Equitable Algorithms: Examining Ways to Reduce AI Bias in Financial Services

Testimony by
Rayid Ghani,
Distinguished Career Professor
Machine Learning Department and the Heinz College of Information Systems and Public Policy
Carnegie Mellon University

Before the
Committee on Financial Services
Artificial Intelligence Task Force
U.S. House of Representatives

The Honorable Maxine Waters, Chairwoman
The Honorable Patrick McHenry, Ranking Member

Wednesday, February 12, 2020

Chairwoman Waters, Ranking Member McHenry, Members of the Committee, thank you for hosting this important hearing today, and for giving me the opportunity to submit this testimony.

My name is Rayid Ghani and I am a Distinguished Career Professor in the Machine Learning Department and the Heinz College of Information Systems and Public Policy at Carnegie Mellon University. I've worked in the private sector, in academia, and extensively with government agencies and non-profits in the US and globally on developing and using Machine Learning and AI systems to tackle social and public policy problems across health, criminal justice, education, public safety, human services, and workforce development in a fair and equitable manner.

Artificial Intelligence (or Machine Learning)[1] has a lot of potential in helping tackle critical problems we face in society today, ranging from improving the health of our children by reducing their risk of lead poisoning[2], to reducing recidivism rates for people in need of mental health services[3], to

[1] I will use the terms AI and Machine Learning interchangeably in this testimony
[2] Predictive Modeling for Public Health: Preventing Childhood Lead Poisoning. Potash et al. Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2015)
[3] Reducing Incarceration through Prioritized Interventions. Bauman et al.. ACM SIGCAS Conference on Computing and Sustainable Societies, 2018.

improving educational outcomes for students at risk of not graduating from school on time[4,5], to improving police-community relations by identifying officers at risk of adverse incidents[6], to improving health and safety conditions in workplaces[7] and in rental housing[8]. AI systems have the potential to help improve outcomes for everyone and result in a better and more equitable society. At the same time, any AI (or otherwise developed) system that is affecting people's lives has to be explicitly built to focus on increasing equity and not just optimizing for efficiency. It is important to recognize that AI can have a massive, positive social impact but we need to make sure that we put guidelines in place to maximize the chances of the positive impact while protecting people who have been traditionally marginalized in society and may be affected negatively by the new AI systems.

An AI system, designed to explicitly optimize for efficiency, has the potential to result in leaving "more difficult or costly to help" people behind, resulting in an increase in inequities. **It is critical for government agencies and policymakers to ensure that AI systems are developed in a responsible and collaborative manner,** including and incorporating input from all groups of stakeholders including: developers who build and deploy AI systems, decision-makers who implement the systems in their workflows, and the community being impacted by these systems. **Integrating input from these diverse voices is a critical element of ensuring that new AI systems result in equitable outcomes for everyone**.

Equitable outcomes and not "just" unbiased or equitable algorithms

Contrary to a lot of work in this area today, I believe that "simply" developing AI algorithms that better account for fairness and bias is generally not sufficient to achieve more equitable decisions or outcomes. Rather, the goal of these efforts should be to make entire systems and their outcomes equitable. Since algorithms are typically not (and should not be) making autonomous decisions in critical situations, the entire decision-making system includes the AI algorithms, the decisions that are being taken by humans using input from those algorithms, and the impact of those decisions. It is entirely possible to have a perfectly fair and equitable algorithm providing fair and equitable recommendations but the human decisions following them may be biased or the interventions allocated as a result of that human decision are not as effective for certain people as they are for others, resulting in inequity in outcomes.

At the same time, it is possible to design a system that contains an algorithm that is not fair but coupled with the appropriate bias mitigation and intervention plan, can result in increasing equity in

[4] A Machine Learning Framework to Identify Students at Risk of Adverse Academic Outcomes. Lakkaraju et al. Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining
[5] http://www.dssgfellowship.org/project/identifying-factors-driving-school-dropout-and-improving-the-impact-of-social-programs-in-el-salvador/
[6] Early Intervention Systems – Predicting Adverse Interactions Between Police and the Public. Helsby et al. Criminal Justice Policy Review, 2017.
[7] http://www.dssgfellowship.org/project/improving-workplace-safety-through-proactive-inspections/
[8] http://www.datasciencepublicpolicy.org/projects/public-safety/san-jose-housing/

outcomes. In some recent preliminary work we did with Los Angeles City Attorney's office, we found that by careful consideration and analysis, we can mitigate the disparities that a potentially biased algorithm may create and coupled with a tailored intervention strategy, the system has the potential to result in equitable criminal justice outcomes across racial groups[9].

