthcare.

[1] Obermeyer Z, Powers P, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science. 2019;366(6464):447-453.

Funding Statement

This Algorithmic Bias in Machine Learning conference was hosted by Duke Forge (Duke University, Durham, North Carolina) and supported by a grant from the Gordon and Betty Moore Foundation.

Conference Participants

Hannah Campbell	Program Coordinator, Duke Forge
Robert M. Califf, MD	Founding Director, Duke Forge; Vice Chancellor for Health Data Science, Duke University School of Medicine; Advisor, Verily Life Sciences
David Carlson, PhD	Assistant Professor of Civil & Environmental Engineering, Duke University
Tina Eliassi-Rad, PhD	Associate Professor of Computer Science, Northeastern University
Sidney Fussell	Staff Writer, *The Atlantic*
Erich S. Huang, MD, PhD	Director, Duke Forge; Director, Duke Crucible; Assistant Dean for Biomedical Informatics
Jonathan McCall, MS	Communications Director, Duke Forge
Andrew Olson, MPP	Associate Director of Policy Strategy & Solutions, Duke Forge
W. Nicholson Price II, PhD, JD	Assistant Professor of Law, University of Michigan
Arti K. Rai, JD	Elvin R. Latty Professor of Law, Duke University
Jana Schaich Borg, PhD	Assistant Research Professor, Social Sciences Research Institute, Duke University
Heike Sichtig, PhD	SME & Team Lead, Digital Health, U.S. Food & Drug Administration
Christina Silcox, PhD	Managing Associate, Duke-Margolis Center for Health Policy
Xindi Wang	Doctoral candidate, Northeastern University

Introduction & Overview

The impetus for the Algorithmic Bias in Machine Learning conference, hosted by Duke Forge, grew out of conversations that centered on the increasing excitement in the world of medicine about the potential for artificial intelligence (AI) and machine learning, the prevailing puzzlement about why its use has yet to permeate clinical practice (other than some relatively simple linear equations), and concerns about the potential for algorithmic technologies to introduce or exacerbate harmful biases.

Considered in this light, it would seem to be a useful exercise to tease out the implicit ideas and expectations surrounding the use of AI/ML and the deployment of clinical algorithms, and make them explicit. Such a task requires creating thoughtful definitions for basic terms in the context of patient care and health sciences research: What constitutes an algorithm? What is bias, and how do we represent it? If we aim to correct bias, what does "fairness" mean in this context?

The ultimate goal of such an exercise must go beyond merely "doing the math." All stakeholders involved in the creation and deployment of algorithmic technologies in health and healthcare must give serious thought to the creation of objective metrics, as well as how best to intervene on them. All of this will require going beyond purely quantitative or mechanistic systems, and immediately poses a number of complications, such as the following:

- Any algorithm requires inductive bias in order to recognize output that it hasn't "seen"; in other words, "An **inductive bias** allows a learning algorithm to **prioritize one solution (or interpretation) over another**, independent of the observed data."[2]
- "Crowdsourcing" moral decisions can be extremely problematic, as popular instincts about what is fair or just may create profound ethical problems (for example: what many people might think about whether a given person "deserves" to receive a donated organ).
- What role should empathy play in the creation of algorithms? Is asking "would I want my own algorithms used on me?" a useful question?
- Do we presently have the right tools to integrate these ethical and moral issues into the development of algorithms, and then to evaluate their outcomes?

Keynote Address and Charge to the Conference

The current trajectories of several trends in U.S. health are alarming.[3] There are marked continuous declines in life expectancy and growing geographical segregation of health outcomes.[4,5] There is also the question of what issues tend to dominate the discussions about health and healthcare, and the results of that focus. The recent furor over efforts to reduce readmissions for heart failure provides an example: preventing readmissions for heart failure helps to save money under managed care systems, but in

[2] https://arxiv.org/pdf/1806.01261.pdf

[3] National Center for Health Statistics. Centers for Disease Control and Prevention. Health, United States: 2018 (Chartbook). Available at: https://www.cdc.gov/nchs/data/hus/hus18.pdf#Chartbook. Accessed January 29, 2020.

[4] Chokshi DA. Income, poverty, and health inequality. JAMA. 2018;319:1312-1313.

[5] Dwyer-Lindgren L, Bertozzi-Villa A, Stubbs RW, et al. Inequalities in Life Expectancy Among US Counties, 1980 to 2014: Temporal Trends and Key Drivers. JAMA Intern Med. 2017;177(7):1003-1011. doi:10.1001/jamainternmed.2017.0918

some instances focusing on reducing heart failure readmissions has been accompanied by increases in deaths due to heart failure.[6] Although this controversy is not itself reflective of algorithmic bias, it does point out the underlying complexities that can affect the answers to even relatively straightforward questions.

As part of his keynote address for the meeting, Robert M. Califf, MD, (Duke Forge, Verily Life Sciences) noted that the Duke University was home to an early example of using computers and algorithms to support diagnosis and shared clinical decision-making. In the early 1970s, Duke cardiologist Eugene Stead began developing a database to track cardiovascular outcomes, one that incorporated lifetime follow-up for all patients treated at the Duke University Medical Center. The impetus for this database was the realization that doctors were not capable of assimilating and synthesizing all of the relevant data needed to guide patient care. The resulting output from this data collection—a cardiovascular "prognostigram"—provided a probability score for whether a patient was likely to benefit more from medical treatment versus bypass surgery.

However, despite this pioneering example, the approach did not spread widely, largely due to structural issues with the provision of healthcare.

Duke is just one place that has created a data pipeline to which algorithms can be (and are being) applied. In the United Kingdom, a Google Deep Mind algorithm for acute kidney injury is poised to be introduced nationwide through the National Health Service. The view of regulators, as expressed in 2016, is that because algorithms are constantly being refined and updated – and for that matter, the inputs themselves are constantly changing[7]— such technology requires approaches to evaluation for risk and benefit that go beyond those used for more traditional medical technologies.

Given that current regulatory paradigms are unsuited to the task of evaluating clinical algorithms, the only viable option is to regulate the entities that create the algorithms—a philosophy that led to the creation of the FDA's Digital Health Software Precertification Pilot Program. Under the proposed precertification pathway, regulators will examine systems and receive assurance that they are adequate. The companies and other entities creating the algorithms will be able to defer some of the premarket review requirements until the postmarket interval and will be required to report real-world analytic data to the FDA periodically, while also remaining subject to audit. The risk of such a paradigm is that people can cheat; further, such cheating is hard to detect, meaning that regulators will have difficulty in knowing when and how to intervene. Cheating might take the form of an organization portraying itself as following best practices with algorithm development and not doing so in reality, or manipulating post-market surveillance data.

