- Eric Loomis pled guilty to driving a stolen vehicle & fleeing police
- Before sentencing, the court looked at an ML assessment (COMPAS) of his likelihood to reoffend. Loomis was rated a high risk
- COMPAS was not trained for sentencing, but for resource allocation after release
- COMPAS is biased in favor of whites. Loomis is white
- The court sentenced him to 8 years in prison, in part because of his high risk rating

Fair?

- Houston SD purchased an ML tool (EVAAS) to evaluate teachers' effectiveness. The system is proprietary
- Houston fired 25% of teachers with ineffective scores, based solely on their scores
- There is no process for teachers to inspect, challenge, or correct their data or scores
- Known errors cannot be corrected because of how the model normalizes: fixing one teacher's data has the potential to change all other teachers' scores

Fair?

- MSU uses cookies, ML, and third party datasets to score possible students who visit their site
- The scores favor those:
  - likely to accept an admissions offer to MSU
  - applying to majors the school is already strong in
  - wealthy out-of-state students
- The scores are used to determine who to focus recruitment efforts on, not who to accept

Fair?

What's Fair?

Mariah A. Knowles

Tue, Dec 6, 2022

When we say something is fair...

Fairness is a conclusion

Fair amounts to Good, Justified, Okay, Fine, Morally Acceptable, ...

Fair amounts to 👍

Fair is a thick evaluation, colloquially

Fair amounts to 👍 + some set of morally relevant criteria

Problem 1: Fair covers a range of moreorless related concepts, some competing

Problem 2: It's easy to talk past each other

1. Technology creates risks
2. We have responsibilities w.r.t. those risks
3. One type of risk is:
   People will be subject to an environment
   that runs afoul
   of fair terms of social cooperation

Some Fair Terms

- Malice
- Respect for Individuals
- Acceptable Social Arrangements
- Deception
- Acceptable Judgment

We ought not hold malice towards other groups

MACHINE BIAS

Facebook Enabled Advertisers to Reach 'Jew Haters'

After being contacted by ProPublica, Facebook removed several anti-Semitic ad categories and promised to improve monitoring.

by Julia Angwin, Madeleine Varner and Ariana Tobin, Sept. 14, 2017, 4 p.m. EDT

We ought to respect others as individuals

Power and Freedom ought to be distributed justly

&

Harms and Goods ought to be distributed "equally"

Minority Neighborhoods Pay Higher Car Insurance Premiums Than White Areas With the Same Risk

Our analysis of premiums and payouts in California, Illinois, Texas and Missouri shows that some major insurers charge minority neighborhoods as much as 30 percent more than other areas with similar accident costs.

by Julia Angwin, Jeff Larson, Lauren Kirchner and Surya Mattu, ProPublica
April 5, 2017

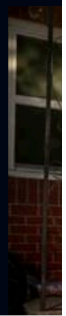
This story was co-published with Consumer Reports.

OTIS NASH WORKS SIX DAYS A WEEK AT TWO JOBS, as a security guard and a pest control technician, but still struggles to make the $190.69 monthly Geico car insurance payment for his 2012 Honda Civic LX.

"I'm on the edge of homelessness," said Nash, a 26-year-old Chicagoan who supports his wife and 7-year-old daughter. But "without a car, I can't get to work, and then I can't pay

Pasco’s sheriff created a futuristic program to stop crime before it happens.

It monitors and harasses families across the county.



