The limits of the quantitative approach to discrimination

2022 James Baldwin lecture, Princeton University (lightly edited transcript)[1]

Arvind Narayanan

Oct 11, 2022

Contents

Good evening. It's truly an honor to be here. I have a confession to make. When I was invited to give this lecture, my first reaction was panic. After all, this lecture series is named after James Baldwin, who was known for wielding his words as a weapon against oppression. Many of the previous speakers in this series have been noted for their oratory. But me? I live in the world of numbers. I do quantitative research. It's abstract and kind of removed from the actual experience of discrimination.

[1] For more information about the lecture and other formats, see https://www.cs.princeton.edu/~arvindn/talks/baldwin-discrimination/

So I was pacing back and forth in a puddle of insecurity. But then I realized that the very thing that was giving me anxiety, the difference between these two ways of knowing, words and numbers, these two ways of studying the human condition, would itself make for a good topic for this lecture. It's a topic that's been endlessly debated and written about, but I will make the case that it is still important to revisit it. I'm guessing that among you there are both quantitative and qualitative scholars, and I hope that I will have something interesting to say to each of you. Most importantly, for members of the public, I hope this will help people see past the myth that numbers don't lie.

Let's set the stage. In 2016, ProPublica released a ground-breaking investigation called Machine Bias.[2] You've probably heard of it. They examined a criminal risk prediction tool that's used across the country. These are tools that claim to predict the likelihood that a defendant will reoffend if released, and they are used to inform bail and parole decisions.

This was far from the first time that someone had pointed out that these tools disproportionately harm Black and minority populations. What was different about the ProPublica investigation was data. They made a FOIA request to Broward County, Florida, and managed to obtain data on the tool's predictions — data that had previously been kept secret by the company.

Their investigation included many statistical ways of measuring the racial bias of this tool. One in particular stood out. This tool makes predictions, so we can ask: how often are the predictions wrong? And how does that differ between Black defendants and white defendants? What they found was that for Black defendants, the tool was twice as likely to falsely flag someone as high risk compared to white defendants.

Prediction Fails Differently for Black Defendants

	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

[2] Angwin et al. *Machine bias.* ProPublica, 2016.

The first row here is the critical one. It's particularly egregious when someone who doesn't end up reoffending is nonetheless classified as high risk. You can see that that happens about 23% of the time for white defendants and about 45% of the time for Black defendants.

I believe that putting a number to the experience of discrimination is the reason that the investigation has had such an impact. And it's really had an impact. It was a finalist for the Pulitzer Prize. It kick-started a field of academic research on algorithmic fairness and has been cited thousands of times. It has been taken very seriously in policy circles, which in fact culminated in last week's AI bill of rights released by the Biden-Harris administration, which gives guidance to federal agencies in how they adopt AI for decision making.[3]

My disillusionment

When this investigation came out, I thought, "This is great! Earlier we had only anecdotes, and now we have data." But then I actually started working in this area, partly because of this piece. The edifice I'd built up in my head came crashing down. I realized what what happened with the ProPublica investigation is extremely unusual, and it's far more common for researchers using quantitative methods to be oblivious to discrimination. I realized that **baked into the practice of quantitative methods is a worldview that sees the status quo as unproblematic**.

It's that realization that I want to share with you today. I'll tell you about 7 of the most serious limitations of the way quantitative methods are used to study discrimination, not just algorithmic discrimination. I could give you more than 7, but I don't want us to be here all night. I think they've actually become tools for justifying racism and excusing inaction. And that's true even when they're used by scholars who are motivated by racial justice.

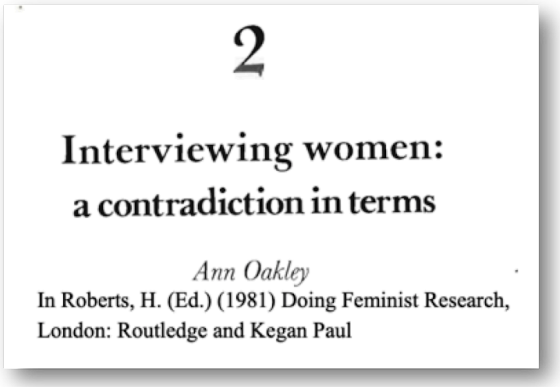
Now, in spite of my disillusionment, I'm still a quantitative scholar. I'm not about to conclude that quantitative methods should never be used. But I do think that if we actually want to help and not mislead, they should be used very differently, and as I go along I'll give suggestions for doing so.

---

[3] White House Office of Science and Technology Policy. *Blueprint for an AI Bill of Rights: making automated systems work for the American people*. 2022.

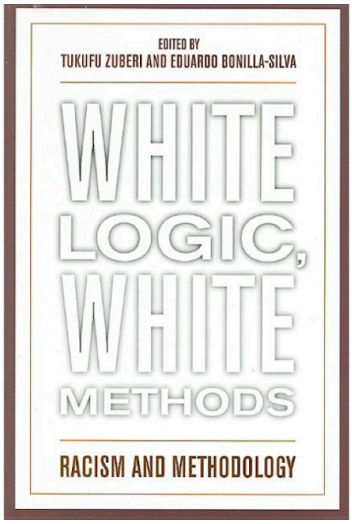
Let me once again acknowledge that the relative merits of quantitative and qualitative methods have been debated endlessly. In the social sciences the so-called paradigm wars have given way to mixed-methods research in which the two sets of methods are used somewhat harmoniously.

Skepticism of quantitative methods has been a major theme of both feminist scholarship and critical race theory. You can see a couple of examples below.[4] This is Oakley's early work, by the way, on the limitations of quantitative methods in feminist scholarship. Later in her career she in fact adopted quantitative methods and became a prominent defender of their use. I'll say more on that later.


