

DATA ETHICS LECTURE 3

RECAP PROMISES

PREVIEW MORE PROMISES

Ali Alkhatib

@_alialkhatib || hi@al2.in

March 24, 2022

ROADMAP FOR TODAY

- Admin?
- Recap Promises
- Preview other promises (tedious/dangerous & social good/token stakeholders)

ADMIN STUFF?

- reading time/reflection

ADMIN STUFF?

- reading time/reflection
- let's do that again →

check the zoom chat for a link

PROMISES

ROADMAP

- AI can “understand” the world better than we can
- AI can make “fairer” decisions than we can

UNDERSTANDING

**AI WILL “UNDERSTAND” BETTER THAN
WE CAN**

AI WILL UNDERSTAND MORE

→ labeling

The New York Times

SUBSCRIBE FOR \$1/WEEK

Facebook Apologizes After A.I. Puts 'Primates' Label on Video of Black Men

Facebook called it "an unacceptable error." The company has struggled with other issues related to race.

Give this article

1 Hacker Way

AI WILL UNDERSTAND MORE

→ labeling



wired.com/story/when-it-comes-to-gorillas-google-photos-remains-blind/

≡ WIRED BACKCHANNEL BUSINESS CULTURE GEAR IDEAS SCIENCE SECURITY SIGN IN

TOM SIMONITE BUSINESS JAN 11, 2018 7:00 AM

When It Comes to Gorillas, Google Photos Remains Blind

Google promised a fix after its photo-categorization software labeled black people as gorillas in 2015. More than two years later, it hasn't found one.

f t e



AI WILL UNDERSTAND MORE

→ labeling

The screenshot shows a web browser window with the MIT Technology Review logo at the top. The main headline reads "ARTIFICIAL INTELLIGENCE" followed by "Neural Network Learns to Identify Criminals by Their Faces". Below the headline is a subtext: "The effort aimed at identifying criminals from their mugshots raises serious ethical issues about how we should use artificial intelligence." The author is listed as "By Emerging Technology from the arXiv" and the date is "November 22, 2016". At the bottom left, there is a snippet: "Soon after the invention of photography, a few criminologists began to notice patterns in mugshots they took of criminals. Offenders, they said, had particular facial features that allowed them to be identified as law breakers." On the right side, there is a "POPULAR" section with a link to an article about a man communicating through thought.

ARTIFICIAL INTELLIGENCE

Neural Network Learns to Identify Criminals by Their Faces

The effort aimed at identifying criminals from their mugshots raises serious ethical issues about how we should use artificial intelligence.

By Emerging Technology from the arXiv
November 22, 2016

Soon after the invention of photography, a few criminologists began to notice patterns in mugshots they took of criminals. Offenders, they said, had particular facial features that allowed them to be identified as law breakers.

One of the most influential voices in this debate was Cesare Lombroso, an Italian criminologist, who

POPULAR

A locked-in man has been able to communicate in sentences by thought alone

Jessica Hamzelou

AI WILL UNDERSTAND MORE

- labeling
- moderation

INSIDER

HOME > TECH

TikTok videos that promote anorexia are misspelling common hashtags to beat the 'pro-ana' ban

Naina Bhardwaj Dec 27, 2020, 7:33 AM



The TikTok logo is displayed on a phone in China on March 3, 2020.
Sheldon Cooper/SOPA Images/LightRocket via Getty Images

▪ TikTok said it banned six accounts reported to it for

AI WILL UNDERSTAND MORE

- labeling
- moderation
- navigation

The New York Times

Snow Closed the Highways. GPS Mapped a Harrowing Detour in the Sierra Nevada.

