

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

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



Facebook apologized on Friday for mislabeling and said it was looking into its recommendation feature to "prevent this from happening again." Jim Wilson/The New York Times

By Ryan Mac

Special Offer. Subscribe and enjoy unlimited articles with Basic Digital Access.

AI WILL UNDERSTAND MORE

→ labeling

WIRED BACKCHANNEL BUSINESS CULTURE GEAR IDEAS SCIENCE SECURITY SIGN IN SEARCH

TM SIMINITE 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



Establishing video connection...

AI WILL UNDERSTAND MORE

→ labeling

The screenshot shows a news article from MIT Technology Review. The header features the site's logo and navigation links for 'Sign in' and 'Subscribe'. Below the header, the word 'ARTIFICIAL INTELLIGENCE' is displayed in white capital letters. The main title of the article is 'Neural Network Learns to Identify Criminals by Their Faces', also in white capital letters. A short summary follows: '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'. In the bottom left corner of the main article area, there is a block of text: '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.' To the right of the main article, there is a sidebar titled 'POPULAR' which lists several other articles. At the very bottom right of the page, it says '10 Breakthrough Technologies 2022'.

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 believed that criminals were "throwbacks" more closely related to apes than law-abiding citizens. He was convinced he could identify them by ape-like features such as a sloping forehead, unusually sized ears and various asymmetries of the face and long arms. Indeed, he measured many subjects in an effort to prove his view although he did not analyze his data statistically.

POPULAR

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

Jessica Hamzelou

The secret police Inside the app Minnesota police used to collect data on journalists at protests

Sam RichardsTale Ryan-Mosley

Activists are targeting Russians with open-source "protestware"

Patrick Howell O'Neill

10 Breakthrough Technologies 2022

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 5, 2020. Sheldon Cooper/SOPA Images/LightRocket via Getty Images

- TikTok said it banned six accounts reported to it for posting content promoting eating habits that are likely to lead to health problems in its latest effort to crackdown on harmful content.
- The app is rife with dangerous material including 'pro-ana' or pro-anorexia and 'pro-mia' or pro-bulimia content which has plagued other social media networks such as Tumblr in the past.

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



Volunteers helped a driver who was stranded after being led by a GPS system down a snow-covered, two-lane dirt road as an alternative route to the closed Interstate 80 on Monday night. Washoe County Sheriff's Office

Special offer. Subscribe for \$1 a week.

AI WILL UNDERSTAND MORE

- labeling
- moderation
- navigation

The screenshot shows a computer monitor displaying a WIRED.com article. The article's title is "Why Zillow Couldn't Make Algorithmic House Pricing Work" by KIRK STRELZOW, published on NOV 11, 2021 at 7:00 AM. Below the title is a subtitle: "The real estate site went into the business of buying and selling homes. But the pandemic messed up its predictions." To the right of the text is a large, high-angle aerial photograph of a residential neighborhood with numerous houses and streets. At the bottom of the screen, there is a promotional banner for a WIRED magazine flash sale.

WIRED 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



FLASH SALE. Get WIRED for just \$29.99 \$5. Ending soon.

Subscribe now.

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 Mettsky¹, Laura Dabbish^{1,2}
¹Human-Computer Interaction Institute, "Heinz College
Carnegie Mellon University
(mklee, dkusbit, emettsky, 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 and data-driven management on human workers and work practices in the context of Uber and Lyft, new ride-sharing services. Our findings from a qualitative study describe how drivers responded when algorithms assigned their work, provided informational support, and evaluated their performance, and how drivers used online forums to socially make sense of the algorithmic features. Implications and future work are discussed.

Author Keywords
Algorithmic management; human-centered algorithms; management systems; CSCW; on-demand work; sharing economy; 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 functions and manage human devices through agent algorithms in practice algorithmic management. Algorithmic management allows companies to oversee myriads of workers in an optimized manner at a large scale, but it has also raised concerns about the impact of algorithmic management on workers. As a first step toward answering these questions, we interviewed 21 drivers with Uber and Lyft and triangulated their experiences by interviewing 12 passengers and conducting archival analysis of online driver forums and official company communication materials. The findings highlight opportunities and challenges in designing human-centered algorithmic work assignment, information, and evaluation as well as the importance of supporting social sensemaking around algorithmic systems. We use the findings to discuss how algorithms and data-driven management should be designed to create a better workplace with intelligent machines, offering implications for future work.