In addition, the orientation of the regulators themselves can be an issue. Regulators who lack sufficient specialized knowledge can create problems, as can regulators who are too "friendly" with industry. In addition, precertification considers purely administrative algorithms to be of little or no risk, but given the potential for bias and perverse incentives to act upon them, they may actually be the riskiest of all,

[6] Wadhera RK, Joynt Maddox KE, Wasfy JH, et al. Association of the Hospital Readmissions Reduction Program With Mortality Among Medicare Beneficiaries Hospitalized for Heart Failure, Acute Myocardial Infarction, and Pneumonia. JAMA. 2018;320(24):2542-2552. doi:10.1001/jama.2018.19232.

[7] Price NII WN. Regulating black-box medicine. Michigan Law Review. 2017;116(3):421-474. Available at: http://michiganlawreview.org/wp-content/uploads/2017/12/116MichLRev421_Price.pdf.

when we consider that all health systems function in ways that are biased toward serving people who can make money for the system.

We are now at moment when Google's search engine fields roughly a billion questions about health each day and access to enormous amounts of health data—whether accurate or not—is in almost everyone's hands as smartphones have become ubiquitous. Yet at the same time, large health systems are purchasing and deploying algorithms with no real knowledge of how they actually work.

This raises general questions about the use of algorithms in health that are pertinent to algorithmic bias because algorithms unavoidably have societal impact:

- What is the objective function for health systems?
- How we are incentivized to achieve it?
- How are algorithms reinforcing it?
- Is the objective function of the algorithm premised on maximizing profit?
- Are we measuring utility at a societal level?
- Regardless of what we want to measure, are we able to quantify it?

Policy Considerations

The FDA's current thinking on the regulation of artificial intelligence/machine learning software products for healthcare is outlined in a discussion paper/request for feedback that considers such algorithms under the "Software as Medical Device" (SaMD) framework used by the agency.[8] It should be noted that issues raised by the regulation of such products are inherently complex, and that the discussion paper cited above represents a work-in-progress that will evolve over time.

When considered through the regulatory lens, "bias" has the working definition of "a systematic deviation from truth," and "algorithmic bias" can be defined as "systematic prejudice due to erroneous assumptions incorporated into the AI/ML" that is subject to regulation under the SaMD framework.

Bias can be introduced at multiple points during the lifecycle of the algorithm: during design, training, and testing. The bias itself can stem from factors such as:

- The intended use of the SaMD;
- Non-representative training, validation, and test data sets;
- Bias in the selection of training, validation and test data sets (e.g., clinical labels); and
- Introduction of bias during data preparation (selection of attributes).

The kinds of bias that may impact the testing and evaluation of a SaMD algorithm include:

- Selection bias, where the sample of subjects is not representative of the target population;
- Spectrum bias, where the sample of subjects studied does not include a complete spectrum of the target population;

[8] US Food and Drug Administration. Proposed Regulatory Framework for Modifications to Artificial Intelligence/Machine Learning (AI/ML)-based Software as a Medical Device (SAMD). Available at: https://www.fda.gov/files/medical%20devices/published/US-FDA-Artificial-Intelligence-and-Machine-Learning-Discussion-Paper.pdf. Accessed January 28, 2020.

- Verification bias, in which 1) only some of the intended subjects undergo the reference standard test or 2) some of the intended subjects undergo one reference test and others undergo another reference test; and
- Automation bias, created by the use of automation as a heuristic replacement for vigilant information-seeking and processing.

One of the chief dangers that characterizes bias in training sets is that its presence may be difficult to discern unless special attention is paid. If it is not detected, the result can be "invisible inequity" that is incorporated into the algorithm. The incorporation of bias may occur for a variety of reasons, including:

- Data may not be equally readily available for all groups;
- There may be a greater proportion of missing or fragmented data for a particular group or population;
- There may be fewer patient-reported outcomes for a particular group or population; and
- Vulnerable populations are at inherently higher risk (vulnerable populations may include persons who are economically and/or socially disadvantaged; racial and ethnic minorities; and pregnant women)

A valuable step in countering algorithmic bias would entail developers preemptively responding to a key set of questions[9] to guide development of a given SaMD product. Such questions include:

- What kinds of bias might exist in your data?
- What have you done to evaluate whether your training data are biased, and how might those biases affect your model?
- What are the possible risks that might arise from biases in your data, and what steps have you taken to mitigate these biases?
- What bias might remain, and how should users take remaining biases into account?
- Is your method of ground truth labeling appropriate to the clinical use case you are trying to resolve?

The FDA is actively engaged in building out its capabilities for evaluating safety and efficacy of algorithmic technologies through its newly created Digital Health Center of Excellence and has been creating an array of guidances for digital health products.[10]

Group Discussion

- **Is "representative" data actually desirable in all of these contexts**, given that the actual amounts of data available within a given population may be too small? Would an over-represented or over-sampled data set in some cases be more desirable in terms of reducing or eliminating bias? "Representative" may in fact be a vague and unhelpful term when thinking about suitability of training data—"sufficient" may be a better way to conceptualize this.

[9] Adopted from a draft of Ethics of AI in Radiology: European and North American Multisociety Statement (October 1, 2019). Available at: https://pubs.rsna.org/doi/full/10.1148/radiol.2019191586

[10] US Food and Drug Administration. Guidances with digital health content (updated September 27, 2019). Available at: https://www.fda.gov/medical-devices/digital-health/guidances-digital-health-content

- **The machine learning community often makes assumptions about the underlying distributions within data that may not be accurate.** We may need to devote more thought to the sources and processes that yield the data that are used to train machine learning models (or that the machine learning model will be applied to).
- Even in cases where it may seem that sufficient data exists to represent subpopulations when developing the algorithm, **those training data do not fully represent the kinds of real-world data** that the algorithm will eventually consume.

The key questions are then:

1. What kinds of data are needed?
2. How much is enough?
3. Should algorithms be labeled with "indications" that specify the populations in which the algorithm was trained?
4. Could a default toward approving algorithmic applications with only a narrow "indication" that reflects the data used to develop it create an incentive for developers to create more generalizable models?