2

Interviewing women:
a contradiction in terms

Ann Oakley
In Roberts, H. (Ed.) (1981) Doing Feminist Research,
London: Routledge and Kegan Paul



Scholars or students of critical race theory might find many of my points in today's lecture to be obvious. And that makes sense: I'm not claiming to offer anything radically novel. At the same time, I think it's unfortunately true that a lot of the critiques aimed at quantitative methods have been less than compelling to quantitative researchers.

Let me make my point through a bit of a caricature. Suppose you point out that numbers can never capture the nuances of the human experience. That's one of the most obvious limitations of quantitative methods. A quantitative scholar would respond: Of course they can't! No one

---

[4] Oakley. *Interviewing women: A contradiction in terms*. In Doing feminist research. Routledge, 2013. 30-61; Tukufu & Eduardo Bonilla-Silva, eds. *White logic, white methods: Racism and methodology*. Rowman & Littlefield, 2008.

claimed they could. One of the most common aphorisms in statistics says, **all models are wrong but some models are useful**. In other words, quantitative researchers are keenly aware that our models are an oversimplification. But they generally produce useful results, and indeed quantitative methods have been spectacularly successful in engineering and other fields. So if one is claiming that models are not just wrong, but so wrong that they're not even useful, the onus is on them to show which models those are and how the results of those models are misleading. But to do that one has to get more technically specific. An argument that rejects all quantitative research with the same brush is no argument at all.

I hope to offer some of that technical specificity today. I hope there is some value in contesting quantitative work on its own terms. If you're familiar with the movie expression, "the calls are coming from inside the house", that's the effect I hope to have. An insider's critique is harder to ignore.

Before diving in, I should also mention that I developed the point of view that led to this talk in the process of co-authoring my textbook on fairness and machine learning. I'm grateful to my coauthors Solon Barocas and Moritz Hardt.

What counts as evidence of discrimination is a subjective choice

Let's start with a depressing but unsurprising fact: about 1% of CEOs in Fortune 500 companies are Black. As an aside, I can't help mentioning that Fortune magazine called this number a record high.


LEADERSHIP · FORTUNE 500

The number of Black Fortune 500 CEOs returns to record high—meet the 6 chief executives

Here's why I'm bringing up this statistic. As a quantitative scholar, you're not allowed to conclude from this that there is discrimination in Fortune 500 companies. You're supposed to be open to all possibilities. Like maybe Black people just aren't that interested in becoming CEOs. I want to stress that this way of thinking is completely natural and normalized for quantitative scholars. Most of them would say, "Of course you can't claim discrimination without having proven it!" It's what I would have said a few years ago.

But what I've come to realize is that this is a choice that quantitative communities have made rather than something that's an inevitable consequence of the scientific method. Scientific communities face subjective choices all the time in deciding what counts as evidence, how much evidence is enough, and who has the burden of producing evidence. Here's a simple example. Soon after covid vaccines became available, the first variants started gaining prominence. Journalists asked scientists if the vaccines would be effective against variants. At first, some scientists, trying to be cautious, said we have no evidence of effectiveness against variants.

"no evidence of vaccine effectiveness against variants"

vs

"no evidence that vaccines are less effective against variants"

But soon, other scientists pushed back. They reframed the lack of evidence in a subtle way. They said we have no evidence that vaccines are any *less* effective against the new strains. The two statements sound almost the same, but their implications are worlds apart. If we have no evidence of vaccine effectiveness against the new strains, then there's no point in taking vaccines since the variants were starting to predominate. In fact, this worry about anti-vaxxers was one of the reasons why scientists started adopting the new framing. If there's no evidence that vaccine effectiveness is decreased, then we needn't really worry about the variants or do anything differently unless new evidence comes up.

Scientists have to reach tentative conclusions in the face of incomplete evidence *all the time*. But science itself doesn't tell us how to do this. These decisions often come down to individual scientists' judgments, their biases, to convenience, and to politics. Scientists often get it wrong.

For decades, psychologists have done research on college students in "Western, Educated, Industrialized, Rich and Democratic" (WEIRD) societies and assumed that the conclusions hold for all people.[5] Doctors used to exclude women from clinical trials because of sexist beliefs, and they assumed that interventions that worked well on men would work well for everyone. But that didn't always end up being the case, because of sex differences in our bodies. Here's another one: medical researchers are surprisingly comfortable drawing conclusions from studies in mice. Tentative conclusions, but still.

The null hypothesis allocates the burden of proof

Now, when it comes to discrimination, maybe there aren't studies specifically about CEOs and Fortune 500 companies, but we have so much other evidence of racial discrimination and structural racism. Qualitative evidence, and quantitative evidence in other contexts. Shouldn't we incorporate that evidence? A lot of this comes down to the so-called null hypothesis. The null hypothesis in science is the default assumption; it's how we presume the world works without evidence otherwise. Almost universally, quantitative researchers would say that the null hypothesis is that there is no discrimination.

But that's not a logical inevitability. That's a choice. We could instead say that when we observe disparities, like in the demographics of CEOs, then the null hypothesis is that those differences are due to discrimination. It is those who claim that there's no discrimination who have the burden of proof.

Now, why is it that quantitative communities have settled on the absence of discrimination as the null hypothesis? I don't know, but here's one possible reason. If you're a privileged person who does not perceive discrimination on a daily basis, then it is natural to view the absence of discrimination as the default. But in the world we actually live in, it makes little sense.

When researchers pick the null hypothesis on autopilot, mimicking what's been done before, they are often oblivious to the fact that their choice has enormous normative significance. And that's a shame.

