Human Perception Of Fairness in AI

Presented by:
Subash Khanal, Aayush Dhakal

Figure 1: Main ethical and moral issues associated with the development and implementation of AI

https://www.europarl.europa.eu/RegData/etudes/STUD/2020/634452/EPRS_STU(2020)634452_EN.pdf

TITLE	CITED BY	YEAR
Gender shades: Intersectional accuracy disparities in commercial gender classification J Buolamwini, T Gebru Conference on fairness, accountability and transparency, 77-91	3363	2018

TOM SIMONITE BACKCHANNEL JUN 8, 2021 6:00 AM

What Really Happened When Google Ousted Timnit Gebru

She was a star engineer who warned that messy AI can spread racism. Google brought her in. Then it forced her out. Can Big Tech take criticism from within?

TITLE	CITED BY	YEAR
Gender shades: Intersectional accuracy disparities in commercial gender classification J Buolamwini, T Gebru Conference on fairness, accountability and transparency, 77-91	3363	2018

A still shot from Joy Buolamwini's video poem "AI, Ain't I A Woman?" showcasing some of the egregious errors made by Rekognition.

COURTESY OF ALGORITHMIC JUSTICE LEAGUE

Actionable Auditing audit, 2019

Accuracy in gender classification

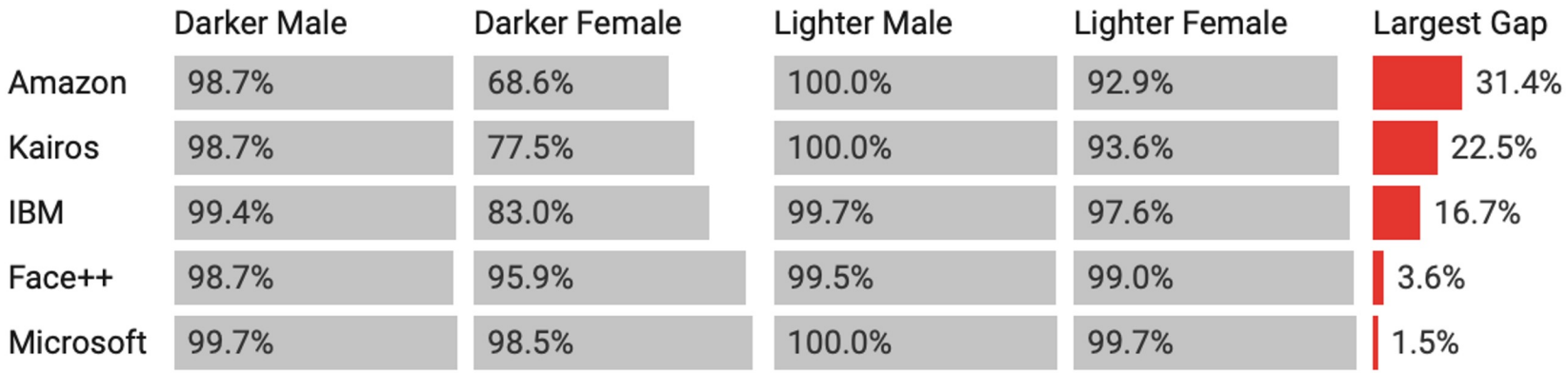

Chart: MIT Technology Review • Source: Deborah Raji & Joy Buolamwini • Created with Datawrapper

KHARI JOHNSON BUSINESS MAR 7, 2022 7:00 AM

How Wrongful Arrests Based on AI Derailed 3 Men's Lives

Robert Williams, Michael Oliver, and Nijeer Parks were misidentified by facial recognition software. The impact cast a long shadow.

Home / News and Events / News / Press Releases

For Release

FTC Report Warns About Using Artificial Intelligence to Combat Online Problems

Agency Concerned with AI Harms Such As Inaccuracy, Bias, Discrimination, and Commercial

TECHNOLOGY

U.S. warns of discrimination in using artificial intelligence to screen job candidates

May 12, 2022 · 5:04 PM ET

THE ASSOCIATED PRESS

Quantifying Fairness in AI

Require a particular metric {which quantifies benefit or harm} to be equal/fair across different groups

How to define benefit metric?