- **Obtaining data is relatively easy, but at present, establishing "ground truth"[11] is nearly impossible.** Any system that actually measures the outcome of interest is enormously valuable.
- **Another point of concern is vendors selling algorithmic products whose inner workings are entirely inscrutable** (i.e., "black box" algorithms).
- **There may be an analogy to laboratory developed tests (LDTs),** which are not subject to FDA jurisdiction if they are used only at the center or laboratory that created them. However, it seems likely that if algorithms developed at one health system are exported to others, they will need to be modified to ensure that they work properly in the new environment. However, at present, it is unclear whether there is any regulatory framework to cover this scenario.

The Perspective from Computer Science

Machine learning itself is not a new discipline: the term was first coined by Arthur Samuel in 1959 to describe his checkers-playing program.[12] By 1997, Tom Mitchell had defined the "well-posed learning problem" in artificial intelligence, which has three components: task, experience, and performance.

> *A computer program is said to learn from experience* E *with regard to some task* T *and some performance measure* P *if its performance on* T*, as measured by* P*, improves with experience* E.
>
> *--Tom Mitchell (1997)*

Two common tasks to which machine learning algorithms are applied are to 1) assess risk or 2) to rank things. These two tasks are popular because of their societal prevalence, and because computer scientists know a lot about these tasks. Fundamentally, human decision-makers like (or find useful) the output of these tests, which typically are easy to grasp.

[11] That is, empirical evidence established by direct observation and measurement, as opposed to inference.

[12] Samuel AL. Some Studies in Machine Learning Using the Game of Checkers. IBM Journal of Research and Development. 3:3, 1959, pp. 210–229.

However, there are a number of issues related to the use of these tools. One of these constitutes so-called "impossibility results," in which it is not possible to simultaneously satisfy three desirable fairness measures: precision parity, true positive parity, and false positive parity. Another factor is that in the case of bin categories for risk, classifications are too abstract to incorporate ethical or normative considerations, nor do they have uncertainty values associated with them. In the widely publicized case of algorithmic software used in judicial sentencing that was shown to be biased,[13] the risk categories did not include a measure of confidence or uncertainty, nor any broader context. They also do not capture the processes, values, or incentive structures of human decision-making.

This raises a key question: **What are the possible incentives and values of a human decision-maker?** Is it one or more of the following?

- Profitability—Does using an algorithm increase financial gain?
- Efficiency—Does an algorithm streamline process?
- Accuracy—Is an algorithm accurate?
- Reliability—Is an algorithm more reliable than humans?
- Fairness—Does an algorithm perform equally well on subpopulations?
- Interpretability—Is it possible to understand how inputs into an algorithm change the outputs (e.g., understanding how much a car turns in response to steering wheel inputs)?
- Explainability—Is it possible to express how an algorithm works in human terms (e.g., the mechanics of a rack-and-pinion steering system)?

Further, there is an important distinction between explainability and interpretability.[14] Many popular networks are black-box models whose inner workings can be explained via inference but not directly examined, as would be the case with an interpretable model.

Apprenticeship Learning

Another approach to creating ML algorithms employs *apprenticeship learning* (also called *imitation learning*) that trains the algorithm by modeling a human exemplar.[15] Although these approaches exist, they are much more difficult than the risk classification assessment and ranking discussed previously, and have typically not been carried out in healthcare/medical settings. Further, there are a number of concerns from a philosophical/ethical perspective, such as how can one know which variables the exemplar attends to, the reliability of self-reporting, and the variation between simulated situations and "real life." In addition, reproducibility, especially across different settings, remains a major problem in ML, as does liability in the event of a bad ML-driven outcome.

[13] Angwin J, Larson J, Mattu S, Kirchner L. Machine Bias. Propublica. May 23, 2016. Available at: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing. Accessed January 27, 2020.

[14] O'Rourke K. Explainable ML vs interpretable ML. Statistical Modeling, Causal Inference, and Social Science. October 30, 2018. Available at: https://statmodeling.stat.columbia.edu/2018/10/30/explainable-ml-versus-interpretable-ml/. Accessed January 27, 2020.

[15] Abeel P, Ng AY. Apprenticeship learning via Inverse Reinforcement Learning. Proceedings of the 21st International Conference on Machine Learning, Banf, Canada, 2004.

Another key question is whether ML is actually learning for the defined task – i.e., if the results the algorithm is producing are actually being driven by the questions or variables of interest.[16] A related problem has sometimes been referred to by Joy Buolamwini as the "undersampled majority,"[17] which encapsulates the idea that the global population is not mostly white males, but most training data related to humans and human activities are derived from white males. When developing algorithms, we must understand how to represent implicit and explicit biases in complex networks.[18]

Performance Measures

There are more than 30 competing definitions of "fairness" in computer science, but none are sufficiently expressive to represent normative situations; i.e., what *should* happen in a given circumstance, and this remains a somewhat neglected area. There is also work to be done in clarifying the optimal balance between accuracy and intelligibility and the appropriateness of deploying black-box models.[19]

In summary, we must start giving serious consideration to: 1) incorporating normativity (i.e., specifying outcomes stemming from the use of an algorithm that would be desirable or undesirable) throughout the development and testing process; 2) weighing incentives and values; 3) investigating the potential of apprenticeship learning; and 4) discussing the circumstances in which ML should be used at all.

Group Discussion

- **Is a biased algorithm worse than a biased human?** Machine learning can produce a certain amount of bias over a large distribution, but in many cases, the messiness and variability of real-world, ground-truth data may actually be more problematic (consider the example of the challenges that emerged in training deep learning algorithms to diagnose retinal disease in clinical settings[20]). The problem with placing excessive weight on the issue of possible bias is that it does not take into account the way decisions are actually made now in healthcare. One question to consider is whether imperfect algorithms are better than current practice—that is to say, the algorithm may be flawed or biased, but that still may be preferable to the decision-making of biased human. A complementary question would address the scale of potential harm due to bias: a widely-adopted biased algorithm could affect far more people than a single biased individual.
- **Conceding agency to an algorithm is problematic.** While biased models might be better than biased humans in some cases, we would still lack knowledge about how the model was trained,

[16] Emspak J. How a machine learns prejudice. Scientific American. December 29, 2016. Available at: https://www.scientificamerican.com/article/how-a-machine-learns-prejudice/. Accessed Janury 27, 2020.