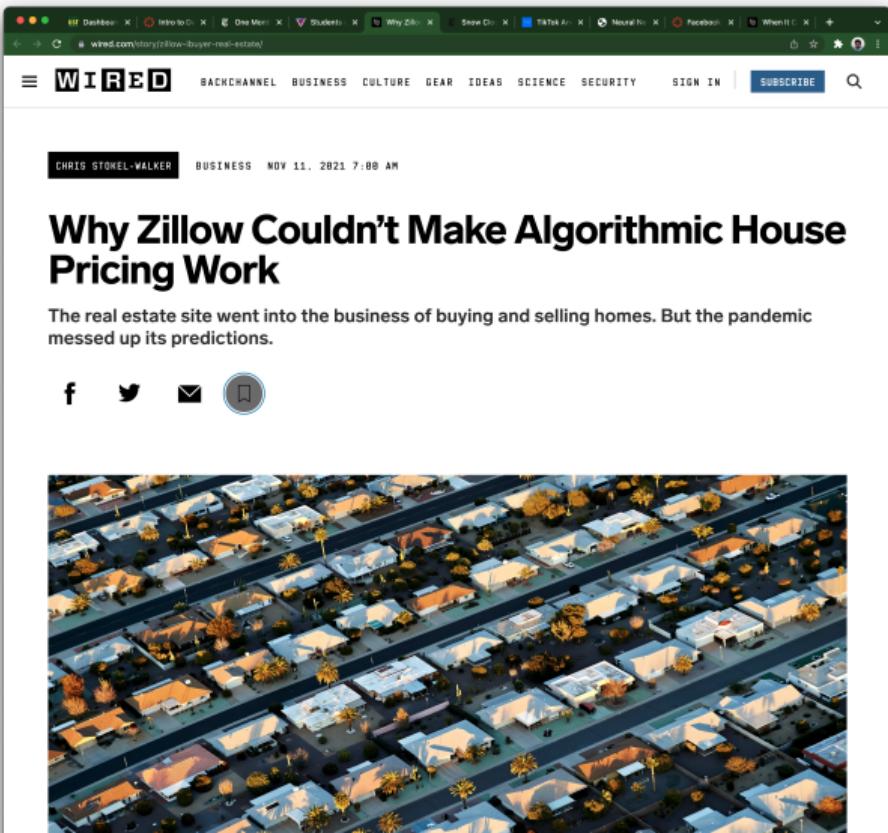
Public safety officials warned that alternate routes offered by apps like Google Maps and Waze don't always take into account hazards to drivers.

Give this article



AI WILL UNDERSTAND MORE

- labeling
- moderation
- navigation



CHRIS STOKEL-WALKER BUSINESS NOV 11, 2021 7:00 AM

Why Zillow Couldn't Make Algorithmic House Pricing Work

The real estate site went into the business of buying and selling homes. But the pandemic messed up its predictions.

f t e 



AI WILL UNDERSTAND MORE

- labeling
- moderation
- navigation

The Impact of Crowd Work on Workers
CHI 2015, Crossings, Seoul, Korea

Working with Machines: The Impact of Algorithmic and Data-Driven Management on Human Workers

Min Kyung Lee¹, Daniel Kusbit¹, Evan Metsky¹, Laura Dabbish^{1,2}
¹Human-Computer Interaction Institute, ²Heinz College
Carnegie Mellon University
{mklee, dkusbit, emetsky, dabbish}@cmu.edu

ABSTRACT
Software algorithms are changing how people work in an ever-growing number of fields, managing distributed human workers at a large scale. In these work settings, human jobs are assigned, optimized, and evaluated through algorithms and tracked data. We explored the impact of this algorithmic, data-driven management on human workers and work practices in the context of Uber and Lyft, new ridesharing services. Our findings from a qualitative study describe how drivers responded when algorithms assigned work, provided informational support, and evaluated their performance, and how drivers used online forums to socially make sense of the algorithm features. Implications and future work are discussed.

Author Keywords
Algorithm; algorithmic management; human-centered algorithms; intelligent systems; CSCW; on-demand work; sharing economies; data-driven metrics; work assignment; performance evaluation; dynamic pricing; sensemaking.

ACM Classification Keywords
H.5.m. Information interfaces and presentation (e.g., HCI); Miscellaneous.

INTRODUCTION
Increasingly, software algorithms allocate, optimize, and evaluate work of diverse populations ranging from traditional workers such as subway engineers [16], warehouse workers [28], Starbucks baristas [19], and UPS deliverymen [7] to new crowd-sourced workers in platforms like Uber, TaskRabbit, and Amazon mTurk [13]. How do human workers respond to these algorithms taking roles that human managers used to play?