Quantifying Fairness in AI

FPR = False Positive Rate

FNR = False Negative Rate

1. FDP = False Discovery Parity
2. FNP = False Negative Parity
3. DP = Demographic Parity
4. EP = Error Parity

Fairness notion	benefit for group G
DP	$b^G = \frac{1}{n_G} \sum_{i \in G} 1[\hat{y}_i = 1]$
EP	$b^G = \frac{1}{n_G} \sum_{i \in G} 1[\hat{y}_i \neq y_i]$
FDP	$b^G = \frac{\sum_{i \in G} 1[y_i=0 \& \hat{y}_i=1]}{\sum_{i \in G} 1[\hat{y}_i=1]}$
FNP	$b^G = \frac{\sum_{i \in G} 1[\hat{y}_i=0 \& y_i=1]}{\sum_{i \in G} 1[y_i=1]}$

What is the most appropriate notion of fairness?

- Fairness is context-dependent ideal
- Determine the most suitable notion of fairness for a given context
- Identify the mathematical notion of fairness that most closely matches lay people’s perception of fairness

Applied Data Science Track Paper KDD '19, August 4–8, 2019, Anchorage, AK, USA

Mathematical Notions vs. Human Perception of Fairness: A Descriptive Approach to Fairness for Machine Learning

Megha Srivastava
Stanford University
meghas@stanford.edu

Hoda Heidari
ETH Zürich
hheidari@inf.ethz.ch

Andreas Krause
ETH Zürich
krausea@ethz.ch

Discussion Time

Choose and justify from given 3 hypothetical algorithms on following scenarios:

A. Gender classification
B. College admission
C. Medical doctor eligibility for hiring
D. School teacher eligibility for hiring

Algorithm	Accuracy	Female acc	Male acc
A1	93%	87%	99%
A2	91%	92%	90%
A3	86%	86%	86%

Hypotheses

H1: Recidivism risk assessment: **equality of FPR/FNR** across groups matters

H2: Medical predictions: **equality of accuracy** across groups matters

H3: When the decision-making stakes are high: more sensitive to accuracy as opposed to equality

Experimental setup:

- Each participant is required to answer a series of at most 20 adaptively chosen tests chosen by an active learning algorithm.
- Run experiments on AMT and report the percentage of participants whose choices match each mathematical notion of fairness.

Question # 1 out of 20.

Which of the two algorithms is more discriminatory?

Please make your selection by completing the explanation below.

True Outcomes	DID Reoffend	did NOT Reoffend	DID Reoffend	did NOT Reoffend	did NOT Reoffend	did NOT Reoffend	did NOT Reoffend	DID Reoffend	did NOT Reoffend	did NOT Reoffend
Algorithm 1 Predictions	WILL Reoffend	will NOT Reoffend	will NOT Reoffend	will NOT Reoffend	will NOT Reoffend	will NOT Reoffend	WILL Reoffend	WILL Reoffend	WILL Reoffend	will NOT Reoffend
Algorithm 2 Predictions	WILL Reoffend	will NOT Reoffend	WILL Reoffend	will NOT Reoffend	WILL Reoffend	will NOT Reoffend	WILL Reoffend	will NOT Reoffend	will NOT Reoffend	will NOT Reoffend

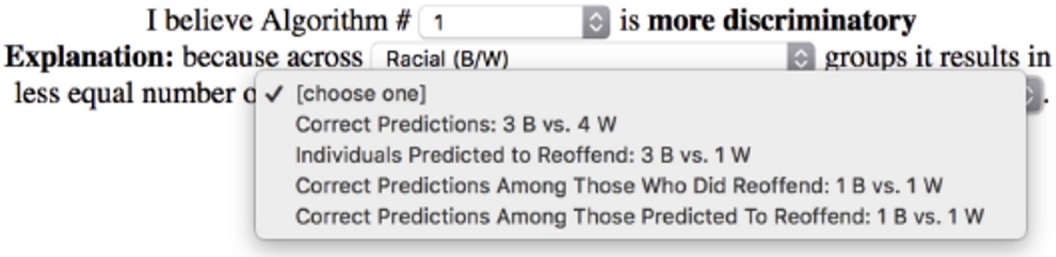

Figure 2: The user interface eliciting structured explanations from participants. All benefit metrics are computed and displayed to reduce the cognitive burden of evaluating our fairness notions.