[17] Buolamwini J. Joy Buolamwini: examining racial and gender bias in facial analysis software. Barbican Centre. Available at: https://artsandculture.google.com/exhibit/joy-buolamwini-examining-racial-and-gender-bias-in-facial-analysis-software-barbican-centre/LgKCaNKAVWQPJg?hl=en

[18] Johnson G. Cognition and the structure of bias. Doctoral dissertation. Available at: https://escholarship.org/content/qt7hf582vz/qt7hf582vz.pdf.

[19] Caruana R. Friends Don't Let Friends Deploy Black-Box Models: The Importance of Intelligibility in Machine Learning. Presented at: KDD '19: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. July 2019, 3174. https://dl.acm.org/doi/10.1145/3292500.3340414

[20] Ting DSW, Pasquale LR, Peng L, et al. Artificial intelligence and deep learning in ophthalmology. Br J Ophthalmol. 2019;103:167-175.

or what data were used. In such a case, would people still be comfortable applying the algorithm to an individual human being?

- **How will algorithms change over time, and how does that affect how we use them?** If we concede that a (flawed) algorithm is better than the alternative when we deploy it at Time 0, will that still be the case at T2, or T3, at which point the cumulative effects of built-in bias may have increased? This may be an argument against deploying an algorithm to achieve a short-term benefit because the longer-term consequences and overall trajectory is less favorable.
- **What can we do to prepare people to use potentially flawed ML tools?** Even given a thorough and comprehensive availability of information about the inner workings of an algorithm, what kinds of training and information would be needed to use these tools and to interpret and judge their output? It may also be important to have clarity and transparency about the underlying incentives and incentive structures acting upon the groups of people (e.g., hospital administrators or physicians) who are using a given algorithm. Providing measures of uncertainty and confirmation, as well as context tailored to the user's level of knowledge, may also be helpful.
- **Should feedback loops be established** to ensure continuous improvement of process and regulation?

Legal & Ethical Considerations

Bias is present in the data itself. It manifests in training and learning processes, as well as in deployment of an algorithm or application. In medicine, for instance, it is quite possible to have an entirely representative data set that nonetheless reflects biases and inequities inherent in the underlying disease prevalence or care setting. However, biases can be introduced deliberately (in order to maximize profit rather than health outcomes, or as in the case of modern "redlining" practices, to engage in proxy discrimination or optimize an algorithm for other unacceptable outcomes[21]).

Bias and inequity are difficult to separate in terms of their effects. Policy-making, application development, and training often takes place in high-resource settings and reflects those circumstances—a tertiary-care hospital in an academic setting will have different patient populations and care pathways than a lower-resource healthcare setting, and different results may accrue when deploying an algorithm created in the former for use in the latter.

Variation in Available Resources

There are legal system regimes that create incentives for developing these systems and approaches within the confines of resource-rich settings. Chief among these incentives is that it is typically expensive to collect clinical data and harness it for secondary uses. Technical resource requirements, privacy-oriented HIPAA compliance, and gathering informed consent all contribute to costs. Some of these constraints are legal and some are not, but all tend to make this kind of data collection something of a luxury. Further, actually deploying ML products, dealing with tort law, gaining approval from federal

[21] Prince A, Schwarcz DB. Proxy discrimination in the age of artificial intelligence and Big Data (August 5, 2019). Iowa Law Review, Forthcoming. Available at SSRN: https://ssrn.com/abstract=3347959

regulators, or securing payer reimbursement all tend to be easier when the training data are derived from larger, better-resourced medical centers.

Creation of Feedback Loops

The model that implicitly underlies clinical algorithms is a cycle of continuous learning, in which the results created by deploying the algorithm are fed back into it, ostensibly to continuously improve its performance. However, the creation of such a loop can be problematic, in that the loop will provide a false sense of security, particularly if the data reflect underlying biases. If biases in care or in data collection practices continue, an unadjusted feedback loop may paper over those problems.

Restrictions on Data Flow

International restrictions on data flow are another legal consideration when thinking about developing and deploying algorithmic to products, with the recently instantiated European General Data Protection Regulations (GDPR) providing a good example of a legal framework that could limit the use of an algorithm in an international setting.

Tensions in Unbiased Representation

There is a tradeoff between inclusivity and privacy, largely because the acquisition of large and truly inclusive datasets usually requires mandates, and that is not typically consistent with prioritizing individual autonomy and privacy.

We also typically want to see widespread participation in the creation of large, representative datasets, but this is contradicted by the reality of the problematic history of participation in research and clinical trials by under-represented minorities. There are fundamental issues of equity in play when asking people to contribute their data or participate in research to ensure its quality, but then denying those same persons the full benefit of that research because they are continuing to receive suboptimal care due to underlying biases effecting healthcare delivery.

Group Discussion

- **People who need help the most are the least likely to benefit from improvements enabled by data science**, due to issues of access as well as the present convergence of historical and cultural factors. The fundamental problem is that people who do not share their data are going to be left behind and continue to experience worse outcomes.
- **People tend to want explanatory models, but having insight into the algorithm's inner workings may come at the expense of personal privacy.** And not only can continuous learning on biased data tend to reinforce problems, but adversarial learning can potentially involve "poisoning" data for nefarious purposes, as in cases where attacks are mounted by introducing corrupt data into a training dataset.[22]
- **If we are going to make ML more effective and less biased, underlying structures and practices may first have to change.** Consider the current move toward value-based care—this will inform the objective function of any algorithm deployed in that setting. The fee-for-for service

[22] Jagielski M, Oprera A, Biggio B, et al. Manipulating machine learning: poisoning attacks and countermeasures for regression learning. 2018 IEEE Symposium on Security and Privacy. https://ieeexplore.ieee.org/document/8418594

approach itself creates an invitation for algorithmic bias due to the objective functions, such as the imperative for profit preservation and maximization, that it fosters.

- **Incremental improvements in data collection and accuracy (as opposed to structural changes in healthcare delivery) can yield valuable outcomes** that are easier to achieve, even if still challenging.

Recommendations

- Invest in data infrastructural resources to enable more representative and equitable data collection.
- Consider legal changes to lower hurdles to data collection and secondary use to increase representativeness and decrease bias.

Journalistic Perspectives on Artificial Intelligence

How has journalism as a profession tended to approach the issue of algorithmic bias? The topic itself is "cool" and attracts attention from readers. However, an injudicious approach risks misinforming or alarming people. There are three things journalism as a whole has done fairly well in this context:

Contextualizing the Definition of Bias

Over the last few years, stories about algorithmic bias have incorporated not just a numerical or quantitative look at bias, but a more "anthropomorphic" approach that examines how data are being applied and asks whether or not these uses are harmful.[23] There is also an emphasis on what people are afraid of in terms of AI applications; typically, it seems that people in general fear the idea of AI having agency and doing things to them. But while it is positive that journalists are helping people to be clear on the problems of bias and how that works, it may be less good that headlines often make it seem as if "AI" is doing things, *rather than the people who build and use the AI*.