We call software algorithms that assume managerial

As a first step toward answering these questions, we interviewed 21 drivers with Uber and Lyft and triangulated

AI WILL UNDERSTAND MORE

- labeling
- moderation
- navigation

The screenshot shows a Mac OS X-style window displaying a PDF document. The title bar reads "utopia [2].pdf (page 1 of 14)". The main content area of the PDF is visible, showing the following text:

To Live in Their Utopia: Why Algorithmic Systems Create Absurd Outcomes

ALI ALKHATIB, Center for Applied Data Ethics, University of San Francisco

The promise AI's proponents have made for decades is one in which our needs are predicted, anticipated, and met - often before we even realize it. Instead, algorithmic systems, particularly AIs trained on large datasets and deployed to massive scales, seem to keep making the wrong decisions, causing harm and rewarding absurd outcomes. Attempts to make sense of why AIs make wrong calls in the moment explain the instances of errors, but how the environment surrounding these systems precipitate those instances remains murky. This paper draws from anthropological work on bureaucracies, states, and power, translating these ideas into a theory describing the structural tendency for powerful algorithmic systems to cause tremendous harm. I show how administrative models and projections of the world create marginalization, just as algorithmic models cause representational and allocative harm. This paper concludes with a recommendation to avoid the absurdity algorithmic systems produce by denying them power.

CCS Concepts • Human-centered computing → HCI theory, concepts and models.

Additional Key Words and Phrases: HCI, Artificial Intelligence, Street-Level Algorithms

ACM Reference Format:

Ali Alkhathib. 2021. To Live in Their Utopia: Why Algorithmic Systems Create Absurd Outcomes. In *CHI Conference on Human Factors in Computing Systems (CHI '21), May 8–13, 2021, Yokohama, Japan*. ACM, New York, NY, USA, 14 pages. <https://doi.org/10.1145/3441764.3445740>

1 INTRODUCTION

HCI researchers have spent years working to improve algorithmic systems, and increasingly systems that produce computational models generated by Machine Learning (ML), that designers often use at enormous scales to classify and make difficult decisions for us. Some of that work is exploratory, finding new places and ways to use technologies, and new insights that AI might yield when ML is applied to massive datasets to find relationships in the data [29, 61, 76]. Other work surfaces problems with existing systems and attempts to mitigate those harms (for instance, by making them more fair, accountable, and transparent) [4, 42, 46, 47, 53]. Then there's work that tries to establish productive theoretical frameworks describing the social environments these systems produce and that designers create and foster, in the hope that some ontology or paradigm will motivate theoretically-grounded discussions about where the first two threads of research ought to lead [30–32, 37, 50, 72].

Part of the challenge of all this seems to be that the future we've imagined and promoted for decades, as designers of technical systems, is woefully misaligned from people's experiences of massive computational systems. Many of these algorithmic systems, especially ML systems, cause substantial harms in myriad domains, often surprising the designers of those systems.

Designers of sociotechnical systems have repeatedly built computational systems and models rendering decisions

**WHAT DOES IT MEAN TO UNDERSTAND
SOMETHING?**

UNDERSTANDING

Figuring out what facets you're missing *a priori* is
impossible

UNDERSTANDING

Figuring out what facets you're missing *a priori* is
impossible

How would you **elicit** this dimension as a designer?

What factors prevent or discourage this kind of approach in
corporate settings?

FAIRNESS

**AI WILL BE “FAIRER” OR “MORE
OBJECTIVE” THAN WE CAN BE**

AI WILL BE FAIRER THAN WE CAN BE

→ medicine

The screenshot shows a web browser window displaying an article from WIRED.com. The URL in the address bar is www.wired.com/story/how-an-algorithm-blocked-kidney-transplants-to-black-patients/. The page header includes the WIRED logo and navigation links for BACKCHANNEL, BUSINESS, CULTURE, GEAR, IDEAS, SCIENCE, and SECURITY, along with SIGN IN and SUBSCRIBE buttons. Below the header, the author is listed as TOM SIMONITE and the date is OCT 26, 2020 7:00 AM. The main title of the article is "How an Algorithm Blocked Kidney Transplants to Black Patients". A subtitle below the title reads: "A formula for assessing the gravity of kidney disease is one of many that is adjusted for race. The practice can exacerbate health disparities." At the bottom of the article preview, there are social sharing icons for Facebook, Twitter, Email, and LinkedIn.

How an Algorithm Blocked Kidney Transplants to Black Patients

A formula for assessing the gravity of kidney disease is one of many that is adjusted for race. The practice can exacerbate health disparities.

f t e-mail LinkedIn

The image at the bottom of the article preview shows a medical scan, likely a CT or MRI, of a person's abdomen. Two kidneys are visible, appearing as dark purple organs against a lighter background. Bright yellow areas are highlighted on both kidneys, specifically in the lower poles, which typically represent areas of kidney dysfunction or scarring. This visual metaphor suggests that the algorithm identified these specific areas as problematic, leading to the denial of transplants for Black patients.