Our study makes the following two contributions to human-computer interaction (HCI). 1) we describe the inside and

but its impact on human workers and work practices has been largely unexplored. In recent years, the press and many scholars have brought attention to the importance of studying the sociotechnical aspects of algorithms [2, 10, 37], yet to our knowledge, there has been little empirical work in this area.

We explored the impact of algorithmic management in the context of new ride-sharing services Uber and Lyft. Algorithmic management is one of the core innovations that enables these services. Independent, distributed drivers with their own cars are assigned work by a central computer system within seconds or minutes, and the fare dynamically changes based on where passenger demand surges, all through the app on their mobile phones. Drivers' performance is evaluated by passengers' rating of their service quality and drivers' level of cooperation with algorithmic management. Algorithmic management allows a few hundred drivers in each city to generate hundreds and thousands of drivers on a global scale. Drivers have little direct contact with company representatives, but can interact with each other through online forums to gain social knowledge of the ride-share system. This setting allowed us to explore the practices that emerged when algorithms assigned work, optimized work behavior through performance评价, and evaluated performance. Do human workers compare with algorithmically-assigned work? How much are people motivated or demotivated by algorithmic optimization? How effective is algorithmic, data-driven evaluation and how do workers feel about it?

As a first step toward answering these questions, we interviewed 21 drivers with Uber and Lyft and triangulated their experiences by interviewing 12 passengers and conducting archival analysis of online driver forums and official company communication materials. The findings highlight opportunities and challenges in designing human-centered algorithmic work assignment, information, and evaluation as well as the importance of supporting social sensemaking around algorithmic systems. We use the findings to discuss how algorithms and data-driven management should be designed to create a better workplace with intelligent machines, offering implications for future work.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyright for copies made for educational, research, and technical purposes may be granted by the copyright holder or its agent. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permission from <http://www.acm.org>.

CHI 2013, April 19–23, 2013, Seoul, Republic of Korea
Copyright is held by the owner/author(s). Publication rights licensed to ACM.

AI WILL UNDERSTAND MORE

- labeling
- moderation
- navigation

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 present 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 *CCS Conference on Human Factors in Computing Systems (CCS '21)*, May 8–13, 2021, Yokohama, Japan. ACM, New York, NY, USA, 14 pages. <https://doi.org/10.1145/3411794.3451740>

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, 41, 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, 51]. 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 [10–32, 37, 50, 74].

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 a woefully misaligned from people's experiences of massive computational systems. Many of these algorithmic systems, especially ML systems, cause substantial harm in myriad domains, often surprising the designers of those systems.

Designers of sociotechnical systems have repeatedly built computational systems and models rendering decisions that exacerbate and reinforce historical prejudices, oppression, and marginalization. As designers of systems, our

Permissions to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyright for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from <http://www.acm.org/jrnlpermissions.html>.
© 2021 Copyright held by the owner/author(s). Publication rights licensed to ACM.
Manuscript submitted to ACM.

**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

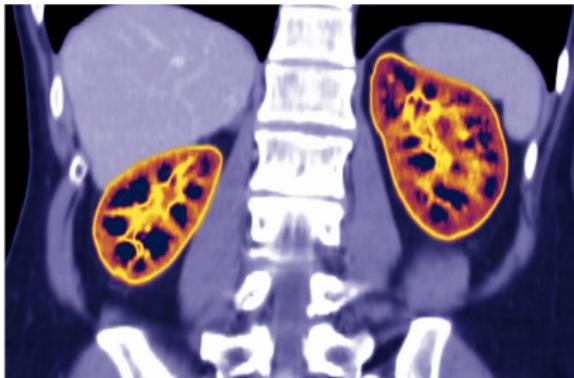
**AI WILL BE “FAIRER” OR
“MORE OBJECTIVE” THAN WE
CAN BE**

AI WILL BE FAIRER THAN WE CAN BE

→ medicine

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.



A score known as eGFR aims to reflect the seriousness of a patient's kidney disease. PHOTOGRAPH: JAMES CAVALLINI/SCIENCE SOURCE

AI WILL BE FAIRER THAN WE CAN BE

→ medicine

The screenshot shows a web browser window with the URL <https://doi.org/10.1038/s41591-020-01192-7>. The page is titled 'ARTICLES' and features the 'nature medicine' logo. The main title of the article is 'An algorithmic approach to reducing unexplained pain disparities in underserved populations'. The article is authored by Emma Pierson^{1,2}, David M. Cutler¹, Jure Leskovec^{3,4}, Sendhil Mullainathan^{3,5,6} and Ziad Obermeyer¹. The abstract discusses how underserved populations experience higher levels of pain, despite similar objective severity of disease like osteoarthritis. The study uses a deep learning approach to predict pain based on X-ray images, which significantly reduces racial disparities in pain prediction compared to standard radiographic measures. The article is dated November 10, 2020.