We ought not deceive others

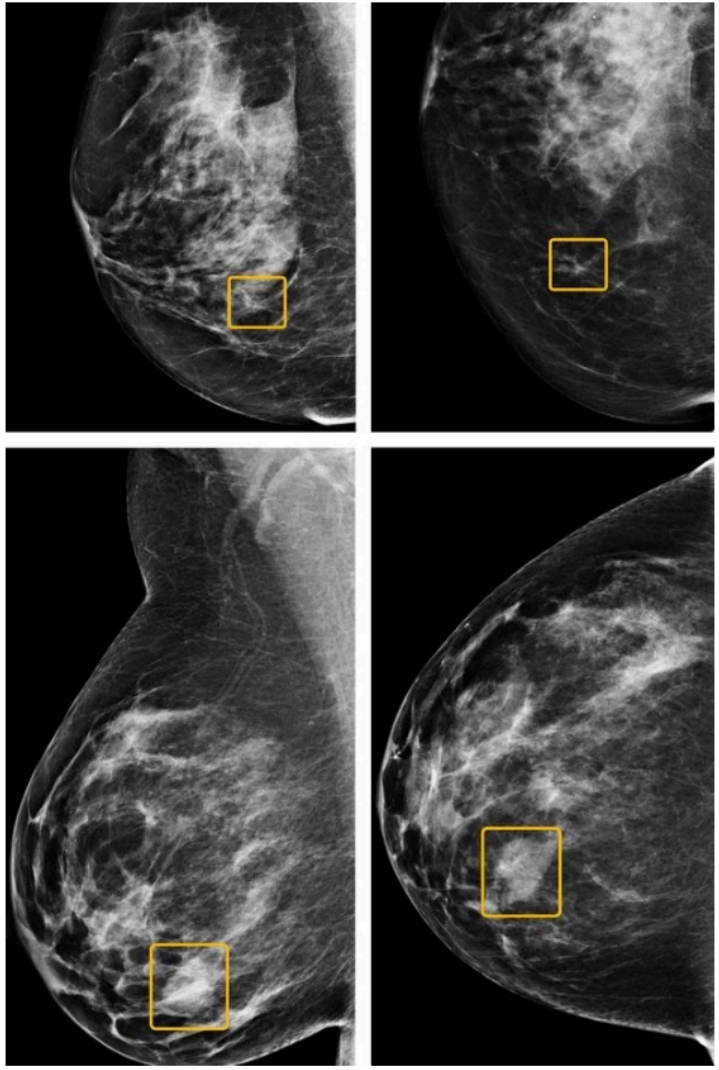
The algorithm is innocent

Google and Facebook deflect responsibility onto algorithms, as if they don't control their own code.

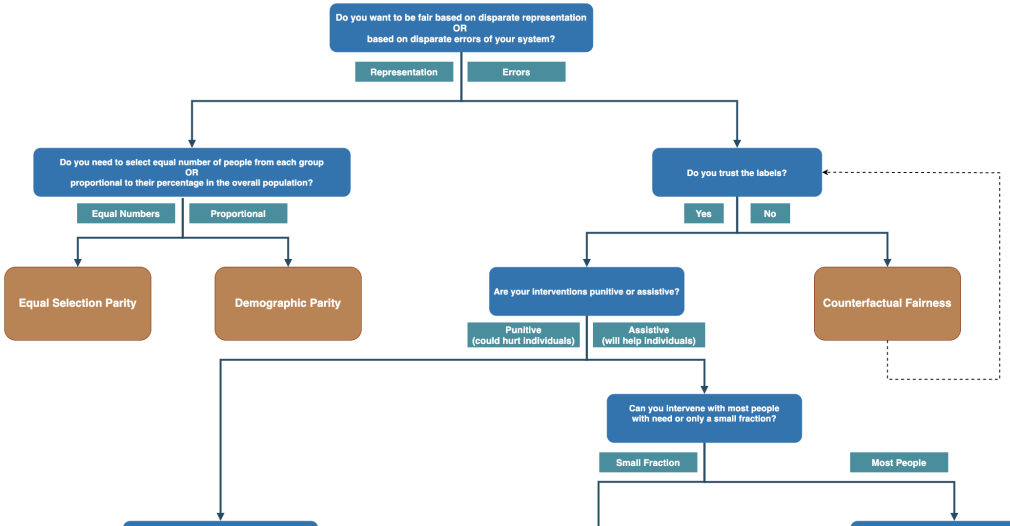
We ought not judge others on things outside their control

&

We ought not judge others on things that bear no relevance to the outcome



FAIRNESS TREE



Motivating Idea:

What are your chances of being wrongly denied bail just given your race?

Among people denied bail, what are the chances you're innocent given your race?