AI systems optimize for what the developers tell them to optimize for

AI algorithms are neither inherently biased nor unbiased (in the societal sense). They typically work by taking historical data and attempting to build a "model" that replicates some outcome that is specified in that historical data while attempting to also generalize in to the future. The developers of such a system often have to specify how to manage the two tradeoffs – how much of the past to replicate and how much to generalize to the (unseen) future. When such a system is built, the developers of the system also specify what metric(s) to optimize for. If the system is asked to correctly predict as many of the past decisions that were provided for it to "learn" from as possible, that is exactly what it attempts to do, regardless of the race, gender, age, or income of the people who these decisions were about. That is one step where a lot of bias may come in to the decisions recommended by this system.

The AI developer can, in fact, tell the algorithm to balance replicating as many human decisions correctly as possible with ensuring fairness and equity across certain protected attributes of people that we care about. Sometimes, the developers fail to incorporate equity considerations in building their AI models, which is of course equivalent to choosing a metric that attempts to replicate as many human decisions as possible, possibly resulting in re-creating and reinforcing historically biased decision processes. In these cases, it is important to remember that the human processes that designed the AI system should own the blame rather than passing it off to an AI algorithm that is being guided and optimized incorrectly, for the wrong goals.

AI is forcing us to make societal (and public policy) values explicit

Because an AI system requires us to define exactly 1) what we want to optimize it for, 2) which mistakes are costlier (financially or socially) than others, and 3) by how much, it forces us to make these ethical and societal values explicit. It is important to know that these values are of course implied in any decision-making process, including all the human decision-making processes that exist today, but are not necessarily made explicit. These implicit values coded in humans making decisions when biased and unfair, result in inequitable outcomes. For an AI system to function, these values need to be provided as a critical input. For example, for a system that is recommending lending decisions, we may have to 1) specify the differential costs of flagging someone as unlikely to pay back a loan and being wrong about it versus predicting that someone will pay back a loan and being wrong about it, and 2) specify those costs explicitly in the case of people who may be from

[9] Predictive Fairness to Reduce Misdemeanor Recidivism Through Social Service Interventions. Rodolfa et al. Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (ACM FAT*) 2020.

different gender, race, income, or education level groups. While that may have happened implicitly in the past and with high levels of variation across different human decision makers (loan officers in this case), with AI-assisted decision-making processes, we are forced to define them explicitly.

One key question we have to answer here is who and how should we come up with these sets of values for a given problem setting. Unfortunately, today, these decisions are too often left essentially by default to the AI system developer or an arbitrary set of individuals who define those values in an AI algorithm (explicitly or implicitly). The recommendations at the end of this testimony go into more detail on what I recommend should be done but it certainly should not be left to the AI system developer making those choices alone; the team and process should include all stakeholders including policymakers and the community being impacted by this system.

All Types of Biases are Not Equal

An AI system (or human) can be unfair in a variety of ways and there is no universally-accepted definition of what it means for an AI system to be fair. Take the example of a system being used to make loan determinations. Different people might consider it "fair" if:

- It makes mistakes about denying loans to an equal number of white and black individuals
- The chances that a given black or white person will be wrongly denied a loan is equal, regardless of race
- Among the population who were denied loans, the probability of having been wrongly denied a loan is independent of race
- For people who should be given loans, the chances that a given black or white person will be denied a loan is equal
- The lending decisions serve to reduce or eliminate disparities in home ownership rates across black and white individuals

These different notions of fairness have formal names and definitions in research literature[10] and a great deal of research has been done describing these fairness notions in different fields. In different contexts, reasonable arguments can be made for each of these potential definitions, but unfortunately, not all of them can hold at the same time[11,12]. In general, understanding which type of bias should be prioritized and weighted more than others requires consideration of the societal goals and a detailed discussion between decision makers, AI developers, and most importantly those who will be affected by the application of the model. Each perspective may have a different concept of fairness and a different understanding of harm involved in making different types of errors, both at individual and societal levels. Practically speaking, coming to an agreement on how fairness should

[10] Sahil Verma and Julia Rubin. 2018. Fairness Definitions Explained. In Fair- Ware'18: IEEE/ACM International Workshop on Software Fairness, May 29, 2018, Gothenburg, Sweden. ACM, New York, NY, USA, 7 pages. https://doi. org/10.1145/3194770.3194776

[11] Alexandra Chouldechova. 2017. Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. Big Data 5, 2 (6 2017), 153–163. https://doi.org/10.1089/big.2016.0047