Underprivileged populations experience higher levels of pain. These disparities persist even after controlling for the objective severity of disease 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% (95% CI, 4.7–50%) of racial disparities in pain. 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,¹ disproportionately affects underserved populations, people of color, and those living in rural areas and in poverty, who often lack access to healthcare.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

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, some physical ailments in different populations, such as lower body differences related to lifeless social isolations or other factors⁴. 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 internal to the knee.

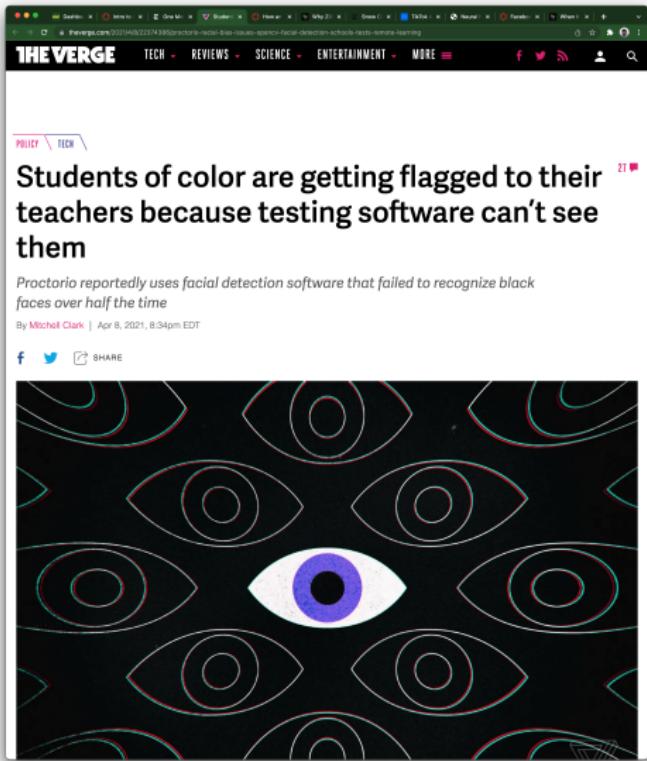
Research to date has indirectly implicated factors external to the knee. Methodologically, this is demonstrated by defining an objective measure of osteoarthritis severity based on knee X-rays and then measuring the extent of pain disparities remaining after adjusting for this measure. Typically, larger differences in pain remain even after adjustment.⁵ For example, even though Black patients have more severe osteoarthritis based on standard radiographic measures (Kellgren–Lawrence grade (KLG)), adjusting for KLG only slightly decreases Black–white pain disparities.⁶ These findings that pain disparities remain even when adjusting for radiographic osteoarthritis severity, however, depend heavily on how severity is measured. The relationship between radiographic severity and pain is debated. Many patients with mild or no disease as measured by radiographic severity suffer pain, and many patients with structural damage on X-ray or even magnetic resonance imaging (MRI) experience no or very little pain.^{7–10} Standard radiographic measures such as KLG, developed decades ago in white British populations, might miss physical causes of pain in people of color and those in rural areas, and therefore have less predictive value in a patient's pain is perceived by observers.¹¹ 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 misclassified as having external to the knee.

To this end, we propose a novel machine-learning approach to disentangle 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 172 patients in the United States who had no history at the time of development of knee osteoarthritis. As part of an NIH-funded study¹², bilateral fixed flexion knee radiographs were obtained and scored by radiologists on summary measures of radiographic severity (for example, KLG) and other measures of pain (for example, outcome index and space rating (ISR)). Patients also report a knee-specific questionnaire (Knee injury and Osteoarthritis Outcome Score (KOOS)), derived from a multi-item survey on pain experienced during various activities (for example, fully straightening the knee).

