ILO INSTITUTE

December 2020

Research Note: Proxy Discrimination

"Proxy Discrimination" is emerging as a critical barrier to deploying artificial intelligence to help assess risk in financial services, criminal sentencing, allocation of healthcare resources, and a broad range of additional areas.

The root of the problem is two-fold.

First, actual risk variance often reflects decades and even centuries of intentional bias that has left groups identifiable by race, income, area of origin and other identity characteristics that our laws and our culture make unacceptable. AI systems will often correctly identify actual differences in risk that reflect past discrimination, and acting on that data will only reinforce past bad actions and social structures.

Second, the easiest to access pools of data with which to train AI systems are often the most biased. Thus facial recognition systems generally perform better distinguishing among the facial characteristics of racial and social groups most like the engineers who build them.

Proxy discrimination also brings with it incentives for behavior change that increase rather than decrease risk, like not connecting to friends and family via social media or not exploring personal genomic data in order to avoid vulnerability to algorithmic bias.

This chart produced by data scientist Emma Pierson (now at Stanford) and her colleagues illustrates the ethical challenges in the data gathering and analysis process, and specific

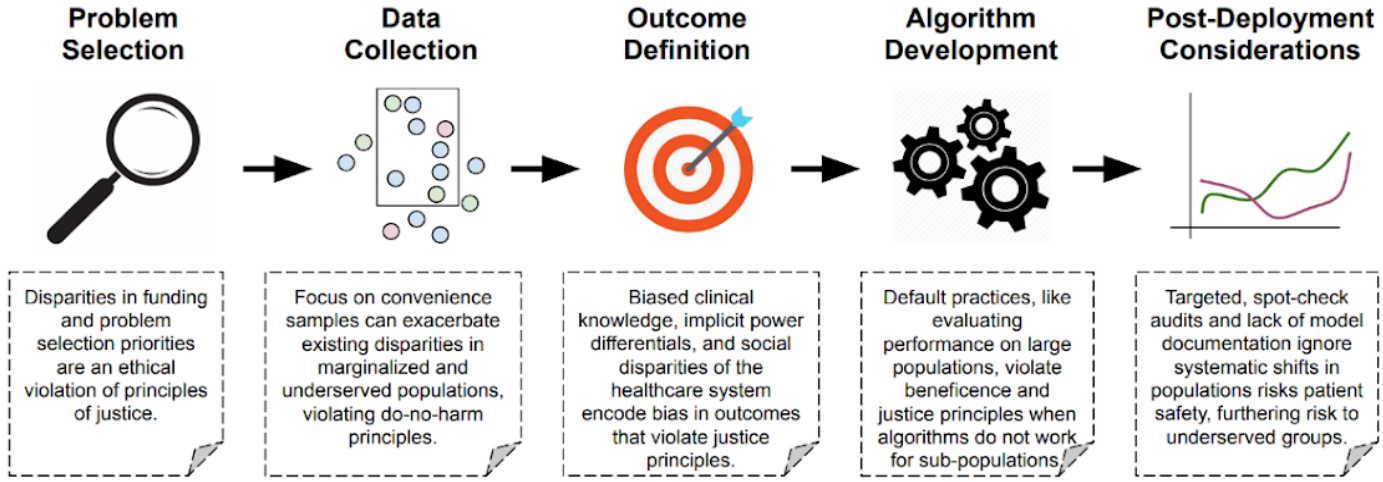


Chart: October 2020 [2009.10576.pdf (arxiv.org)]

There are remedies.

Among the most common now being explored and deployed are

1) Explicitly balancing outcomes to protect special classes most likely to be harmed by "unbalanced" bias (creating "ethical algorithms" with thumb-on-the-scale prisms built-in, or practicing DADM, "Discrimination-Aware Data Mining);

2) Aggressively seeking significantly more data, deeper and wider, to get past the surface-layer correlations that often contain the greatest dimensions of proxy

discrimination (Google's Federated Learning ML model may prove deeply helpful here); and

3) The active promotion of pro-social behaviors among groups and individuals likely harmed by proxy discrimination, to "cure" their risk profiles.

We see great limitations in option 1; a valuable but difficult hill to climb in option 2; and great opportunities not only for social good but also for competitive advantage in option 3.

