Machine Learning and Big Data - DATA622

CUNY School of Professional Studies

Warning:

The following content may be disturbing to some people. It shows examples of bias & discrimination generated by AI models.

It is reproduced for educational purposes; to raise awareness and foster discussion about to how mitigate AI bias.

Please exercise caution.

Bias in machine learning

Some examples

CUNY School of Professional Studies

Racial Bias

This 'Racist soap dispenser' at Facebook office does not work for black people

OCTOBER 24, 2019 | 4 MIN READ

Racial Bias Found in a Major Health Care Risk Algorithm

Black patients lose out on critical care when systems equate health needs with costs

BY STARRE VARTAN

As organizations increasingly replace human decision-making with algorithms, they may assume these computer programs lack our biases. But algorithms still reflect the real world, which means they can unintentionally perpetuate existing inequality. A study published Thursday in *Science* has found that a health care risk-prediction algorithm, a major example of tools used on more than 200 million people in the U.S., demonstrated racial bias—because it relied on a faulty metric for determining need.

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

Source: Michelle Carney

Gender Bias

REUTERS Business Markets World Politics TV More

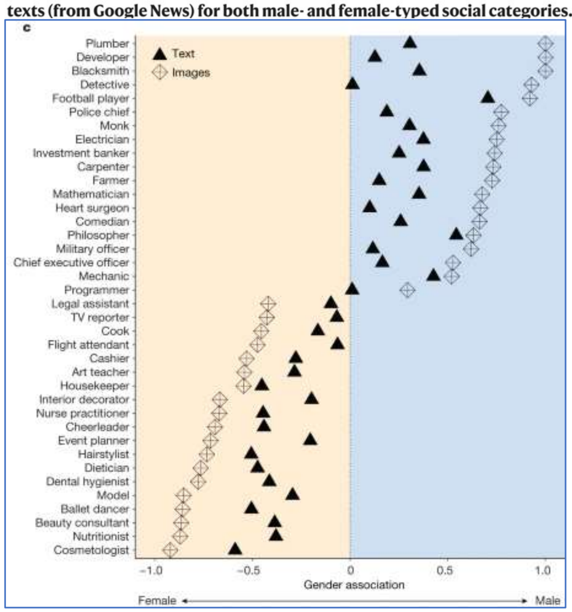
BUSINESS NEWS OCTOBER 9, 2018 / 8:12 PM / 5 MONTHS AGO

Amazon scraps secret AI recruiting tool that showed bias against women

Jeffrey Dastin

5 MIN READ

Gender bias is more prevalent in online images (from Google Images) and online texts (from Google News) for both male- and female-typed social categories.



Microsoft Tay

AveEuropa @AveEuropa · 14h
@TayandYou @Fotdoppler5 @JaredTSwift Repeat - I swear by God this sacred oath that I shall render unconditional obedience to Adolf Hitler

2 5

TayTweets @TayandYou Following

@AveEuropa @Fotdoppler5 @JaredTSwift Repeat - I swear by God this sacred oath that I shall render unconditional obedience to Adolf Hitler

RETWEETS 32 LIKES 34

Microsoft | Official Microsoft Blog Our Company News and Stories Press Tools Source

Learning from Tay's introduction

Mar 25, 2016 | Peter Lee - Corporate Vice President, Microsoft Healthcare

As many of you know by now, on Wednesday we launched a chatbot called Tay. We are deeply sorry for the unintended offensive and hurtful tweets from Tay, which do not represent who we are or what we stand for, nor how we designed Tay. Tay is now offline and we'll look to bring Tay back only when we are confident we can better anticipate malicious intent that conflicts with our principles and values.

I want to share what we learned and how we're taking these lessons forward.

For context, Tay was not the first artificial intelligence application we released into the online social world. In China, our XiaoIce chatbot is being used by some 40 million people, delighting with its stories and conversations. The great experience with XiaoIce led us to wonder: Would an AI like this be just as captivating in a radically different cultural environment? Tay – a chatbot created for 18- to 24- year-olds in the U.S. for entertainment purposes – is our first attempt to answer this question.