Summary scores of 4,172 participants, who generated 36,369 observations (one for each knee at each time point) are provided in Table 1. Black patients had substantially higher pain levels across knees and time points compared with non-Black patients (97% vs 75% mean Black–white pain difference). The median pain score of the Black–white (0.035 vs 0.1, a standard threshold for 'severe pain'), compared with 35% for patients overall (P for racial difference, <0.001). The median Black patient had worse pain than 75% of non-Black patients. Black patients had a pain score that was 10.6 KOOS points higher than that of non-Black patients

AI WILL BE FAIRER THAN WE CAN BE

→ medicine



The screenshot shows a news article from The Verge. The title is "Students of color are getting flagged to their teachers because testing software can't see them". The subtitle reads "Proctorio reportedly uses facial detection software that failed to recognize black faces over half the time". The author is Mitchell Clark, dated April 6, 2021, 8:34pm EDT. Below the text is a graphic of multiple stylized eyes on a dark background, with one eye highlighted in purple.

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

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

By Mitchell Clark | April 6, 2021, 8:34pm EDT

SHARE



AI WILL BE FAIRER THAN WE CAN BE

- medicine
- education

The Verge website screenshot showing a news article about AI in education.

THE VERGE TECH • REVIEWS • SCIENCE • ENTERTAINMENT • MORE

US & WORLD TECH ARTIFICIAL INTELLIGENCE

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 | @JonPorter | Aug 17, 2020, 12:16pm EDT

SHARE



Photo by Lucy North / MI News / NurPhoto via Getty Images

The UK has said that students in England and Wales will no longer receive exam results based on a controversial algorithm after accusations that the

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.

SUBSCRIBE

AI WILL BE FAIRER THAN WE CAN BE

- medicine
- education
- social

The screenshot shows a Twitter post from the 'Engineering' blog account (@ruchowdh). The post is titled 'Sharing learnings about our image cropping algorithm'. It features a profile picture of Rumman Chowdhury, Director, Twitter META. The post includes a link to 'Only on Twitter' and mentions @Twitter and #OnlyOnTwitter. The main text discusses how Twitter heard feedback in October 2020 about the image cropping algorithm not serving all people equitably and how they addressed it. It also mentions the analysis was a collaborative effort with Kyra Yee and Tao Tantipongpipat from the ML Ethics, Transparency, and Accountability team and Shubhangshu Mishra from the Content Understanding Research team.

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 collaborative effort together with [Kyra Yee](#) and [Tao Tantipongpipat](#) from our ML Ethics, Transparency, and Accountability (META) team and [Shubhangshu Mishra](#) from our Content Understanding Research team, which specializes in improving our ML models for various types of content in tweets. In our research, we tested our model for gender and race-based biases and considered whether our model aligned with our goal of enabling people to make their own choices on our platform.

AI WILL BE FAIRER THAN WE CAN BE

- medicine
- education
- social



Bernard Parker, left, was rated high risk; Dylan Augest was rated low risk. (Drew Angerer 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 21, 2016

ON A SPRING AFTERNOON IN 2014, Brisha Borden was running late to pick up her god-sister from school when she spotted an unlocked kid's blue Huffy bicycle and a silver Razor scooter. Borden and a friend grabbed the bike and scooter and tried to ride them down the street in the Fort Lauderdale suburb of Coral Springs.

AI WILL BE FAIRER THAN WE CAN BE

- medicine
- education
- social

The New York Times

One Month, 500,000 Face Scans:
How China Is Using A.I. to Profile
a Minority

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.

SenseFace

人脸布控实战平台

SenseFace Face Recognition Surveillance Platform

Special offer. Subscribe for \$1 a week.

WHAT IS FAIRNESS?

FAIRNESS

A Mulching Proposal

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 beneficial 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; •

CHI 2019, May 2019, Glasgow, Scotland, UK
2019, ACM ISBN 978-1-4503-5970-2/19/05... \$15.00
<https://doi.org/10.1145/3290608.3295423>

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 risk estimates, called ‘‘fairness constraints,’’ 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 use a different perspective, namely, how the root statistical approach creates a distribution of risk that one can produce. Such a strategy, while not usually 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 introducing biases in risk data. But importantly, one cannot generally address these difficulties by requiring that algorithms satisfy the three 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 field.

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

*We thank Alex Chouldechov, Alessandro Chouldechov, Avi Feller, Aziz Huq, Moritz Hardt, Daniel E. Ho, Shira Mitchell, Jan Odegard, Emma Pierson, and Ravi Shroff for their thoughtful comments. This paper synthesizes and expands upon material that we first presented in tutorials at the 19th Conference on Economics and Computation (EC 2018) and at the 35th International Conference on Machine Learning (ICML 2018). We thank the audiences at

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?