Option One: The Folly of Counter-Bias

Creating "ethical algorithms" that don't try to be directly fair and accurate, but purposefully bias results toward a social aim is a tough challenge, and one that we believe offers false hope. The kinds of counter-bias we deploy in doing this rely on very slow-moving, often quite subjective sources of what the engineered social good actually is. Questions about whose truth standard or whose fairness standard to deploy, and non-dynamic hardening of approved paths and weights, inevitably lead these kinds of interventions to unintended negative consequences. The better path is not seeking counterbalancing unfairness, but actually achieving fairness.

Option Two Thinking: More Data Decreases the Proxy Distance, and Gets to Real Relevance at the Level of the Individual

Proxy discrimination may well prove to be an artifact of an interim stage in technology development. As we have access to much more data about individuals we might insure, hire or incarcerate than we've had before, we can easily fool ourselves into believing that we have enough data to make better decisions when in fact we have very little – it only

seems like a lot, and perhaps it might even seem adequate, because it is so much more than we've had before.

Data available for risk rating is a tiny fraction even today of what it will be in a few years, and more data, better analyzed, will reveal truer correlations between the complex web of identity and individual actions and real risk than we can imagine today. As a helpful article in the Iowa Law Review (discussed further, below) outlines the issue:

> *"AIs may, in fact, decrease the incidence of statistical discrimination—by proxy or otherwise—by reducing the costs of acquiring and processing data about directly relevant characteristics.*
>
> *"For instance, to the extent that past incarceration rates are indeed directly predictive of job performance for a particular employer, the AI might either be able to directly access this information or else to construct more reliable proxies than race for this information.*
>
> *"It follows that increasing AI's access to relevant data could decrease the program's need to rely on proxies for suspect characteristics by allowing it to more directly measure the factors that most directly relate to risk.*
>
> *"Adding data to AI models would also minimize situations where actors implicitly accept the possibility of decreased accuracy in their models to save costs on collecting or verifying information."*

In other words, it's not the "discrimination" side of the "proxy discrimination" equation that will shift over time, but the "proxy" side, as we are able to address individuals in their full complexity based on rich sources of data that are still largely untapped.

Google's Federated Learning ML model offers a glimpse of this possibility, built as it is to analyze pools of data – like all the data on an android smar- phone, or a genome – to find

correlations and data structure, without identifying data-source individuals or copying their data, but reviewing it all to find correlations at deep levels and allowing correlative and causal models to become much richer and more accurate.

Finding the Cure

The head of new-product development at one of the largest integrated financial firms in the world shared with us his approach.

> *"We can look at equitable use of data in risk-management as a multi-step process. We have the opportunity to foster pro-social behavior – to create incentives to balance the indicator of bad risk with acquisition of good-risk characteristics.*
>
> *"We can focus on what you do with a ballot lacking a signature, as we've seen and heard so much talk of during this election cycle, to 'cure' the ballot, that is to say, to add what's missing, to confirm the initially risk-filled provenance of the ballot and remove that risk. How we do that is part of what my office is exploring in some details right now."*

Option Three Thinking: Prove risk rather than predict it

Something sneaky in the micro-finance world offers a very good example of financial risk-modeling that overcomes proxy discrimination in two ways. Grameen Bank of Bangladesh and Kashf in Pakistan have notably claimed repayment rates by unbanked, lightly documented women in rural areas that Western banks have been unable to duplicate.

Grameen and Kashf succeed where Western banks tend to fail because of a bit of gaming the system. While local micro-credit organizations still make claims – and supply data – about repayment rates of over 90%, that number is an artifact of the common practice of the screening loan.

That is, the first loan is not recorded as a loan, but as a membership initiation, and only those – generally about 50% - who successfully pay that first screening loan are put on the books as first-time lenders. That more elite group repays at a much higher rate than the larger initial group.

A powerful lesson here is that the initial loan is an effective tool for proving real risk – not predicting real risk. This may be a powerful way to avoid proxy discrimination by creating challenges that can in effect cure the poor rating that a discriminatory application of data analysis initially proposes, or at least reveals an individual's actual behavior, rather than predicted behavior.

The model for this kind of intervention becomes far more effective – and far less expensive to deploy – if the cure is automated in some way. For example, a two-step screening process that begins with the filling out of a simple form, and then the user identifying a time at least 24-hours hence for a review interview, can both screen for, and create an incentive for, timely behavior even if the review interview turns out to be merely a proforma note of approval.

The kept promise of prompt return is a signal of lower risk, and the performance of the task is itself generative of more similar behavior.