[12] Moritz Hardt, Eric Price, ecprice, and Nati Srebro. 2016. Equality of Opportunity in Supervised Learning. In Advances in Neural Information Processing Systems 29. Neural Information Processing Systems Foundation, Barcelona, Spain, 3315– 3323.

be measured in a purely abstract manner is likely to be difficult. Often it can be instructive instead to explore different options and metrics based on preliminary results, providing tangible context for potential trade-offs between overall performance and different definitions of equity and helping guide stakeholders through the process of deciding what to optimize[13]. While we have a more comprehensive set of guidelines we term the Fairness Tree[14], some of the guidelines we have developed and use in our work include:

- If the intervention is punitive in nature (e.g., determining whom to deny loan), individuals may be harmed by intervening on them in error so we may care more about metrics that focus on false positives.
- If the intervention is assistive in nature (e.g., determining who should receive loan forgiveness), individuals may be harmed by failing to intervene on them when they have need, so we may care more about metrics that focus on false negatives.
- If the available resources are significantly constrained such that we can only intervene on a small fraction of the population at need, a different set of metrics may be of more use (see Fairness Tree[10] for more details).

Bias in AI systems can come from a lot of sources and it's important to separate them out

Bias may be introduced into an AI system at any step along the way and it is important to carefully think through each potential source and how it may affect the results. In many cases, some sources may be difficult to measure precisely (or even at all), but this doesn't mean these potential biases can be readily ignored when developing interventions or performing analyses. These sources include

1. **Biased data sources**: due to either data being used to build an AI system not being representative of the population it will be used to make decisions for, or having incorrect/biased outcomes for certain people (based on historical biases in society and/or human decision-making such as over-policing black communities), or the unknowability of certain outcomes from past decisions(for instance, you can't know whether or not an individual who was denied a loan would actually have repaid it had it been granted).
2. **Bias due to decisions made by AI developers when designing the system**: I will not go into detail here but would refer to other literature[15] that describes different analytical decisions that are made when developing an AI system that can lead to biases.
3. **Application Bias**: This is often not due to the AI algorithm being biased but because of the way the results of an AI algorithm are applied. One way this might happen is through heterogeneity in the effectiveness of an intervention across groups. For instance, imagine an AI system built to identify individuals most at risk for developing diabetes in the near future for a particular preventive

[13] Predictive Fairness to Reduce Misdemeanor Recidivism Through Social Service Interventions. Rodolfa et al. Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (ACM FAT*) 2020.

[14] http://www.datasciencepublicpolicy.org/projects/aequitas/

[15] https://textbook.coleridgeinitiative.org/chap-bias.html

treatment. If the treatment is much more effective for individuals with a certain genetic background relative to others, the overall outcome of the effort might be to exacerbate disparities in diabetes rates even if the AI algorithm itself is unbiased.

What does it take to create AI systems that lead to equitable outcomes for society?

The following steps need to be taken to attempt to create AI systems that are likely to lead to equitable outcomes for society:

1. **Defining** the (equitable) outcomes we want to achieve in society (which includes the societal values and a collaborative, multi-stakeholder process).
2. **Translating/Mapping** those desired societal outcomes into analytical requirements that the AI system should optimize for.
3. **Building** an AI system that fulfills those analytical requirements and releasing documentation on how it was built to achieve those goals. This step includes
    A. **Detecting** biases in intermediate/iterative versions of the system
    B. **Understanding** the root causes of the biases
    C. **Improving** the system by reducing the biases (if possible) or selecting tradeoffs across competing objectives
    D. **Mitigating** the impact (and coming up with an overall mitigation plan) of the residual biases of the system
4. **Validating** through a trial (and providing evidence) that the AI system did, in fact, fulfil those requirements and achieve the initial outcomes defined in step 1 before deploying the system.
5. **Continuous Monitoring & Evaluation** of the entire system (AI algorithm followed by human decisions) during its lifetime to ensure that it continues to achieve equitable outcomes from step 1.

It is important to note that the steps above are **not purely technical**, but rather involve understanding existing **social and decision-making processes and systems** as well as **collaboratively coming up with solutions for each step**. These steps may require new data to be collected and existing processes to be modified in order to ensure equitable outcomes. For example, if data about race or gender was not being collected in the past, and the goal is to monitor and achieve equity across different groups of gender and race, it will require new data collection processes. Likewise, goals surrounding fairness and equity must be actively integrated into the modeling, evaluation, and decision-making processes. A considerable body of work[16] has

---

[16] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In Proceedings of the 3rd Innovations in Theoretical Computer Science Conference. ACM Press, New York, New York, USA, 214–226. https://doi.org/10.1145/2090236.2090255