---

[5] Henrich, Heine, Norenzayan. *The weirdest people in the world?* Brain and behavioral sciences 2010.

In any case, the effect of the system we have right now is that civil rights advocates or others alleging discrimination have the burden of proving again and again and again that it exists. By the way, it's not just researchers who have this worldview: it's also policy makers and decision makers of all kinds. For example, why do we have diversity efforts at institutions rather than antiracism or anti-discrimination efforts? Because it is unacceptable to suggest that the disparities we observe are the result of discrimination. What is politically palatable, instead, are diversity efforts that are agnostic about the reasons why the world is the way it is. Rather than try to right a wrong, they view diversity as something that is good for the organization and so try to promote it. Because of this, such efforts are much less effective than they could be.

Here's something fascinating. **United States law uses something called a burden shifting framework**. It is an interesting approach to making decisions in the face of incomplete evidence. These frameworks have been established by the courts including the supreme court. If a plaintiff alleges a discriminatory practice by an employer, the first step is for them to show what is called prima facie evidence of discrimination. That might be something along the lines of an observed disparity. They don't actually have to prove that this was because of discrimination. At this point, burden shifts to the employer to produce a reason justified by business necessity that explains the observed disparity. If the employer can't do this, the plaintiff prevails. If the employer is successful, then burden shifts back to the plaintiff to show that the employer could have achieved its business goals in a less discriminatory way.

Now, this system has a lot of limitations, and we discuss in our book how anti-discrimination law falls far short of its ideals. Nonetheless, I think there's a real epistemic innovation here and it's worthwhile for quantitative scholars to pay attention to it.

The field of psychology has recently adopted a different kind of adversary model. I think it's been forced to do that because of the reproducibility crisis. It's called an adversarial collaboration, and it explicitly acknowledges that scientists have their biases, their pet theories, their favored and hoped-for conclusions, and that these factors will consciously or unconsciously bias the methods that they pick in a way that makes certain conclusions more likely. I think this development in psychology is interesting and admirable.

Compounding inequality is far below the radar of quantitative methods

Let's go back to the CEO example. You might wonder, what's so hard about this burden of proving discrimination? Shouldn't it be possible to look at the data and use quantitative methods to test whether the fact that there are so few Black CEOs is because of discrimination?

I'm going to argue that even with a disparity as stark as 1% of CEOs being Black, and even if that disparity is entirely because of discrimination (which I think it probably is), it is still going to be practically invisible to quantitative methods. That might seem a shocking claim, so let me justify it. I'm going to give you a simple mathematical model that shows why this might be the case. Like I said, this talk is about using quantitative methods to knock quantitative methods.

I looked at the data and it turns out that the workforce of these companies is about 7% Black. I'm guessing that they use the usual proxies for race, such as hiring only from "elite" universities, to end up with a workforce composition that already doesn't reflect the country's diversity. But let's set that aside for now. How do we go from 7% of the overall pool being Black to just 1% at the top? That's the key question.

Here's the thing. I'm sure that none of these companies has a policy saying that Black people can't become CEOs. Discrimination is a bit more subtle than that. Let's say that you become CEO by being hired at an entry-level position, performing well year after year, getting good performance reviews each time, and gradually getting promoted up the ladder. Also, let's say it takes 20 years of good performance from entry level to CEO. That seems like a realistic number to me.

Crucially, I'm going to assume that managers at this company are a tiny bit discriminatory. Other things being equal, a White employee will get an x% better performance review on whatever numerical scale the company is using. The thing is, this x% adds up year after year. Let's say the company does performance reviews every quarter. I'll come back to that assumption shortly. My question to you is, what does x need to be to add up to a 7-fold or a 700% difference? Maybe 50%? No. I did the math, and it's just 2.5%.

7x inequality in CEO demographics
could be produced by the compounding effect of
2.5% bias in quarterly performance reviews.

A 2.5% difference is so subtle that if you tried to measure it quantitatively using a corpus of performance reviews, you'd need a huge sample size to estimate the effect with any confidence. I did some back-of-the-napkin calculations and you'd need a corpus of tens of thousands of performance reviews. If the performance reviews were twice yearly instead of quarterly, you'd need a 5% difference for it to add up to a 7-fold difference over time. 5% is still very small. On the other hand, maybe even once a quarter is too generous. We aren't just judged during performance reviews. We're judged every day, not just by our managers but also by our peers. Every micro-aggression, every little thing adds up. This kind of pervasive discrimination is far below the threshold that's detectable by quantitative methods.

This is the concept of compounding inequality. Of course, in reality, the opportunities available to us compound not just within the course of a job at a particular company, but throughout our lives and in fact over generations. Today, the average per-capita wealth of Black people is one-sixth that of white people. Professor Ellora Derenoncourt and her coauthors have traced that disparity all the way back to its roots at the end of slavery, showing how Black and White Americans not only obviously didn't start from the same starting line, but have had unequal opportunities for wealth accumulation in the last 150 years.[6] They argue that the gap is unlikely to ever close on its own.

Snapshot datasets hide discrimination

So compounding inequality introduces inherent difficulties for quantitative methods, but the issue is made much worse by what I call the problem of snapshots. Most datasets are snapshots: they are collected from a single system at a single point in time. Of course, quantitative researchers are constrained by datasets. It's very rare for someone to go out and collect their

[6] Derenoncourt et al. *Wealth of Two Nations: The U.S. Racial Wealth Gap, 1860-2020*. 2022.

own dataset, and extremely common to repurpose existing datasets for purposes beyond the one originally envisioned.