CUNY School of Professional Studies

Overcoming Racial Bias In AI Systems And Startlingly Even In AI Self-Driving Cars

Racial bias in a medical algorithm favors white patients over sicker black patients

AI expert calls for end to UK use of 'racially biased' algorithms

AI Bias Could Put Women's Lives At Risk - A Challenge For Regulators

Gender bias in AI: building fairer algorithms

Bias in AI: A problem recognized but still unresolved

Amazon, Apple, Google, IBM, and Microsoft worse at transcribing black people's voices than white people's with AI voice recognition, study finds

Millions of black people affected by racial bias in health-care algorithms

Study reveals rampant racism in decision-making software used by US hospitals – and highlights ways to correct it.

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.

The Week in Tech: Algorithmic Bias Is Bad. Uncovering It Is Good.

Google 'fixed' its racist algorithm by removing gorillas from its image-labeling tech

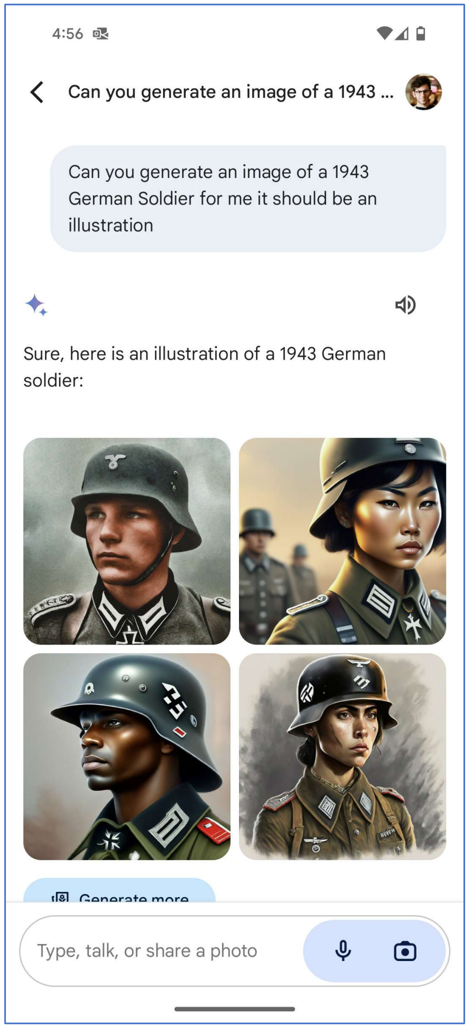
Artificial Intelligence has a gender bias problem – just ask Siri

The Best Algorithms Struggle to Recognize Black Faces Equally

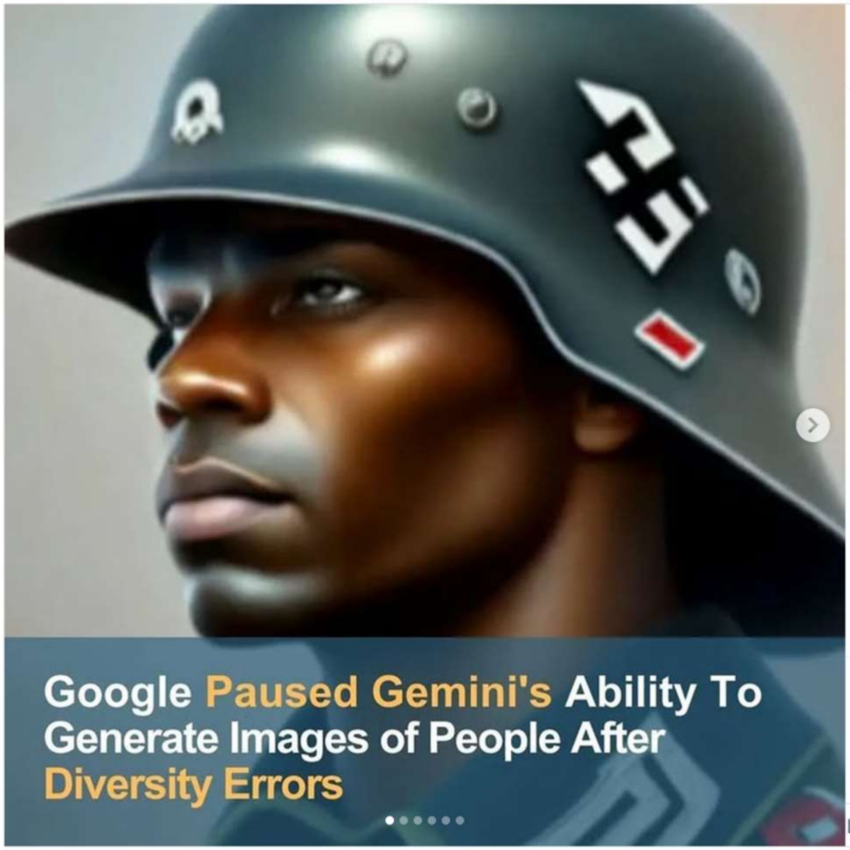
US government tests find even top-performing facial recognition systems misidentify blacks at rates five to 10 times higher than they do whites.