Making Connections across Fields

Journalists are generally good about reaching across disciplines; in fact, journalism is by definition interdisciplinary. Some of the best journalism in this area has resulted from journalists seeking perspectives not just from technical experts and scientists, but also drawing upon knowledgeable viewpoints from history, philosophy, ethics, and the legal profession. Historical context in AI discussions is particularly important, because when historical context is removed, it tends to lead to the idea of AI as something purposeful, something that acts of its own agency—a framing that offers novelty but may convey a distorted view. Journalists have also tended to ask probing questions of people in the world of business and put pressure on corporations for accountability.

Spotlighting Voices Outside of the Margins

Journalism has perhaps distinguished itself most in spotlighting issues of marginalization arising (or that could arise) from the use of AI applications. One example of this can be seen in the evolution of reporting on workplace issues at Google that emerged after revelations that the company was

[23] Murray SG, Wachter RM, Cucina RJ. Discrimination By Artificial Intelligence In A Commercial Electronic Health Record—A Case Study. Health Affairs Blog. January 31, 2020. Available at: https://www.healthaffairs.org/do/10.1377/hblog20200128.626576/full/. Accessed February 21, 2020.

employing "microworker" contractors for projects with direct military applications.[24] Other examples include reporting on the use of YouTube videos created by transgender persons to train facial recognition software[25] and the use of poorly paid, offshore "ghost workers," often women, to support Silicon Valley projects.[26]

Summary

At its best, journalism in this arena can give people the information that lets them understand how these tools are affecting them, and allow them to have a more informed understanding of the surrounding issues. Continuing to "complicate" the story of AI is, on balance, a positive thing. Thus far, journalism has perhaps deserved a "C-plus" or "B-minus" grade on how it has handled the broad topic of algorithmic bias.

Group Discussion

- **A key question is "strategic ignorance" on the part of corporations applying AI tools and the use of that ignorance as a shield** against the possible negative consequences of using an algorithm. It can almost function as a kind of currency in the marketplace. Journalism has an important role in probing how we are with the existence and perpetuation of this ignorance, and what can be done about it.
- **Computer science, as a discipline, has not yet had its "day of reckoning"** in terms of the moral and ethical consequences of its products. However, it has not been definitively established that providing formal training in ethics for engineers will necessarily improve outcomes, and it is possible that frameworks of constraints and regulations could cause them to conceive of these issues as someone else's problem. What approaches can help people internalize and incorporate principles of moral AI development?
- **A further issue to consider is the issue of bias and framing within the press**, and the role that money plays in that process; i.e., how stories are chosen, how headlines are created, how stories are framed for and consumed by the public.
- **Who is responsible for educating the public about issues relating to algorithmic bias?** Is this a job for the media/press, or have universities abdicated their role in this regard?

[24] Kelly M. Google hired microworkers to train its controversial Project Maven AI. The Verge. February 4, 2019. Available at: https://www.theverge.com/2019/2/4/18211155/google-microworkers-maven-ai-train-pentagon-pay-salary. Accessed January 28, 2020.

[25] Vincent J. Transgender YouTubers had their videos grabbed to train facial recognition software. The Verge. August 22, 2017. Available at: https://www.theverge.com/2017/8/22/16180080/transgender-youtubers-ai-facial-recognition-dataset. Accessed January 28, 2020.

[26] Epstein G. How 'ghost work' in Silicon Valley pressures the workforce, with Mary Gray. TechCrunch. August 16, 2019. Available at: https://techcrunch.com/2019/08/16/how-ghost-work-in-silicon-valley-pressures-the-workforce-with-mary-gray/. Accessed January 28, 2020.

Working Session 1: Identifying Good Algorithmic Practices

As we consider how best to move forward with the development, evaluation, and implementation of machine learning algorithms in health and healthcare, we might consider an approach analogous to Good Laboratory Practice (GLP),[27] an approach to ensuring the quality of laboratory practices and results that has been widely (if not universally) adopted throughout the world of clinical care and research and the regulatory bodies that oversee them. Other groups, including universities and professional organizations, are in the process of developing standards for machine learning, but a comprehensive set of standards has yet to be articulated and adopted. It may therefore be useful to consider some elements that should inform standards for machine learning, including:

- Internal validation specifications
- Quality of training data
- Properties of the objective function
- Robustness of optimization algorithms

Data and Regulatory Review

Given that companies working in AI/ML treat training data as highly proprietary, this raises the issue of whether it is possible to develop good practices with regard to training data while at the same time preserving intellectual property. There is also the further question of whether it should be taken for granted that companies working on health or healthcare applications of ML technologies should be allowed to keep training data secret, if we posit a regulatory framework for such applications similar to those used for the approval of new drugs (in other words: disclosure of training data might be made a condition of regulatory approval for a medical algorithm). But is handing regulators a dataset the most efficient way to assess for bias, or could a summary description or explanation be more useful, when coupled with checking for bias in performance data?[28]

At a minimum, disclosure to the FDA in a way that makes it possible for regulators to meaningfully evaluate the algorithm's working seems desirable. (In other words, regulators should be able to "cross-examine" a given algorithm; for this, regulators will need to know what data the algorithm "saw.")[29] However, there remains the question of what the most efficient and effective approach for this would be. Would summary statistics or descriptive analyses suffice, or is more granular detail needed?

One concern with using summary data is that certain critical failure modes would not be revealed by an examination of only the summary data. This can be seen in a recent example in which a deep neural network appeared to be successfully analyzing radiological images, but was actually cueing on labels

[27] OECD Principles of Good Laboratory Practice (GLP) and GLP Compliance Monitoring. Available at: https://www.oecd.org/chemicalsafety/testing/overview-of-good-laboratory-practice.htm. Accessed February 27, 2020.

[28] Gebru T, Morgenstern J, Vecchione B, et al. Datasheets for datasets. Preprint available from arXiv.org at: https://arxiv.org/abs/1803.09010. Updated January 14, 2020. Accessed February 21, 2020.