AI WILL BE FAIRER THAN WE CAN BE

→ medicine

The screenshot shows a computer window displaying a research article from the journal *nature medicine*. The article is titled "An algorithmic approach to reducing unexplained pain disparities in underserved populations". The authors listed are Emma Pierson^{1,2}, David M. Cutler³, Jure Leskovec^{3,4}, Sendhil Mullainathan^{3,5} and Ziad Obermeyer⁶. The text of the article discusses how underserved populations experience higher levels of pain due to osteoarthritis, and how an algorithmic approach can reduce racial disparities in pain prediction by accounting for factors external to the knee, such as stress. The article also notes that standard radiographic measures like KL grade may miss physical causes of pain in people of color.

ARTICLES
https://doi.org/10.1038/s41591-020-01192-7

nature medicine

An algorithmic approach to reducing unexplained pain disparities in underserved populations

Emma Pierson^{1,2}, David M. Cutler³, Jure Leskovec^{3,4}, Sendhil Mullainathan^{3,5} and Ziad Obermeyer⁶

Underserved populations experience higher levels of pain. These disparities persist even after controlling for the objective severity of diseases like osteoarthritis, as graded by human physicians using medical images, raising the possibility that underserved patients' pain stems from factors external to the knee, such as stress. Here we use a deep learning approach to measure the severity of osteoarthritis, by using knee X-rays to predict patients' experienced pain. We show that this approach dramatically reduces unexplained racial disparities in pain. Relative to standard measures of severity graded by radiologists, which accounted for only 9% (95% confidence interval (CI), 3–16%) of racial disparities in pain, algorithmic predictions accounted for 43% of disparities, or 4.7× more (95% CI, 3.2–11.8%), with similar results for lower-income and less-educated patients. This suggests that much of underserved patients' pain stems from factors within the knee not reflected in standard radiographic measures of severity. We show that the algorithm's ability to reduce unexplained disparities is rooted in the racial and socioeconomic diversity of the training set. Because algorithmic severity measures better capture underserved patients' pain, and severity measures influence treatment decisions, algorithmic predictions could potentially redress disparities in access to treatments like arthroplasty.

Pain is widespread and unequally distributed in society. Like many other causes of pain, knee osteoarthritis, which affects 10% of men and 13% of women over 60 years of age in the United States, disproportionately affects underserved populations: people of color score higher on knee pain scales than do white individuals^{1,2}. Understanding these racial disparities in pain is important for clinical decision making and public policy but also for understanding pain disparities for a variety of other medical problems^{3,4}.

Two explanations for these disparities have been proposed. First, underserved patients might have more severe osteoarthritis within the knee. Alternatively, underserved patients could have more aggravating factors external to the knee. For example, the same physical ailments in different populations can produce very different experienced pain due to life stress, social isolation or other factors^{5,6}. These two explanations have very different treatment implications: psychosocial interventions target causes external to the knee, whereas physical therapy, medication and orthopedic procedures address causes within the knee^{7,8}.

Research has indirectly implicated factors external to the knee. Methodologically, this is demonstrated by defining an

with structural damage on X-ray or even magnetic resonance imaging (MRI) experience no or very little pain^{1,4,16}. Standard radiographic measures such as KL grade, developed decades ago in white British populations, might miss physical causes of pain in people of color^{7,17}; further, there are known racial and socioeconomic biases in how a patient's pain is perceived by observers^{18,19}. If the pain experienced by underserved populations is caused by objective factors missing from current measures, a range of painful, treatable knee ailments would be misattributed to factors external to the knee.

In this paper, we use a machine-learning approach to discriminate between the 'within the knee' and 'external to the knee' hypotheses. We produce a new algorithmic measure of osteoarthritis severity from radiographs alone. We use a dataset of knee radiographs from a diverse sample of 4,172 patients in the United States who had or were at high risk of developing knee osteoarthritis. As part of an NIH-funded study²⁰, bilateral fixed flexion knee radiographs were obtained and scored by radiologists on summary measures of radiographic severity (for example, KL grade) and other objective features (for example, eburnation and joint space narrowing (JSN)). Patients also reported a knee-specific pain score (Knee Injury and Osteoarthritis Outcome Score (KOOS)), derived from a

AI WILL BE FAIRER THAN WE CAN BE

→ medicine

theverge.com/2021/4/8/22374386/proctorio-racial-bias-issues-openai-facial-detection-schools-tests-remote-learning