R. G. Fryer, G. C. Loury, and T. Yuret. 2007. An Economic Analysis of Color-Blind Affirmative Action. Journal of Law, Economics, and Organization 24, 2 (11 2007), 319–355. https://doi.org/10.1093/jleo/ewm053

Toon Calders and Indre Žliobaitė. 2013. Why Unbiased Computational Processes Can Lead to Discriminative Decision Procedures. In Discrimination and Privacy in the Information Society. Studies in Applied Philosophy, Epistemology and Rational Ethics, Volume 3, Springer Press, Berlin, Germany, 43–57.

demonstrated that notions of "fairness through unawareness" (e.g., simply excluding or ignoring these protected attributes in AI systems) is insufficient for achieving equitable results, both because these attributes are often highly correlated with other predictors and due to historical disparities in outcomes themselves.

Moving Forward to a More Equitable Society: Our Recommendations

It is critical and urgent for policymakers to act and provide guidelines and/or regulations for both the public and private sector organizations using AI-assisted decision-making processes in order to ensure that these systems are built in a transparent and accountable manner and result in fair and equitable outcomes for society. As initial steps, we recommend:

1. Expanding the existing regulatory environment to account for AI-assisted decision-making

Instead of creating a Federal AI regulatory agency across policy areas, we should expand the already existing regulatory frameworks in different policy areas to account for AI-assisted decision-making. A lot of these regulatory bodies already exist, including SEC, FINRA, CFPB, FDA, FEC, FTC, and FCC that are each responsible for ensuring compliance with existing regulations. These bodies typically regulate the inputs that go into a decision-making process (for example, what attributes cannot be used such as race or place of residence) and often the processes themselves, but do not always focus on the outcomes produced by these processes. We recommend expanding these regulatory bodies to:

1. Update the regulations to make them outcome-focused.
2. Update the regulations to ensure they apply to AI-assisted decision making.
3. Define the set of artifacts an organization (government or industry) should publicly release before deploying (and ideally during the development phases of) an AI system. This includes information on how the system was built, what it was designed to optimize for, what tests were run to check if it did, what types of people is it effective for, who does it fail for, how long was it in trials for, and how did the effectiveness change over time. Ideally this should be put in place for any process involving decision making of any kind, whether human decisions or AI-assisted decisions but becomes critical in cases where the scale of deployed AI systems increases the risk. This set will need to vary based on the impact this system can have on people's lives.
4. Define a set of risks that could lead to inequities that need be considered when building an AI system and a mitigation plan for each of these risks.
5. Set up an extended data collection process and infrastructure to collect additional data attributes (such as race, gender, or income) that may not already be collected but are necessary to measure equity outcomes (to deal with the "fairness through unawareness" issue described earlier.

6. Set up evaluation standards to compare the performance of these systems to the human decision-making processes currently being used.
7. Define standards around the explainability of the AI systems in order to provide recourse to individuals who may be adversely impacted by the decisions made using the system.

These expanded bodies should be responsible for defining standards as well as for continuous monitoring, audits, and compliance with the standards and regulations.

2. Creating Trainings, Processes, and Tools to Support Regulatory Agencies in their Expanded Roles

As these agencies expand their role, they will need to be supported by increasing their internal capacity to fulfil this role and ensure that regulations are being effectively complied with. We recommend creating trainings, processes, and tools to help them

1. Understand where existing regulations may and may not be well-adapted to applications involving AI-assisted decision-making.
2. Understand and define what equitable outcomes standards to set.
3. Understand how to evaluate whether the requirements created for an AI system were in fact aligned with the identified societal equitable outcomes.
4. Understand how to evaluate whether the AI system did in fact do what it was designed to do.
5. Develop a continuous monitoring and audit process and tools (such as Aequitas[17]) to support the audit process.
6. Create standards for when a system should "expire" and a corresponding renewal process.

3. Procuring AI systems should include Key Requirements in the Request for Proposals (RFP) Process

Government agencies and corporations putting out RFPs for AI systems that are making critical decisions and affecting people should require proposers/bidders to include:

- An explicit initial project phase to gather requirements for what it would mean to have equitable outcomes and what they should be. This process should include a diverse team and work with stakeholders including: developers who build and deploy AI systems, decision-makers who implement the systems in their workflows, and the community being impacted by these systems.
- A detailed plan and methodology for Steps 1-5 in the previous section of this testimony titled "What does it take to create AI systems that lead to equitable outcomes for society?"
- A continuous improvement plan to ensure that the system continues to not only be evaluated but also improved upon to achieve equitable outcomes.

[17] http://www.datasciencepublicpolicy.org/projects/aequitas/