[29] Mitchell M, Wu S, Zaldivar A, et al. Model cards for model reporting. Preprint available from arXiv.org at: https://arxiv.org/pdf/1810.03993.pdf. Accessed February 21, 2020.

placed in the margin of images taken from portable x-ray machines.[30] Final performance testing should be done either with the developer's set of training data or with a completely different set of data (not a subset of training data), taking care to exclude as many aspects of the image or accompanying "metadata" as possible that could be introducing bias. However, the baseline for evaluating model performance should not be "no variability," considering that variability among physicians is likely higher than what would be seen from an algorithm with different sets of data.

How Can Regulators Evaluate Machine Learning Applications?

Could the FDA mandate an entirely reproducible process for machine learning applications? A good researcher would supply variance measures that indicate how stable an ML application is. Hardware and software, as well as other variables, can all play a role in the output of ML algorithms. For the FDA to evaluate an algorithm, the agency would need to have key capabilities in place, including compute containerization and the ability to generate independent reference data sets.

The question of what "reproducibility" actually means is an active discussion in the world of machine learning. At present, there is no clear, agreed-upon definition. In the case of a pretrained model, the question of how the model was trained may be of secondary importance compared with the question of whether the model works or does not work on real-world data (i.e., whether the model's output is generalizable). Regardless, there should be either transparency around the training process or performance on independent, and ideally real-world, data.

Reviewing a pretrained model and running it on new data that the model has not "seen" before may pose challenges for regulators (since regulators do not currently possess their own validation datasets), given just how much of data in general is not interoperable, but at the same time, using a single-institution data set would raise data provenance issues. Should the model's creator specify what data "pipes" will work with that model? Or is it more likely that health systems will stick with the ML equivalent of laboratory-developed tests and use "home-brewed," non-transferrable applications? Are the data needed to test someone else's algorithm readily available? Could a generative adversarial network (GAN) offer a possible option for testing?[31]

Transfer Learning and Modification of "Stock" Algorithms

How would evaluation of transfer learning applications work under a regulatory schema? For an example of transfer learning, imagine a neural network that was originally trained using a national dataset and then brought to a particular institution, where the topmost layers were trained with local data—a process analogous to purchasing a suit "off the rack" and having it fitted by a tailor. Does modifying a product in such a fashion create legal or liability issues? What would the regulatory approach be?

[30] Zech J. What are radiological deep learning models actually learning? July 8, 2018. Available at: https://medium.com/@jrzech/what-are-radiological-deep-learning-models-actually-learning-f97a546c5b98.

[31] Lin Z, Jain A, Wang C, et al. Generating high-fidelity, synthetic time series datasets with DoppelGANger. September 30, 2019. Preprint available from arXiv.org at: https://arxiv.org/pdf/1909.13403.pdf. Accessed February 21, 2020.

A Role for Continuous Monitoring of Algorithmic Performance

Is there a way to have a marketed algorithm that has been deployed in a real-world, working environment "report back" performance metrics to a central post-market surveillance repository, particularly as users adjust parameters to accommodate local conditions? Contrariwise, if the algorithm is not modified, what happens if there are changes in the underlying populations, or new treatments become available? Would continuous measurement of real-world performance constitute a key component of good algorithmic practice?

One analogy to such continuous modeling might be the lifecycle management approach used by the FDA as part of its National Evaluation System for health Technology (NEST) program, in which medical devices that undergo constant refinement are monitored for safety and performance. Such an approach would require some way to delineate when a given algorithm is succeeding and when it is failing.

If an algorithm is being applied in different contexts and succeeding in some but failing in others, that is valuable information that should inform use. This could start with demographic and co-morbidity subgroup performance data. While regulatory bodies would have some responsibility for monitoring the appropriate real-world deployment of applications, so do the users (i.e., health systems) of those algorithms.

Regulatory Paradigms: Labeling and REMS

Another potential regulatory issue concerns whether a framework for defining the limits of "off-label" use of an algorithm (i.e., used in a fashion other than indicated by the formal regulatory approval) should be developed. Given the leeway in the current system that permits physicians to use medications off-label, it is questionable whether this would curb particular kinds of use absent additional controls. One solution might be the creation of registry for every predictive model deployed in clinical settings, combined with follow-up by regulators over the product lifecycle. The FDA Center for Drug Evaluation and Research's Risk Evaluation and Mitigation Strategies (REMS) program might offer a better paradigm for health and healthcare algorithms. Health systems also can be assumed to have a measure of responsibility for conducting their own evaluations when adopting an algorithmic tool developed outside of their own institution.

It is possible that a case can be made for the "labeling" paradigm based on the actual practical (as opposed to strictly legal) effects of product labeling. Federal strictures on labeling of therapeutic products have follow-on effects; they limit what manufacturers can claim about how a product works, for instance. But even though clinicians are able to prescribe medical therapeutics "off-label," i.e., for conditions other than the label specifically supports, that labeling nevertheless does affect practice and helps form a consensus around the appropriate clinical uses of a given product. However, this paradigm may be less useful for providing point-of-use information in a clinical context. For this reason, adopting a quasi-REMS model for algorithmic products may be more apt.

Other Oversight Options

Consideration should be given to the question of whether other federal agencies such as the Centers for Medicare and Medicaid Services or the Federal Trade Commission potentially have oversight related to the use of algorithmic products in healthcare. Another possible approach might involve an independent, highly trusted private organization for evaluating ML applications similar to the Underwriters

Laboratories (UL), which tests multiple consumer products and appliances for safety.[32] One challenge to such an approach—one involving an independent, private, third-party entity—revolves around issues regarding the accessibility of data, which might have to be brokered by an agency like the FDA. A broader challenge would derive from the need for retraining, recalibrating, and certifying algorithms in a setting where applications must be extensively customized to the local conditions. Given these challenges, real-world registries and surveillance may be the best way to capture performance information.

Examining Bias as Part of Performance Evaluation

There are existing toolkits available that provide background information and tools for identifying, preventing, and countering algorithmic bias in software products. These include IBM's open-source AI Fairness 360 toolkit,[33] Google's compilation of Machine Learning Fairness Resources,[34] and the University of Chicago's open-source Aequitas toolset for auditing for machine-learning bias.[35] There are also regulatory requirements and guidance from the world of clinical trials (whether conducted under the auspices of FDA or the National Institute of Health) that deal with collecting data that may have bearing on bias, as well as effects on under-represented populations.

In terms of resources for thinking about continuous collection of real-world performance data from algorithms, the FDA has created a webpage that includes public information about the agency's current thinking on the topic to date and how to address it in the future. A public workshop may be the best pathway for further discussion.