The masters of automated interaction to change behavior are, today, the large social media platforms. While it is true that their data-driven behavioral scientists largely craft interactions to increase time-on-platform, the major players have invested in explorations of automated processes to create more prosocial behavior. One or more of these players would make valuable partners to experiment with pro-social behavior change at scale to help defeat proxy discrimination.

Two Types of Proxy Discrimination, as revealed from traffic-stop data

Based on an initial set of reviews of traffic-stop data available at the Stanford Open Policing Project, we believe that proxy discrimination can be helpfully seen as falling into two types in physical-world policing.

Type 1 proxy discrimination is the result of training and attitudes that can be shifted or counterbalanced by training differently and putting limits on police behavior via monitoring, rewards and punishments.

Type 2 is the impact of long-term structural biases that lead, for example, poor people and people of color to be more likely to be operating an automobile without a license.

The difference, roughly, is between stopping a car because it "feels out of place in a neighborhood like this" (Type 1 proxy discrimination, where an old, poorly maintained car is a proxy for race and income); and stopping a car because of an expired registration tag (Type 2 bias).

We believe it will be helpful for insurance companies to view their own proxy discrimination challenges through a similar lens. Identify type 1 instances of proxy discrimination and you will generally find a path to curing that discrimination with more data. Find the type 2 instances, and you will have a tougher challenge, shifting people to more pro-social behavior, likely using challenge screening and other similar techniques.

Resource: The Stanford Open Policing Project This online resource is rich in state-by-state traffic-stop data as well as several rigorous studies on the nature of bias these statistics reveal.

Resource: Reducing bias in AI-based financial services (brookings.edu) This Brookings Institution report on an approach to mitigating proxy discrimination in financial services

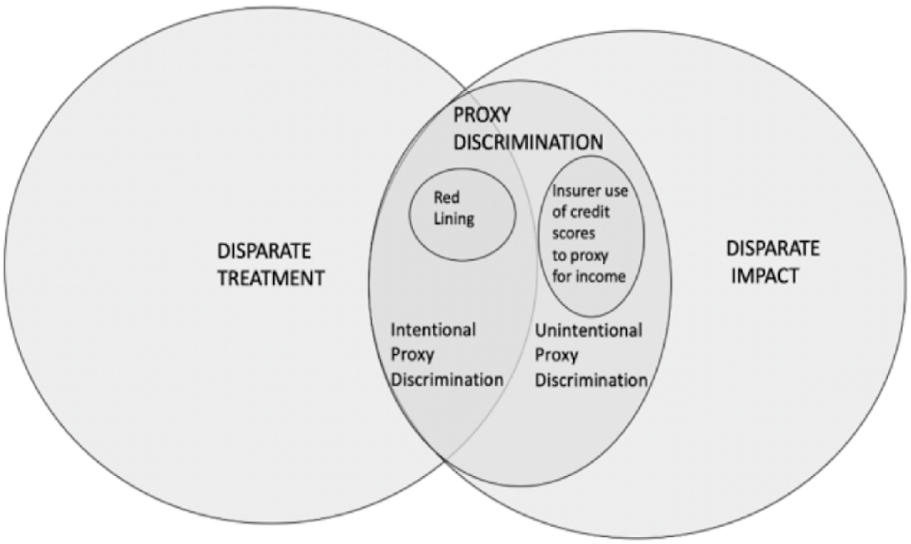
ultimately advocates “less accuracy for less bias,” which we believe is not a necessary tradeoff. Still, it offers a helpful discussion.

Undoing Incentives for Alienating Behavior

Following the principle that a good measure redeployed as a goal then loses its value as a measure, a quite useful paper published in the Iowa Law Review notes that the challenges of proxy discrimination have

> *“led prominent newspapers like the Wall Street Journal to recommend that individuals post on social media pictures of themselves exercising and eating healthy, while avoiding posts of themselves smoking or engaging in extreme sports. As proxy discrimination by AI becomes more common, it is easy to imagine similar newspaper stories warning individuals not to join Facebook groups associated with suspect characteristics like genetic conditions or domestic violence, because doing so might result in future adverse consequences for insurance, credit, or employment.”*

The article includes this useful chart:



Resource: Proxy [Discrimination in the Age of Artificial Intelligence and Big Data - Iowa Law Review - The University of Iowa College of Law (uiowa.edu)] This law-review article offers a sound review of key liability issues.