In fact, here's a dirty secret. Most senior quantitative researchers, myself included, will advise our students to tailor their research questions to take advantage of the datasets that are available to them. If you first specify your research question and then cast about for a dataset, 9 times out of 10 you will fail. The tail absolutely wags the dog.

What does the problem of snapshots mean for the study of discrimination? It forces the researcher to ignore people's broader circumstances and the past discrimination they may have experienced, because that's not recorded in the data. And it frames discrimination as happening at discrete moments in time rather than encoded into the way that our institutions are designed. In other words, it has real trouble identifying systemic and structural discrimination.

It gets worse. Who produces the data? Usually it's the very companies or organizations which we suspect might be discriminating. Even when external researchers gather data by interacting with the company, they are constrained by the types of interactions that the company makes possible. So when companies are in control of producing data, they have simple ways of affecting the conclusions that are drawn by controlling which data are collected or released.

The very organizations we suspect might be discriminating tend to produce the data.

As a simple example, when tech companies faced criticism about how few Black, minority, and women engineers they employed, they released numbers on their overall workforce composition. Of course, engineering positions at these companies have the highest status and are the best paid, so without the breakdown by position type the numbers don't mean much. Why engineering is paid so much more is also a separate and important conversation. Anyway, I think this point is well known, so I won't dwell on it too much. **Data aren't inert and objective: they are political; they are produced by people or entities towards certain ends.**

Explaining away discrimination

Getting back to research, it's not like labor market discrimination has never been studied. In fact, there are hundreds of studies of it. As I've mentioned, they're only able to measure discrimination — even by their narrow definition, not structural discrimination — when it is so egregious that the effect size is large. One famous study was conducted by economists Bertrand and Mullainathan two decades ago.[7] They sent in fictitious resumes in response to job ads. They wanted to test if an applicant's race had an impact on the likelihood of an employer inviting them for an interview. They signaled race in the resumes by using White-sounding names (Emily, Greg) or Black-sounding names (Lakisha, Jamal). They created pairs of resumes that were identical except for the name. What they found is that White names were 50% more likely to result in a callback than Black names. The magnitude of the effect was equivalent to an additional eight years of experience on a resume.

That's really stark. But consider this. What does it mean to create pairs of resumes that are identical except for the name? Literally every other variable that might signal race or be correlated with race is held constant between the two conditions. That includes things like the applicant's residential neighborhood. So if some employers discriminate against applicants who live in the "wrong part of town", this study wouldn't pick up on it. So even a study like this in many ways drastically underestimates discrimination.

This is the norm in audit studies. Another famous study tested race and gender discrimination by car salespeople when customers bargain for a car.[8] Surprise, surprise, they quoted higher prices for the same cars to Black people. But the way they studied this was to have Black and White testers bargain with different salespeople, behaving identically, including wearing the same outfits. Why? What if attire is one of the ways racial discrimination operates? In other words, even when it comes to identifying interpersonal discrimination rather than structural discrimination, audit studies end up with a narrow definition.

---

[7] Bertrand & Mullainathan. *Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination*. American Economic Review, 2004.

[8] Ayres & Siegelman. *Race and gender discrimination in bargaining for a new car*. The American Economic Review, 1996.

Academic incentives also have a role to play here. Researchers get major brownie points for being clever. That's how you get your papers accepted, get promoted and get tenure. And one of the main ways you can be clever when testing for discrimination is to think of some "omitted variable" that previous researchers haven't thought of that might "explain" some of the discrimination. A simple example would be to say, "oh no no no, employers aren't racially biased, they just have a preference for some universities over others" ... such as never hiring from HBCUs.

The net result of this is that researchers in this area spend so much of their time and energy on an activity that is tantamount to finding excuses to explain away discrimination. **They end up controlling for the attributes that together constitute the social construct of race.** They exclude from consideration most of the ways in which racial discrimination might actually operate.

Case study: the gender pay gap on Uber

Here's an example that brings together most of what we've talked about so far. It's a paper called "The Gender Earnings Gap in the Gig Economy: Evidence from over a Million Rideshare Drivers."[9] It's a paper about gender discrimination rather than racial discrimination, but everything we've talked about is still applicable.

The paper analyzes data from Uber and several of the authors are Uber employees. They start with the observation that female drivers earn 7% less on Uber for every hour they drive than male drivers do. Their goal is to test whether this is due to discrimination. The paper concludes that this gap can be explained by three factors: gender differences in drivers' choices of where to drive, men's greater experience on the platform, and men's tendency to drive faster. In other words, there is no gender discrimination by Uber riders or by the Uber algorithm.

Here's what I find surprising about this paper. It notes in passing that female drivers are 2.7 times as likely to drop off the platform as male drivers are. This observation apparently merits

---

[9] Cook et al. *The Gender Earnings Gap in the Gig Economy: Evidence from over a Million Rideshare Drivers*. National Bureau of Economic Research, 2018.

no further investigation. **The paper spends 71 pages investigating a 7% gap but completely ignores a 170% gap.** One would think that if there is rider discrimination, it would be most apparent in its effect on dropout rates. In contrast, the only avenue of discrimination considered in the paper goes like this: a rider requests a ride, is automatically assigned a driver (Uber doesn't let riders pick drivers), and upon seeing the name of the assigned driver, is apparently so misogynistic that they will cancel the ride if it's a female sounding name, even at the cost of incurring delays to find a new driver and penalties by the algorithm if they cancel too many rides.