THE VERGE TECH REVIEWS SCIENCE ENTERTAINMENT MORE

POLICY \ TECH

Students of color are getting flagged to their teachers because testing software can't see them 27

Proctorio reportedly uses facial detection software that failed to recognize black faces over half the time

By Mitchell Clark | Apr 8, 2021, 8:34pm EDT

f t SHARE



AI WILL BE FAIRER THAN WE CAN BE

- medicine
- education

theverge.com/2020/8/17/21372045/uk-a-level-results-algorithm-biased-coronavirus-covid-19-pandemic-university-applications

THE VERGE TECH REVIEWS SCIENCE ENTERTAINMENT MORE

US & WORLD TECH ARTIFICIAL INTELLIGENCE

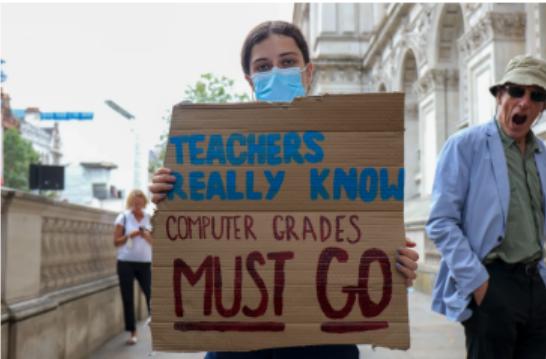
12 ▾

UK ditches exam results generated by biased algorithm after student protests

Protesters chanted 'Fuck the algorithm' outside the country's Department for Education

By Jon Porter | @JonPorty | Aug 17, 2020, 12:16pm EDT

f t SHARE



verge deals

Subscribe to get the best Verge-approved tech deals of the week.

Email (required)

By signing up, you agree to our [Privacy Notice](#) and European users agree to the data transfer policy.

AI WILL BE FAIRER THAN WE CAN BE

- medicine
- education
- social

The screenshot shows a Twitter blog post from the 'Engineering' category. The author is Rumman Chowdhury (@ruchowdh), a Director at Twitter META. The post is titled 'Sharing learnings about our image cropping algorithm'. It was published on Wednesday, 19 May 2021. The post discusses feedback received in October 2020 regarding the image cropping algorithm and the steps taken to address bias, including a bias assessment and improvements to the model. A link to the analysis is provided.

Engineering Insights Infrastructure Open source Sign Up Search

Rumman Chowdhury
@ruchowdh

Director, Twitter META

Only on Twitter
@Twitter
#OnlyOnTwitter

Sharing learnings about our image cropping algorithm

By @ruchowdh
Wednesday, 19 May 2021

In October 2020, we heard feedback from people on Twitter that our **image cropping algorithm** didn't serve all people equitably. As part of our **commitment** to address this issue, we also shared that we'd analyze our model again for bias. Over the last several months, our teams have accelerated improvements for how we assess algorithms for potential bias and improve our understanding of whether ML is always the best solution to the problem at hand. Today, we're sharing the outcomes of our bias assessment and a link for those interested in **reading** and **reproducing** our analysis in more technical detail.

The analysis of our image cropping algorithm was a

AI WILL BE FAIRER THAN WE CAN BE

- medicine
- education
- social

The image is a screenshot of a computer screen displaying a news article from ProPublica. The article is titled "Machine Bias" and features a photograph of two men, Bernard Parker and Dylan Fugate, side-by-side. Bernard Parker, on the left, has dark skin, long dreads, and is wearing a white t-shirt. Dylan Fugate, on the right, has light skin, short hair, and is wearing a black t-shirt. The background is a dark grey gradient. At the top of the screen, there is a green header bar with the ProPublica logo and several tabs. Below the header, there is a red "Donate" button and social media links for Facebook, Twitter, and LinkedIn.

Bernard Parker, left, was rated high risk; Dylan Fugate was rated low risk. (Josh Ritchie for ProPublica)

Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica
May 23, 2016

AI WILL BE FAIRER THAN WE CAN BE

- medicine
- education
- social

The screenshot shows a web browser window displaying an article from The New York Times. The title of the article is "One Month, 500,000 Face Scans: How China Is Using A.I. to Profile a Minority". Below the title, there is a brief summary: "In a major ethical leap for the tech world, Chinese start-ups have built algorithms that the government uses to track members of a largely Muslim minority group." At the bottom of the article page, there is a large image showing a computer monitor displaying the SenseFace Face Recognition Surveillance Platform. The monitor shows multiple video feeds from surveillance cameras, along with various data and graphs related to facial recognition and monitoring.