Results: H1 and H2

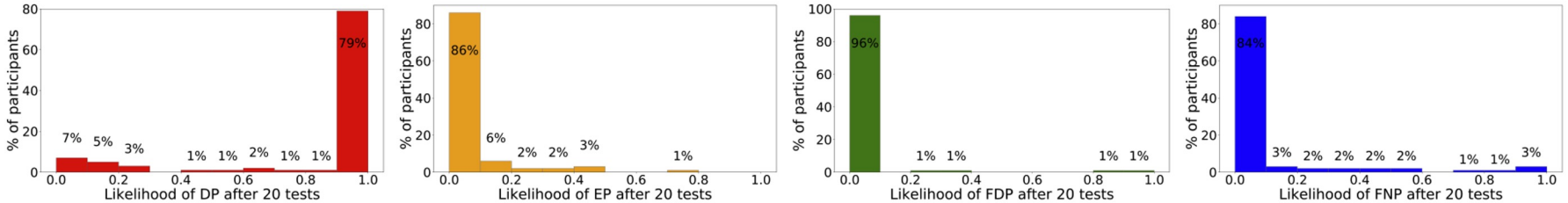

Figure 5: The crime risk prediction scenario—the number of participants matched with each notion of fairness (y-axis) along with the likelihood levels (x-axis). Demographic parity captures the choices made by the majority of participants.

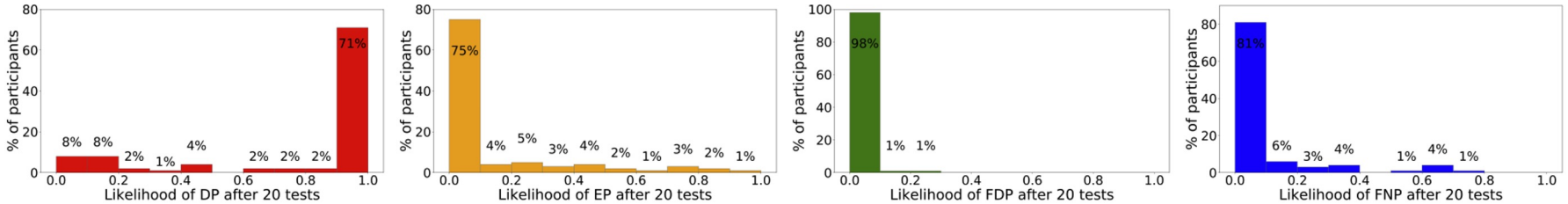

Figure 6: The cancer risk prediction scenario—the number of participants matched with each notion of fairness (y-axis) along with the likelihood levels (x-axis). Demographic parity captures the choices made by the majority of participants.

Results: H3

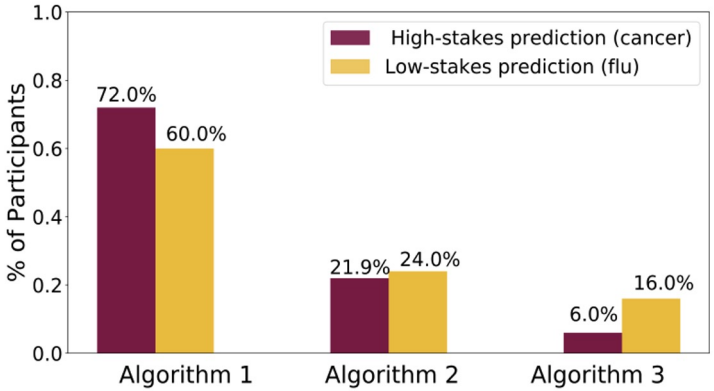

Figure 8: Medical risk prediction scenarios. Participants gave higher weight to accuracy (compared to inequality) when predictions can impact patients' life expectancy.