*Among people who **should** be granted bail, what are the chances you were denied bail given your race?*

If we can only provide assistance to a small fraction of people with need, attempt to ensure it is distributed in a representative way

What are your chances of being wrongly left out of assistance given your gender?

Among people who don't receive assistance, what are the chances you had need given your gender?

Among people with need, what are your chances you don't receive assistance given your gender?

Probabilistic Notion:

$P\left(\hat{Y}=1, Y=0 \mid G\right)$

$P\left(Y=0 \mid G, \hat{Y}=1\right)$

$P\left(\hat{Y}=1 \mid G, Y=0\right)$

$P\left(\hat{Y}=1 \mid G, Y=1\right)$

$P\left(\hat{Y}=0, Y=1 \mid G\right)$

$P\left(Y=1 \mid G, \hat{Y}=0\right)$

$P\left(\hat{Y}=0 \mid G, Y=1\right)$

* Note: Focusing on recall in this case is equivalent to focusing on FNR parity, but may have nicer mathematical properties, such as meaningful ratios. In such cases, you may also want to reconsider the definition of your target variable to ask whether the problem can be redefined to focus on cases with most severe need.

Model Cards for Model Reporting

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, Timnit Gebru
{mmitchellai,simonewu,andrewzaldivar,parkerbarnes,lucyvasserman,benhutch,espitzer,tgebru}@google.com
deborah.raji@mail.utoronto.ca

ABSTRACT

Trained machine learning models are increasingly used to perform high-impact tasks in areas such as law enforcement, medicine, education, and employment. In order to clarify the intended use cases of machine learning models and minimize their usage in contexts for which they are not well suited, we recommend that released models be accompanied by documentation detailing their performance characteristics. In this paper, we propose a framework that we call model cards, to encourage such transparent model reporting. Model cards are short documents accompanying trained machine

KEYWORDS

datasheets, model cards, documentation, disaggregated evaluation, fairness evaluation, ML model evaluation, ethical considerations

ACM Reference Format:
Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, Timnit Gebru. 2019. Model Cards for Model Reporting. In *FAT* '19: Conference on Fairness, Accountability, and Transparency, January 29–31, 2019, Atlanta, GA, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3287560.3287596

Intended Use

Primary intended use

1. Assist NLP researchers in improving detector models for this or other machine-generated models.
2. Provide social media platform moderators and consumers the probability that a selection of text is generated by CTRL, with an aim to detect machine-generated:
    - fake news;
    - texts that manipulate political opinions; and/or,
    - hate speech.

Primary intended users

- NLP researchers (can be beneficial to further develop detector models)
- Social media platform moderators
- Civil society actors with an interest in identifying and countering mis- and disinformation

Out-of-scope use cases

- The CTRL detector should not be used for the adversarial training of fake news generators.
- This software should not be used to promote or profit from:
    - violence, hate, and division;
    - environmental destruction;
    - abuse of human rights; or
    - the destruction of people's physical and mental health.
- The CTRL detector should not be used to automate a production-level system - this is currently in a state of applied research intended for the audiences noted above.

HmntyCntrd: Interactive, cohort-based online course and community for UX professionals who want to learn how to design and advocate for equitable and inclusive user experiences. Created by UX Researcher and Humanity in Tech Advocate Vivianne Castillo.

Responsible design practices to try:

Design Ethically Tooklit: A toolkit for design strategists and product designers with several 30-minute to 1-hour small group exercises to help teams evaluate ethical implications of product ideas, think about consequences of unintended user behaviors, and create checklists for ethical issues to monitor after shipping product. Created by Kat Zhou, Product Designer at Spotify.

Judgment Call: Team-based game for cultivating stakeholder empathy through scenario-imagining. Game participants write product reviews from the perspective of a particular stakeholder, describing what kind of impact and harms the technology could produce from their point of view. Created by Microsoft's Office on Responsible AI.

Harms Modeling: Framework for product teams, grounded in four core pillars that examine how people's lives can be negatively impacted by technology: injuries, denial of consequential services, infringement

ty

- Decision Tree
- Model Cards
- Design Practices
- Loomis & EVAAS
- MSU
- Facebook
- Car Insurance
- Pasco
- Algorithm is Innocent
- Breast Cancer