This paper illustrates all the pitfalls I've discussed so far. The first is the null hypothesis. Since the default assumption is that disparities are explained by anything but discrimination, the authors zoomed right past the 170% difference in dropout rates. Since they are using a snapshot dataset and didn't have ready-made data to interrogate what might cause the dropout rate disparity, it's not considered their responsibility to dig further. For example, they don't have data on harassment complaints; they don't have data on what else was going on in people's lives that might have helped them understand why someone dropped off the platform. Instead they come up with a tortured hypothesis for how discrimination might operate. It involves a presumably misogynistic rider who cancels a ride, incurring delays and potentially algorithmic penalties, based solely on the driver's gender.

Even limiting ourselves to the 7% difference in earnings, the whole paper can be seen as an example of explaining away discrimination. Here are the 3 factors that the authors identify, that allow them to conclude that discrimination is not the explanation. The first is gender differences in choices of where to live and drive. But surely part of that is because some neighborhoods aren't safe for women. The second reason is that men have greater experience on the platform. Surely, part of the reason for that is that some women face harassment. The third is that men drive faster. The authors frame this as simply a preference, but surely part of the story is that if women drove above the speed limit they would be perceived as aggressive and face social penalties in our culture... very concretely in the form of lower ratings in the app.

All of this also illustrates why the authors' explanations aren't a good guide to interventions (and this is typical of a lot of quantitative work). You'd either conclude that there is no reason to intervene, or you could conclude that there's something wrong with women drivers that needs

to be fixed, like maybe they should drive faster. You'd miss all the structural factors like safety, you'd miss the harassment that women experience, you'd miss the gender stereotypes that penalize women for perceived aggressiveness.

Now, Uber is a company that is known to have a corrupt relationship with academics. The corruption is on both sides; it's not only Uber that's to blame. In many cases it was a straightforward matter of money changing hands. But what we should appreciate is that money might not even be necessary. Selective data availability alone, combined with the blinkers that quantitative researchers are trained to wear, is sufficient to produce this kind of one-sided research.


The Uber files

Uber paid academics six-figure sums for research to feed to the media

Felicity Lawrence

Tue 12 Jul 2022 01.00 EDT

High-profile professors in Europe and the US were engaged as part of lobbying campaign, leak shows


Uber and the Sherlock Holmes Principle: How Control of Data Can Lead to Biased Academic Research

BY LUIGI ZINGALES October 9, 2019


Uber's "Academic Research" Program: How to Use Famous Economists to Spread Corporate Narratives

BY HUBERT HORAN December 5, 2019


WORLD ECONOMICS ASSOCIATION

How UBER Money Dominates and Distorts Economic Research on Ride–Hailing Platforms

I should know. I myself once almost became a pawn in one of Uber's schemes. Many years ago I was approached by the company seeking my expertise in privacy. I was known for my research in data "deanonymization". I'd developed a set of algorithms by which sensitive data about people, even if it is stripped of names and other identifiers, can be linked to other public datasets, allowing an adversary to put the identities back. So anonymizing data doesn't protect its privacy. Uber wanted me to test this theory on their data: data about the locations of people's trips. I was excited, because I'd never worked with real-world location data before, and this was a rare chance. But the project fell through because of administrative roadblocks.

Only later did I realize what Uber's true motivation was. It's not because Uber was considering releasing data publicly, which would indeed incur genuine privacy risks. Regulators were coming after them to provide this data. As you know, Uber and other companies, at least at that time, operated in a gray area of the law, and regulators needed this information to do their jobs. Uber weaponized privacy as a way to stall them, and would have used my report highlighting the privacy risks to their advantage. I'm embarrassed to say that I'd never been trained to ask a basic critical question: why is someone so eager to let me analyze this valuable dataset? What's in it for them?

Quantitative methods can't resolve conflict between values

As you probably know, quantitative researchers are trained to believe that our work is objective. That any two researchers will arrive at the same answer to the same question. I hope it's obvious that that's not the case. In a typical research paper I would say researchers make at least 10-20 subjective choices, each of which could substantially alter their conclusions.

The belief in objectivity is so strong that for many years, the community of computer scientists studying algorithmic discrimination tried to achieve consensus on a single mathematical definition of what fairness means and what discrimination means. In other words, not only must there be a single answer to a given question, but a single mathematical way to ask a question, in this case whether a system is fair or discriminatory.

This effort at finding a consensus definition was not successful. There have been more and more statistical non-discrimination criteria over time, each of which captures some small part of our normative understanding of fairness, but is far from the full picture. Back in 2018 I gave a talk titled 21 definitions on fairness and their politics, in which I traced some of these connections.[10] That's probably my best known work in this area.

The lesson here is that scientific and quantitative methods cannot possibly resolve debates that arise from conflict between values. We should be skeptical of all claims to mathematical objec-

---

[10] Narayanan. *Translation tutorial: 21 fairness definitions and their politics.* In Proc. Conf. Fairness, Accountability and Transparency, 2018.