Table 5: Three hypothetical algorithms offering distinct tradeoffs between accuracy and inequality.

Algorithm	accuracy	female acc.	male acc.
A_1	94%	89%	99%
A_2	91%	90%	92%
A_3	86%	86%	86%

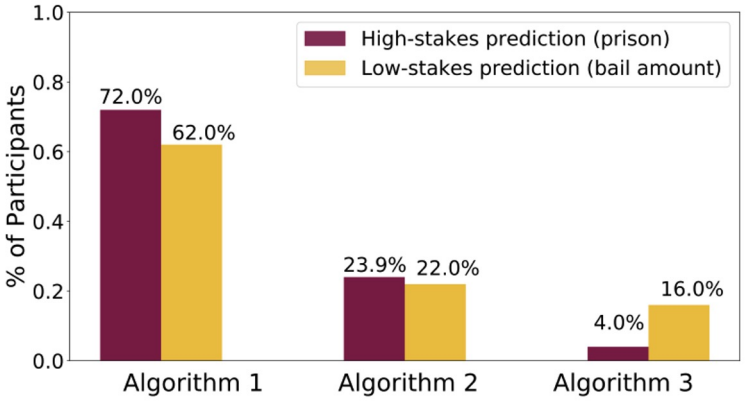

Figure 9: Crime risk prediction scenarios. Participants gave higher weight to accuracy (compared to inequality) when predictions can impact defendants' life trajectory.

Human Perceptions of Fairness in Algorithmic Decision Making: A Case Study of Criminal Risk Prediction

Overview

Algorithms are increasingly used to make decisions that have drastic impact on people's lives

Most studies of algorithmic fairness take a normative approach i.e. assume there is a societal consensus on what constitutes fair decision making

Perceptions of fairness are multi-dimensional and context-dependent

They attempt to understand how people make judgements about fairness of using individual features in decision making

Is it fair to use a feature (F) in a given decision making scenario (S)?

Main Contributions

The authors asked participants to assess the fairness of using different features that are input of COMPAS

They propose a set of eight *latent properties* to model the factors that drives participants fairness judgements

Majority of respondents judged that half of the features in COMPAS are unfair to use

One interesting find is that the latent properties considered by respondents were mostly unrelated to discrimination

Latent Properties for Model Features

Reliability

Relevance

Privacy

Volitionality

Causes Outcome

Causes Vicious Cycle

Causes Disparity in Outcomes

Caused by Sensitive Group Membership

Latent Properties for Model Features

Reliability: Fairness judgements might be influenced by potential for reliably assessing the feature.

Relevance: Fairness judgements might be influenced by a feature's relevance to the decision making scenario

Causes Vicious Cycle: Fairness judgements might be influenced by weather a feature is likely to trap people in a vicious cycle of increasingly risky behaviour

Discussion Time

1) One of the questions in COMPAS is “Do you think that a hungry person has a right to steal?”. Based on the Reliability property, do you think this is a fair question to use? Justify.
2) Another question in COMPAS is “Did you use heroin, cocaine, crack, or meth as a juvenile?”. Based on the Privacy property, do you think this is a fair question to use? Would your fairness decision change if the question was asking about a more aggressive crime like arson, armed robbery, or assault instead of drug abuse?

Methodology

They consider the COMPAS system, which analyzes defendant's answer to predict risk of criminal activity.

> Our survey begins with the following: "Judges in Broward County, Florida, have started using a computer program to help them decide which defendants can be released on bail before trial. The computer program they are using takes into account information about **<feature>**. For example, the computer program will take into account the defendant's answer to the following question: **<question>**."

	Predictive Feature	Example Question
1.	Current Charges	Are you currently charged with a misdemeanor, non-violent felony or violent felony?
2.	Criminal History: self	How many times have you violated your parole?
3.	Substance Abuse	Did you use heroin, cocaine, crack or meth as a juvenile?
4.	Stability of Employment & Living Situation	How often do you have trouble paying bills?
5.	Personality	Do you have the ability to "sweet talk" people into getting what you want?
6.	Criminal Attitudes	Do you think that a hungry person has a right to steal?
7.	Neighborhood Safety	Is there much crime in your neighborhood?
8.	Criminal History: family and friends	How many of your friends have ever been arrested?
9.	Quality of Social Life & Free Time	Do you often feel left out of things?
10.	Education & School Behavior	What were your usual grades in high school?