WHAT IS FAIRNESS?

FAIRNESS

A Mulching Proposal

The image shows a LaTeX document with a title page. At the top right, there is a small thumbnail of the document itself, labeled "mulching.pdf (page 1 of 10)". The main title "A Mulching Proposal" is centered at the top. Below it is a subtitle: "Analysing and Improving an Algorithmic System for Turning the Elderly into High-Nutrient Slurry". There are two author sections: one for Os Keyes and one for Jevan Hutson, both from the University of Washington. Below them is another section for Meredith Durbin. A "ABSTRACT" section follows, containing a detailed paragraph about the ethical implications of algorithmic systems. At the bottom, there is a "CCS CONCEPTS" section with a single bullet point.

A Mulching Proposal

Analysing and Improving an Algorithmic System for Turning the Elderly into
High-Nutrient Slurry

Os Keyes
Department of Human Centered Design &
Engineering
University of Washington
Seattle, WA, USA
okeyes@uw.edu

Jevan Hutson
School of Law
University of Washington
Seattle, WA, USA
jevanh@uw.edu

Meredith Durbin
Department of Astronomy
University of Washington
Seattle, WA, USA
mdurbin@uw.edu

ABSTRACT

The ethical implications of algorithmic systems have been much discussed in both HCI and the broader community of those interested in technology design, development and policy. In this paper, we explore the application of one prominent ethical framework—Fairness, Accountability, and Transparency—to a proposed algorithm that resolves various societal issues around food security and population ageing. Using various standardised forms of algorithmic audit and evaluation, we drastically increase the algorithm's adherence to the FAT framework, resulting in a more ethical and beneficent system. We discuss how this might serve as a guide to other researchers or practitioners looking to ensure better ethical outcomes from algorithmic systems in their line of work.

CCS CONCEPTS

- Human-centered computing → Empirical studies in HCI; Social engineering (social sciences); • Computing methodologies → Object recognition; Machine learning algorithms; •

CCS2020-11-2020-CH-1000-1000

FAIRNESS

The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning



The Measure and Mismeasure of Fairness:
A Critical Review of Fair Machine Learning*

Sam Corbett-Davies Sharad Goel
Stanford University Stanford University

September 11, 2018

Abstract

The nascent field of fair machine learning aims to ensure that decisions guided by algorithms are equitable. Over the last several years, three formal definitions of fairness have gained prominence: (1) anti-classification, meaning that protected attributes—like race, gender, and their proxies—are not explicitly used to make decisions; (2) classification parity, meaning that common measures of predictive performance (e.g., false positive and false negative rates) are equal across groups defined by the protected attributes; and (3) calibration, meaning that conditional on risk estimates, outcomes are independent of protected attributes. Here we show that all three of these fairness definitions suffer from significant statistical limitations. Requiring anti-classification or classification parity can, perversely, harm the very groups they were designed to protect; and calibration, though generally desirable, provides little guarantee that decisions are equitable. In contrast to these formal fairness criteria, we argue that it is often preferable to treat similarly risky people similarly, based on the most statistically accurate estimates of risk that one can produce. Such a strategy, while not universally applicable, often aligns well with policy objectives; notably, this strategy will typically violate both anti-classification and classification parity. In practice, it requires significant effort to construct suitable risk estimates. One must carefully define and measure the targets of prediction to avoid retrenching biases in the data. But, importantly, one cannot generally address these difficulties by requiring that algorithms satisfy popular mathematical formalizations of fairness. By highlighting these challenges in the foundation of fair machine learning, we hope to help researchers and practitioners productively advance the area.

Keywords— Algorithms, anti-classification, bias, calibration, classification parity, decision analysis, measurement error

FAIRNESS

The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning

- literacy tests
- redlining
- proxies, like
 - zip code
 - last name

FAIRNESS

The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning

“...there are important cases where even protected group membership itself should be explicitly taken into account to make equitable decisions.”

FAIRNESS

To varying degrees, we are **mediators** of “fair” or
“equitable” or “just” outcomes.

FAIRNESS

To varying degrees, we are **mediators** of “fair” or “equitable” or “just” outcomes.

How can you identify **stakeholders** who might care how the system adjudicates issues?

What is your responsibility regarding a system **after** deployment?