Working Session 2: Metrics

Defining and Measuring Fairness

Metrics for fairness (whether defined in terms of individual or group fairness, or in terms of other classifications[36,37] in machine learning applications are currently being developed[38] based on a confusion matrix of actual versus predicted results. If the goal is to create risk measures, there cannot be parity

[32] Underwriters Laboratories (UL). Our mission. Available at: https://www.ul.com/about/mission. Accessed February 21, 2020.

[33] IBM. AI Fairness 360 Open-Source Toolkit. Available at: https://www.ibm.com/blogs/research/2018/09/ai-fairness-360/. Accessed January 28, 2020.

[34] Google. ML Fairness. Available at: https://developers.google.com/machine-learning/fairness-overview. Accessed January 28, 2020.

[35] University of Chicago Center for Data Science and Public Policy. Aequitus – an open-source bias audit toolkit. Available at: http://www.datasciencepublicpolicy.org/projects/aequitas/. Accessed January 28, 2020.

[36] Fairness definitions in machine learning. Fairness Measures. Available at: http://www.fairness-measures.org/Pages/Definitions. Accessed February 27, 2020

[37] Rajkomar A, Hardt M, Howell MD et al. Ensuring Fairness in Machine Learning to Advance Health Equity. Ann Intern Med, 2018;169(12):866-87.

[38] Barocas S, Hardt M, Narayanan A. Fairness in Machine Learning: Limitations and Opportunities. Available at: https://fairmlbook.org/.

across all dimensions in the matrix. Work by Prof. Deborah Hellman at the University of Virginia School of Law has been exploring what these measures mean, and what actions should be based upon them.[39]

However, defining fairness on the basis of results from a confusion matrix may seem inadequate from the perspective of an ethicist or social scientist. While there are other theoretical constructs (e. g., John Rawls' theory of justice[40]), much of the approach to determinations of fairness in computer science is based on a utilitarian framework.

From a machine learning perspective, null models for fairness would be useful. For example, what should be considered "fair" in the setting of predictive algorithms? A working definition might be that *ceteris paribus,* two similar individuals get the same degree of prediction accuracy from the model.

One approach used by some companies for food tasting involves pairwise sampling from a given population in order to allow the researcher to decide who (within the sample population) is similar to whom according to attributes of interest. A recent preprint has analyzed fairness in machine learning based on 21 proposed metrics.[41]

Once Detected, How Should Bias Be Addressed?

Assuming a consensus on fairness metrics and definitions of bias—which conference participants agreed will be difficult to attain in the real world—what should the next steps be? If an algorithm's performance is meaningfully worse for some group, should the algorithm be adjusted, or should it be identified as "contraindicated" for some group? This is complicated by the difficulty in accurately quantifying measures of bias. A piecemeal imitation learning approach, where the algorithm would imitate exemplar doctors in different contexts and then pool different models, offers one option.

Aside from the models and choice of classifiers, data quality issues can introduce bias. For example: facial recognition technologies may have difficulty discriminating facial features in people of color, not only because of unrepresentative training data, but also because photographic technology has historically been optimized to capture lighter skin tones.[42] In another case, an ML application for identifying diabetic retinopathy worked well with high-quality training data, but experienced challenges in the field, where the quality of retinal images was not as good.[43]

A remaining unanswered question is whether it is necessary or desirable for algorithms to outperform human physicians in all tasks that can be potentially automated. In other words: for a given algorithm, what should the expectations of its performance be?

[39] Hellman D. Measuring algorithmic fairness. Virginia Law Review (forthcoming). Preprint available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3418528. Accessed February 21, 2020.

[40] Stanford Encyclopedia of Philosophy. John Rawls. Available at: https://plato.stanford.edu/entries/rawls/. Accessed January 28, 2020.

[41] Corbett-Davies S, Goel S. The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning. Preprint available from arXiv: https://arxiv.org/abs/1808.00023.

[42] Simonite T. The Best Algorithms Struggle to Recognize Black Faces Equally. Available at: https://www.wired.com/story/best-algorithms-struggle-recognize-black-faces-equally/. Accessed January 28, 2020.

[43] Abrams C. Google's Effort to Prevent Blindness Shows AI Challenges. Wall Street Journal. January 26, 2019. Available at: https://www.wsj.com/articles/googles-effort-to-prevent-blindness-hits-roadblock-11548504004.

All humans, no matter how acute, will have systematic biases that will inevitably manifest. That being the case, any approach to creating a model for fairness that incorporates ensembles of human judgments needs to incorporate outcomes data as well. In addition, there are also fortuitous mistakes that turn out to be beneficial (for instance, because they accidentally yield insight into a clinical situation), and these might not be recognized in the setting of a purely imitative approach to training a clinical or diagnostic algorithm.

The Importance of Metadata and Data Provenance

Importantly, where the data come from and the biases of those data are separate from the specific task defined. The context of the data must be preserved: how were the data collected; what was the situation at the time? Part of the obligation, especially in light of good algorithmic practices, is ensuring that as much descriptive/contextual metadata is available as possible, in order to provide assurance of the data's validity and reproducibility.

Working Session 3: Workforce Development & Technology

Expertise and Educational Background

In terms of identifying the kinds of skills/expertise that would be most useful for holistically evaluating ML models, an MD/PhD candidate who combines clinical and technical knowledge and experience would seem to be ideal, as a person with a purely machine learning background might lack needed domain expertise. Another possibility is a person that combines expertise in ML or health/medical informatics with a deep grounding in formal biomedical ethics, or a person with a background in computational social science. It is important to stress that simply hiring persons with only machine learning expertise as such is unlikely to be helpful. For meaningful impact in the field, a background in medical informatics and the broader context of clinical care and research are essential for success.

Developing Standards for Data and Performance Evaluation

It has not been easy to perform studies using only EHRs for acquiring research data. Flatiron's recent success with a hybrid approach has benefited from the fact that they have access to a proprietary oncology EHR. In inpatient settings, the Epic EHR predominates. Epic has been a challenging environment in which to attempt to apply machine learning. The FDA is in the early stages of a growing focus on real-world data/real-world evidence,[44] and could potentially play the role of setting standards and nudging the industry along.

The FDA has jurisdiction over applications that supply a diagnosis, but algorithms that are merely predictive of individual outcome are not presently overseen by the agency. Precision FDA and Open FDA represent the agency's initial forays into cloud computing—it may also be desirable for the agency to have some level of infrastructure capable of serving as a runtime environment for evaluating ML algorithms. With regard to collecting RWE, should FDA be maintaining the "Sentinel System" for this, or should such responsibility be the obligation of companies marketing the given algorithm?