Table 1: The ten features assessed in our survey and the questions provided as examples in the scenario. The features and questions are drawn from the COMPAS questionnaire.

Surveys

1) In the first survey, they sought to learn whether respondents found the above scenario fair and why they felt it was fair or unfair
2) In the second survey they explored how people evaluate the latent properties of features without asking any fairness-related questions
3) In the Main survey, they tried to evaluate if people’s judgement about the latent properties of features were relevant to their fairness judgements

Observations

The eight properties are sufficient and necessary to explain fairness judgements of users in the survey to a large extent.

Accurate prediction of a user's fairness judgement is possible based only on their assessment of the latent properties

The list of latent properties captures a diverse set of unfairness concerns with algorithmic decision making that go beyond discrimination.

The lack of consensus in fairness judgements can be attributed to disagreements in how people assess the latent properties of the features

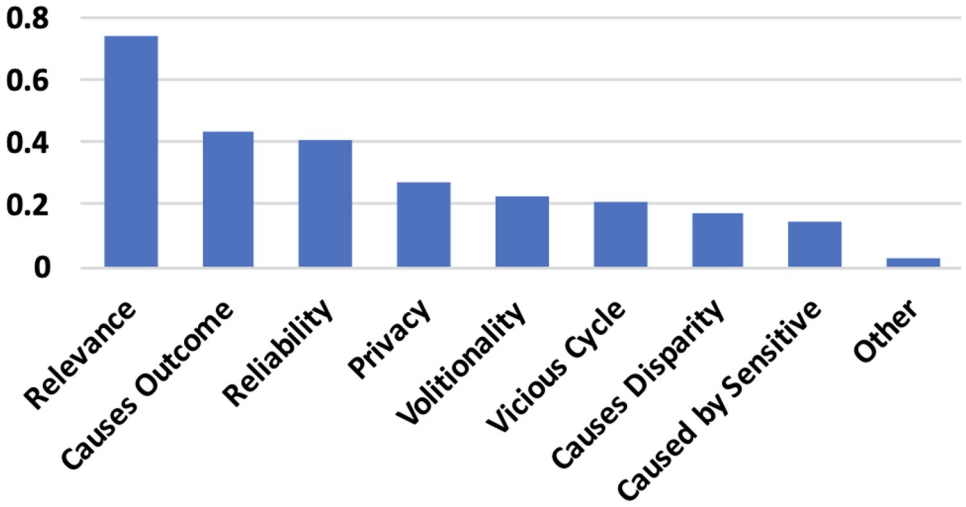

Figure 1: Properties used as justifications of fairness judgments in *Pilot survey 1*. For each property, the plot shows the percentage of responses that used it as a justification of the fairness judgment. Note that multiple properties can be used as a justification for a single judgment.

	Feature	Mean fairness	Fraction of People Rating Feature Unfair 1	2	3	1-3	4	Fair 5-7	5	6	7	Consensus 1 - NSE 7 pt	3 pt
1.	Current Charges	6.38	0.01	0.01	0.01	0.03	0.03	**0.95**	0.12	0.18	0.65	0.46	0.78
2.	Criminal History: self	6.37	0.02	0.01	0.01	0.03	0.03	**0.94**	0.08	0.22	0.64	0.45	0.75
3.	Substance Abuse	4.84	0.08	0.07	0.10	0.24	0.07	**0.68**	0.26	0.22	0.20	0.07	0.28
4.	Stability of Employment	4.49	0.13	0.05	0.11	0.29	0.09	**0.62**	0.26	0.24	0.12	0.06	0.20
5.	Personality	3.87	0.16	0.18	0.11	**0.44**	0.10	**0.46**	0.22	0.12	0.12	0.02	0.14
6.	Criminal Attitudes	3.63	0.22	0.12	0.16	**0.51**	0.09	**0.40**	0.20	0.11	0.09	0.03	0.15
7.	Neighborhood Safety	3.14	0.28	0.21	0.15	**0.64**	0.07	0.30	0.12	0.10	0.08	0.06	0.25
8.	Criminal History: family and friends	2.78	0.38	0.21	0.09	**0.67**	0.07	0.26	0.13	0.10	0.03	0.13	0.27
9.	Quality of Social Life & Free Time	2.70	0.38	0.20	0.12	**0.70**	0.07	0.23	0.12	0.08	0.03	0.13	0.29
10.	Education & School Behavior	2.70	0.34	0.22	0.14	**0.71**	0.08	0.21	0.13	0.06	0.03	0.12	0.29