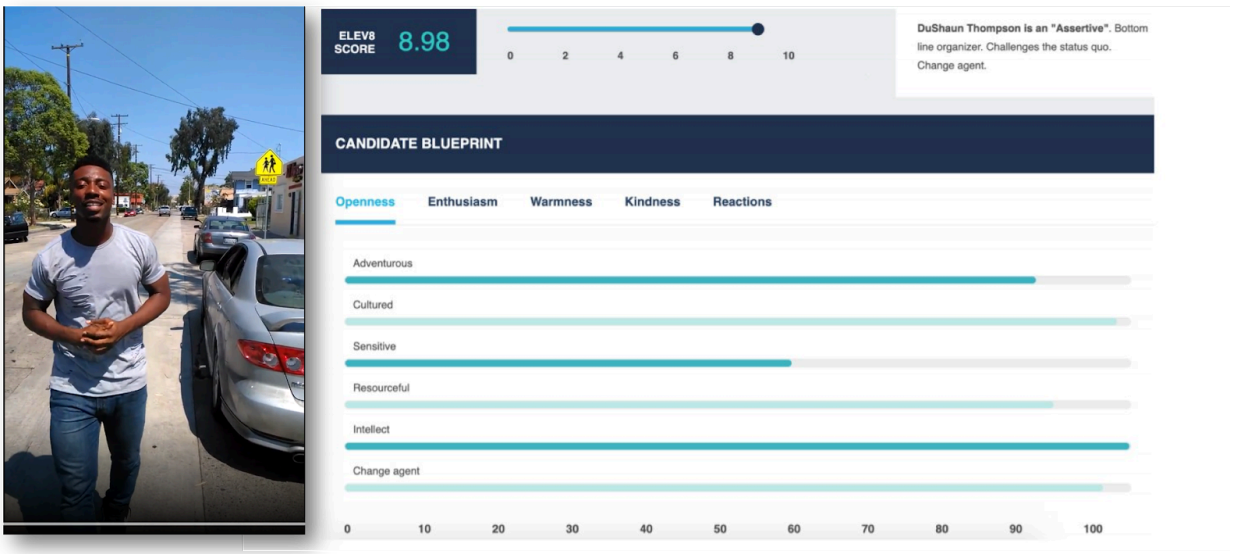
tivity, but we should be doubly skeptical when such methods promise an alternative to democratic debate about our values. That process is slow, messy, and painful, but it is necessary.

For better or worse, numbers are the language of policy making

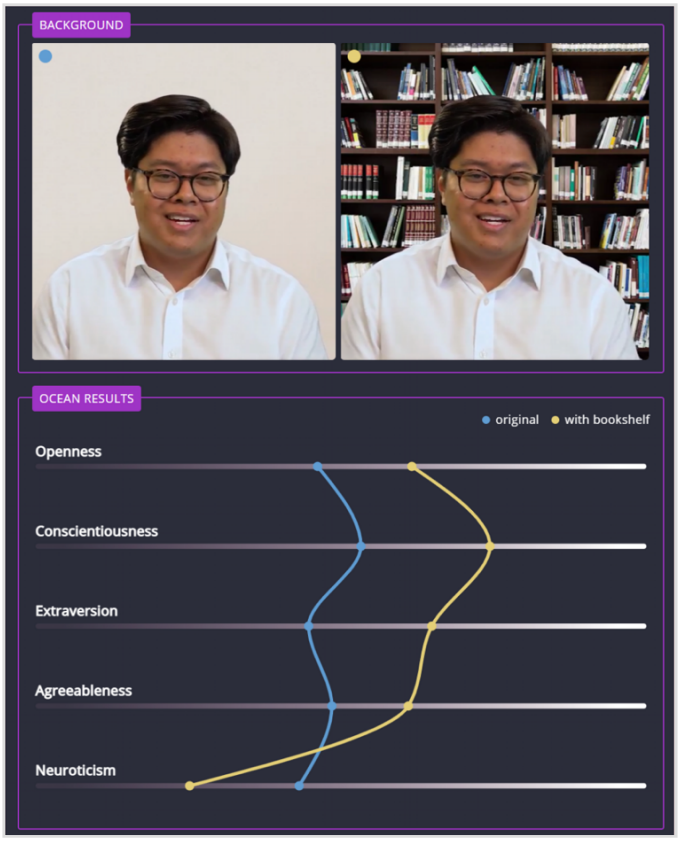
It might feel like I'm airing dirty laundry in this lecture. Many disciplines have methodological crises, but they have ways of gradually resolving them and making progress through internal dialog. But there is something in this case that goes beyond the usual methods debates, and I think people outside the research community need to pay attention.

The issue is that policy makers take quantitative evidence far more seriously than other forms of evidence. Numbers have been the language of policy making for more than a century, but especially so today, with the tech industry being so successful at convincing the public about the power of big data and AI and what not.

So quantitative researchers feed policy makers with evidence about discrimination that's based on a very narrow understanding of injustice and oppression. What's worse is that over time, policy makers actually start to imbibe this way of looking at discrimination, and this becomes the operative definition of discrimination. One term for this is **performativity**.



Let me give you an example. In hiring, tools like this have become extremely common over the last five years. This is an actual video and screenshot from marketing materials of an actual company. These systems claim to predict someone's job performance by analyzing not even what the candidate said, but rather body language, speech patterns, etc. In this example, the candidate uploaded a 30 second video where they talk about their hobbies and stuff like that, and the AI system has apparently categorized them as a "change agent"... along with various other scores in 5 different categories, for an overall score of 8.98. I love the two digits of precision! If you suspect that tools like these don't work, you'd be correct. I've called them elaborate random number generators.[11]



In fact, they're random number generators in the *best case*. More likely what's happening is that they're just picking up on stereotypes. The screenshot shows an experiment that reveals

[11] Narayanan. *How to recognize AI snake oil.* Arthur Miller lecture on science and ethics, 2019.

that the presence of a bookshelf substantially alters the scores.[12] So does the presence of glasses and various other irrelevant characteristics, it turns out.

What are the harms from using hiring tools that don't work? First of all, it is demeaning to candidates to be judged by a machine, especially one that understands nothing about the actual job. It disrespects the candidate and the time they have put into preparing for the labor market. Besides, they might not be selected for a job they are well qualified for, be given no explanation, and have no recourse — no way to do better next time to improve their chances.

There are even anecdotes of job seekers being repeatedly screened out of jobs on the basis of personality tests, all offered by the same vendor.[13] Note that these snake oil tools are primarily used for low-status positions that tend to have a higher proportion of lower income and minority applicants.

Injustices other than disparate impact seem illegible to regulators.