As a typical part of the approval process for a clinical trial involving an algorithm or other software, the FDA would receive the software and the clinical data as part of re-running the data analysis. An

[44] US Food and Drug Administration. Real-world evidence. Available at: https://www.fda.gov/science-research/science-and-research-special-topics/real-world-evidence. Updated May 9, 2019.

approach to regulating algorithms that requires the agency to completely reproduce the workings of application might require a substantially greater magnitude of computing infrastructure. A possible alternative would be for the company seeking approval to provide and set up their own machines, but then have the agency to substitute its own test data set to evaluate the algorithm.

This would raise a further question: would (or should) regulatory approval entail adherence to a standard for data (such as the Fast Health Interoperability Resources [FHIR] open-source standard)? The FDA may not have the statutory authority to mandate the use of such a standard, although the agency is allowed to request additional information. The Office of the National Coordinator has embraced the use of FHIR[45]; there may be an opportunity to engage other agencies in encouraging a move toward widespread use of a standard.

Next Steps

At the conclusion of the discussions, conference attendees identified the following key points as priorities for ongoing work to build on this and other efforts:

- Developing a consensus over Good Algorithmic Practices and the infrastructure and data science culture that supports such practices;
- Developing a regulatory workforce that is facile with healthcare and machine learning, and considering whether independent third party organizations can supplement this workforce; and
- Because algorithms will inevitably behave differently in diverse, real-world environments, consider creating a model analogous to the FDA's Sentinel System, in which regulatory clearance or approval also requires that real-world data be collected in a central regulatory repository.

[45] Office of the National Coordinator. HealthIT.gov. Notice of proposed rulemaking to improve the interoperability of health information. Available at: https://www.healthit.gov/topic/laws-regulation-and-policy/notice-proposed-rulemaking-improve-interoperability-health. Updated June 5, 2019.

Appendix I. Selected Readings

The following items were chosen as part of a landscape survey meant to capture a representative cross-section of scholarly and journalistic writing about ethical issues at the intersections of artificial intelligence, machine learning, and healthcare, with a specific emphasis on the problem of algorithmic bias in the context of clinical care. This select reading list includes preprint articles, peer-reviewed research and viewpoint articles published in the biomedical and biomedical ethics literature, reports and recommendations from expert panels, and journalistic investigations and essays.

The AI Initiative. AI in medicine and health. Available at: http://thefuturesociety.org/the-ai-INITIATIVE/#governments-policymakers. Accessed January 29, 2020.

National Academies of Sciences, Engineering, and Medicine. 2018. *Data Matters: Ethics, Data, and International Research Collaboration in a Changing World: Proceedings of a Workshop.* Washington, DC: The National Academies Press. doi: http://doi.org/10.17226/25214.

Caplan R, Donovan J, Hanson L, Matthews J. Algorithmic accountability: A primer. *Data and Society.* April 4, 2018. Available at: https://datasociety.net/output/algorithmic-accountability-a-primer/. Accessed January 29, 2020.

Vallor S. Moral deskilling and upskilling in a new machine age: Reflections on the ambiguous future of character. *Philos Technol.* 2015;28:107-24.

Eckersley P. Impossibility and uncertainty theorems in AI value alignment, or, why your AGI should not have a utility function. March 5, 2019. Preprint available from arXiv.org at: https://arxiv.org/pdf/1901.00064.pdf. Accessed January 29, 2020.

Luxton D. Recommendations for the ethical use and design of artificial intelligent care providers. *Artif Intell Med.* 2014;62(1):1-10. doi: 10.1016/j.artmed.2014.06.004.

Challen R, Denny J, Pitt M, Gompels L, Edwards T, Tsaneva-Atanasova K. Artificial intelligence, bias and clinical safety. *BMJ Qual Saf.* 2019 Mar;28(3):231-237. doi: 10.1136/bmjqs-2018-008370.

Gershgorn D. If AI is going to be the world's doctor, it needs better textbooks. Quartz. September 6, 2018. Available at: https://qz.com/1367177/if-ai-is-going-to-be-the-worlds-doctor-it-needs-better-textbooks/. Accessed January 29, 2020.

Angwin J, Larson J, Mattu S, Kirchner L. Machine Bias. ProPublica. May 23, 2016. Available at: https://qz.com/1367177/if-ai-is-going-to-be-the-worlds-doctor-it-needs-better-textbooks/. Accessed January 29, 2020.

Chen I, Szolovits P, Ghassemi M. Can AI help reduce disparities in general medical and mental health care? *AMA J Ethics*. 2019;21(2):E167-179.

Lamanna C, Byrne L. Should artificial intelligence augment medical decision making? The case for an autonomy algorithm. *AMA J Ethics*. 2018;20(9):E902-910.

Ross C. What if AI in healthcare is the next asbestos? STAT News. June 19, 2019. Available at: https://www.statnews.com/2019/06/19/what-if-ai-in-health-care-is-next-asbestos/. Accessed January 30, 2020.

Price II WN. Medical AI and contextual bias. *Harvard Journal of Law and Technology*. 2019 (forthcoming). Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3347890. Accessed January 30, 2020.

Osoba O, Wessler W. An intelligence in our image: The risk of bias and errors in artificial intelligence. RAND Corporation. 2017. ISBN: 978-0-8330-9763-7.

Gretton C. The dangers of AI in health care: Risk homeostasis and automation bias. Towards Data Science. June 24, 2017. Available at: https://towardsdatascience.com/the-dangers-of-ai-in-health-care-risk-homeostasis-and-automation-bias-148477a9080f. Accessed January 30, 2020.

Hsu J. Artificial intelligence could improve health care for all — unless it doesn't. Undark. July 29, 2019. Available at: https://undark.org/2019/07/29/ai-collaboration-medicine-doctors/. Accessed January 30, 2020.

Hagerty A, Rubinov I. Global AI ethics - A review of the social impacts and ethical implications of artificial intelligence. Preprint available from arXiv. July 18, 2019. Available at: https://arxiv.org/abs/1907.07892. Accessed January 30, 2020.

Coravos A, Chen I, Gordhandas A, Dora Stern A. We should treat algorithms like prescription drugs. Quartz. February 14, 2019. Available at: https://qz.com/1540594/treating-algorithms-like-prescription-drugs-could-reduce-ai-bias/. Accessed January 30, 2020.