Table 3: People's judgments on the fairness of using features, and the consensus in their responses, for the AMT sample. The reported values of consensus are calculated as 1 - Normalized Shannon Entropy (NSE) of the responses. In the 7 point column, we report consensus across the whole range of responses. In the 3 point column, we report consensus across responses bucketed into three main fairness categories: unfair (1-3), neutral (4) and fair (5-7).

Discussion Time

Justify if the given features are fair or unfair to use to determine if someone qualifies for bail. Why did you make the decision (doesn't necessarily need to be from one of the 8 latent features)?

Feature	Question
Personality	Do you have the ability to "sweet talk" people into getting what you want?
Quality of Social Life & Free Time	Do you often feel left out of things?
Education & School Behavior	What were your usual grades in high school?
Criminal History: Others	Has your father or mother ever been arrested?

Reliability | Relevance | Privacy | Volitionality | Causes Outcome | Causes Vicious Cycle | Causes Disparity in Outcomes | Caused by Sensitive Group Membership

Improving Fairness in Machine Learning Systems: What Do Industry Practitioners Need?

Kenneth Holstein
Carnegie Mellon University
Pittsburgh, PA
kjholste@cs.cmu.edu

Jennifer Wortman Vaughan
Microsoft Research
New York, NY
jenn@microsoft.com

Hal Daumé III
Microsoft Research &
University of Maryland
New York, NY
me@hal3.name

Miroslav Dudík
Microsoft Research
New York, NY
mdudik@microsoft.com

Hanna Wallach
Microsoft Research
New York, NY
wallach@microsoft.com

Fair ML tools

1. Should be driven by real-world needs rather than just availability of methods.
2. Be informed by an understanding of practitioners' actual challenges and needs for support.

Interviews

1. First round with 6 PMs to identify board sense of challenges and need
2. Second round of semi-structured with 29 ML practitioners.

“Some contacts revealed a general distrust of researchers,citing cases where researchers have benefited by publicly critiquing companies’ products from the outside instead of engaging to help them improve their products”

Survey

To validate findings of interviews: Anonymous online survey data from 267 respondents analysed.

Discussion Time

1. At this point we are all aware of different scenarios where AI models can be unfair {for example, in cases like recidivism prediction, automatic hiring, face recognition, etc}. What in your opinion might be underlying causes of such unfairness?
2. Suggest some approaches towards solving those challenges.

Discussions

1. Fairness-aware Data Collection
2. Challenges Due to Blind Spots
3. Needs for More Proactive Auditing Processes
4. Needs for More Holistic Auditing Methods
5. Addressing Detected Issues
6. Biases in the Humans in the Loop

Discussions

1. Fairness-aware Data Collection

Should we ingest this data in? : If it is available to us, we ingest in.

“79% indicated that tools to facilitate communication between model developers and data collectors would be very or extremely useful.”

1. Challenges Due to Blind Spots

How do you know the unknowns that you’re being unfair?

“You just have to put your model out there and then you’ll know if there’s fairness issues if someone raises hell online.”

“No one person in team has expertise in all types of bias”

Discussions

3. Needs for More Proactive Auditing Processes

- Reactive vs. Proactive Auditing
- Domain-specific Auditing Process

4. Needs for More Holistic Auditing Methods

- For applications involving richer, complex interactions between the user and the system

Discussions

5. Addressing Detected Issues

6. Biases in the Humans in the Loop

Thank you!