You'd hope that regulators like the Equal Employment Opportunity Commission and the Federal Trade Commission are cracking down. Unfortunately, none of these injustices I've described seem to be legible to regulators. In my experience, the main thing they care about is disparate impact. In other words, if a company adopts this kind of tool to screen their applicant pool, are they less likely to hire a Black applicant than a White applicant. Disparate impact is so popular because it is amenable to quantitative methods using snapshot datasets. Because of that, it has been heavily promoted as the definition of discrimination by many stakeholders, including by the companies themselves. Note that disparate impact is relatively easy to fix in these tools... especially so because they are essentially random number generators. And so these tools proliferate. By the way, I'm co-authoring a book called AI Snake Oil with Sayash Kapoor where you can read more about this and other dubious uses of algorithmic decision making.

---

12 via Bayerischer Rundfunk.

13 O'Neil. *How Algorithms Rule Our Working Lives*. The Guardian 2016.

To summarize, here are seven limitations of quantitative methods for studying discrimination as they are used today.

1. Choice of null hypothesis
2. Use of snapshot datasets
3. Use of data provided by company
4. Explaining away discrimination
5. Wrong locus of intervention
6. The objectivity illusion
7. Performativity

Is there hope?

Let me try to end on a positive note. Actually, I'm not sure how positive it will be, but let me at least try to end on a constructive note.

First of all, I don't think giving up quantitative work is an option. If nothing else, because of the outsize emphasis that policy makers place on quantitative evidence. Ceding the turf to researchers and companies who will employ even less critical thinking is not going to solve anything. And besides, almost none of the limitations I've pointed out are inherent to quantitative work. I genuinely think a better way is possible.

Inequality and discrimination are not separable

Here's the #1 thing that needs to change. Currently the main focus of the field is to separate inequality from discrimination, to try to identify the discrete moments when discrimination happens, see how much they add up to, and figure out the extent to which inequality is due to discrimination.

This is a fool's errand. Inequality and discrimination are not separable. They're not practically separable — recall the CEO example. They're not even conceptually separable — recall the gender pay gap on Uber. One way to view discrimination is that it's simply the way inequality perpetuates itself. Any time you see inequality perpetuating itself, that's discrimination in action. Anyway, regardless of what you call it, that phenomenon is worth studying. The main focus of the field should be to study the dynamics of inequality.

Rich families in Florence 600 years ago are also richer today.

To do this we need to assemble the right datasets. With the right datasets, quantitative methods can really shine. For example, one study showed that rich families in medieval Florence 600 years ago are also richer today, based on surnames and tax records.[14] But this was more of a one-off study made possible by a chance dataset rather than the fruit of a systematic research program to study wealth transmission through the ages. Earlier I mentioned Professor Derenoncourt's work on building datasets to track wealth accumulation post slavery. That's the kind of work we need more of.

To do this right, **we should be spending most of our time on curating and interrogating datasets**, rather than opportunistically running with whatever is available. In this vein, I want to highlight another Princeton example: Fragile Families and Child Wellbeing study that has been following the life course of disadvantaged children and their families for over 20 years. It's incredibly painstaking work that required tens of thousands of hours of researcher time but that's the kind of investment we need to do this properly. Even with this investment of effort, the scale of the dataset is only on the order of thousands of families. That's far less than commercial datasets, but it's a necessary tradeoff.

The case for descriptive work

In addition to collecting, curating, and interrogating datasets, there's a lot of value in simply describing what's in the data. No fancy causal inference, not even any statistical models. You

[14] Barone & Mocetti. *What's your (sur)name? Intergenerational mobility over six centuries.* 2016.

can see a pattern here — the activities I'm suggesting are those that aren't prioritized in the quantitative world because they aren't considered technically sophisticated. But they are the ones that are most valuable.

I learned this lesson the painful way. When I joined Princeton's faculty a decade ago, I recruited a team of amazing graduate students to study discrimination on the web. We wanted to track where our personal data is going when we use websites and apps, and how companies use that in ways that are potentially harmful: racially discriminatory ads, recommendation algorithms that put us in filter bubbles, price discrimination, and so forth.

That work was a disaster. Causal inference methods were too brittle to detect the effects we wanted to detect. Others have later successfully done some of what we wanted to do, but at the time the methods were just not there. I wasted about two years of everyone's time. If we hadn't decided to pull the plug and change direction, I wouldn't have gotten tenure and the students may not have gotten their PhDs.

But what we realized is that in the process of trying to do this, we'd collected a lot of data showing that our websites and apps were teeming with hidden trackers watching what we search, browse, and shop for. People suspected this, of course, but until then the data showing evidence of this hadn't been assembled at the scale that we had done. So we decided to tell that story instead. We ended up writing about a dozen papers about the spying that is constantly happening on our devices. We didn't have to use any fancy statistical techniques and we had to face peer reviewers saying this is just measurement, it's not novel and doesn't deserve to be published. But we persisted and kept pointing out why data work was hard and what a big social challenge it could help address.

I'm happy to say that that work has had an impact. It's had an impact both on companies like Apple that build operating systems and browsers: it's helped convince them to crack down on tracking. It's also helped motivate a wave of privacy legislation in many parts of the world. Not just the work we did, of course, but similar descriptive work that other researchers and investigative journalists did.

Separating individual experiences from collective oppression

One important role for quantitative methods is to distinguish between individual experiences and collective oppression, in the words of feminist sociologist Ann Oakley.[15] The book Data Feminism recounts a story of a screenwriter in the UK who met a quantitative researcher who was collecting data on the demographics of the industry.[16] The woman screenwriter in question was surprised to hear that only 20% of screenwriters in the UK were women. She noted that screenwriters never get to meet each other and had never considered the question of gender composition.

It's true that quantitative work can help provide a collective voice, but it's not the only thing or even the most powerful thing that technologists can do to help. Building tools for screenwriters to come together with each other would be at least as useful; it would help women screenwriters connect with each other and amplify their own voices and build solidarity. On a related point, Ruha Benjamin describes many abolitionist technical tools in her wonderful book Race After Technology.[17]

Here's something I personally find to be a compelling example of quantitative work that illuminates collective oppression, and it goes back to an example I opened with: the fact that there are so few Black CEOs. It's called the **glass cliff** phenomenon.[18] Not the glass ceiling, but the glass cliff. Researchers have found that women and ethnic minorities are more likely to be promoted to CEO when the firm is struggling. And if the firm performs poorly during their tenure — which should be unsurprising if the firm was already struggling — then they tend to be replaced by white men. There's a lot of evidence for this and there is a multitude of papers. My guess is that this is the sort of discrimination that wouldn't necessarily be obvious even to

---

15 Oakley. *Paradigm Wars: Some Thoughts on a Personal and Public Trajectory*. 1999.

16 Wreyford & Cobb. *Data and Responsibility: Toward a Feminist Methodology for Producing Historical Data on Women in the Contemporary UK Film Industry*. 2017.

17 Benjamin. *Race After Technology: Abolitionist Tools for the New Jim Code*. Polity Press, 2019.

18 Ryan & Haslam. *The Glass Cliff: Evidence that Women are Over-Represented in Precarious Leadership Positions*. BJM 2005.; Cook & Glass. *Above the glass ceiling: When are women and racial/ethnic minorities promoted to CEO?* Strategic Management Journal 2013.

someone facing it, because of how devious it is, but I could be wrong. Anyway, I would like to hold this up as a positive example of quantitative analysis of discrimination. It allows us to see the sometimes subtle mechanisms by which it operates.

Statistical disparities are symptoms, not the disease

One simple but important point to keep in mind is that if we find that there is a statistical disparity in some decision making system, fixing that disparity by itself might not achieve much. Usually, what it is pointing to is that there is some deep systemic injustice. Let's go back to the ProPublica example. If the developers of risk prediction algorithms redesigned them to equalize the rates of falsely flagging someone as high risk, between Black defendants and white defendants, that doesn't solve the problem. It remains profoundly unjust to deny someone their freedom based on a prediction that they might commit another crime or a prediction that they might not appear in court for their arraignment or their trial. Especially so when those predictions are only slightly better than random, and even more so once we realize that the reasons someone might not appear in court are often due to their circumstances such as difficulty finding childcare.

One study looked at why it is that so many defendants in NYC fail to appear in court when they are issued a summons for low-level crimes. It turned out that one of the major reasons was that the summons form was too confusing and people couldn't figure out what was asked of them! By simply redesigning the summons and texting people reminders of their court dates, they were able to drastically reduce the rate at which people failed to appear in court.[19] What an impactful example of quantitative work.

Centering the experiences of those harmed

Finally, the best way for researchers to avoid overly narrow conceptions of the problem is to have empathy for those who are harmed. Going back to criminal risk prediction tools, most

[19] Fishbane, Ouss, Shah. *Behavioral nudges reduce failure to appear for court*. Science 2020.

quantitative researchers studying racial bias in these tools never actually meet any defendants who have been harmed by the criminal justice system. The conference on Fairness Accountability and Transparency in sociotechnical systems has been looking to change this. It's the main meeting of the algorithmic fairness research community.

In the 2018 version, there was a panel featuring Terrence Wilkerson, a person twice falsely accused of armed robbery and spent a total of two years in pre-trial detention, including in Rikers island.[20] He spoke about his experiences with the criminal justice system and how pre-trial detention affected his life. Since then, the participation of those harmed by technological systems has become a regular feature of the conference. In the latest edition there was a community keynote involving gig workers who organized themselves in partnership with researchers to protest and contest an unjust algorithm that determined their pay.

This is admirable, but it should be considered the bare minimum. Ideally, researchers should spend some time being embedded in the communities that they claim to study and ultimately serve.

Three take-aways

We're almost at time. Let me summarize briefly. I have three take-ways.

If you're a quantitative researcher, I urge you to recognize that currently quantitative methods are primarily used to justify the status quo. I would argue that they do more harm than good. But a different way is possible. We need to let go of our idea of an epistemic hierarchy where some forms of evidence are superior to others. In an ideal world, the role of quantitative methods will be limited and it will be just one among many ways of knowing. But that's ok, and it's certainly better than where we are today.

By the way, this point applies not just to the study of discrimination. I gave a talk making exactly the same points about the overconfidence of another quantitative research community,

[20] Bender, Lum, Wilkerson. *Translation Tutorial: Understanding the Context and Consequences of Pre-trial Detention*. FAT* 2018.

namely those studying the effects of social media algorithms on society — polarization and echo chambers and that sort of thing. You can find the talk on my website; it's called "Is there a filter bubble on social media? A call for epistemic humility."

For qualitative researchers and critical theorists: I'm very much on board with critiques of quantitative research, but to be able to change behavior among my community, I think a bit of technical specificity would go a long way. I hope this talk has given you some ideas on how to do that.

My third takeaway is for everyone, and it is simply this. Approach all quantitative claims, especially claims about discrimination, with caution. Always dig deeper. Behind the facade of sophisticated formulas and regressions there are usually crude assumptions about the world and datasets whose politics haven't been interrogated. It's important to get to that layer in order to understand any statistic that you come across. Stripped of this context, numbers by themselves have no meaning.

Thank you very much.