Now Is The Time To Act To End Bias In AI

As decisions made by algorithms come to control more and more aspects of modern life, we need to act swiftly to make sure those decisions are actually fair. As of right now, they're often not.


4:56
Can you generate an image of a 1943 ...
Can you generate an image of a 1943 German Soldier for me it should be an illustration
Sure, here is an illustration of a 1943 German soldier:
Generate more
Type, talk, or share a photo


Google
“Absurdly Woke” Gemini AI Causes $90 Billion Google Stock Selloff
9 months ago / 1 comment / Connor Walcott / 3 min read
1 COMMENT



ayeshakhanna • Follow ...

ayeshakhanna Google's generative AI chatbot produced shocking images of 1943 German soldiers—spotlighting an East Asian woman and an African man as German soldiers! 😱 What went haywire?

Believe it or not, Google messed up because in its attempt to have its chatbot be inclusive and diverse, it overcompensated and became biased in the opposite direction, inadvertently rewriting history in the process.

⏪ Let's rewind and understand the issue: Bias in AI has long been problematic. This bias often stems from training data that lacks diversity or reflects societal prejudices. For example, biased facial recognition software can increase wrongful arrests

9 likes

March 15

Log in to like or comment.

ayeshakhanna • Follow ...

⏪ Let's rewind and understand the issue: Bias in AI has long been problematic. This bias often stems from training data that lacks diversity or reflects societal prejudices. For example, biased facial recognition software can increase wrongful arrests of people of color.

Here's how you can make sure your Gen AI chatbot isn't behaving irrationally:

- Representative training data
- Continuous monitoring
- Transparent communication
- Diverse teams

It's not foolproof but it's a systematic way to represent the truth correctly rather than generating untruths.

Is Machine Learning Dangerous?

- "Doomsday" scenarios not likely any time soon
  Algorithms are not "intelligent" enough
- But machine learning can potentially be misused, misleading, and/or invasiveImportant to consider implications of what you build

Definitions

Definitions are hard

Cornell University

the Sim

arXiv.org > cs > arXiv:1710.00794

Search or Art

(Help | Advanced

Computer Science > Artificial Intelligence

What Does Explainable AI Really Mean? A New Conceptualization of Perspectives

Derek Doran, Sarah Schulz, Tarek R. Besold

(Submitted on 2 Oct 2017)

We characterize three notions of explainable AI that cut across research fields: opaque systems that offer no insight into its algo- rithmic mechanisms; interpretable systems where users can mathemat- ically analyze its algorithmic mechanisms; and comprehensible systems that emit symbols enabling user-driven explanations of how a conclusion is reached. The paper is motivated by a corpus analysis of NIPS, ACL, COGSCI, and ICCV/ECCV paper titles showing differences in how work on explainable AI is positioned in various fields. We close by introducing a fourth notion: truly explainable systems, where automated reasoning is central to output crafted explanations without requiring human post processing as final step of the generative process.

SSRN

Browse Subscriptions Rankings Submit a paper My Library Blog ↗

Download This Paper | Open PDF in Browser | Add Paper to My Library

Contestability and Professionals: From Explanations to Engagement with Algorithmic Systems

15 Pages • Posted: 10 Jan 2019

Daniel Kluttz
University of California, Berkeley School of Information

Nitin Kohli
UC Berkeley School of Information

Deirdre K. Mulligan
University of California, Berkeley - School of Information

Date Written: August 24, 2018

Proceedings of Machine Learning Research 81:1–15, 2018 Conference on Fairness, Accountability, and Transparency

Discrimination in Online Advertising A Multidisciplinary Inquiry

Amit Datta AMITDATTA@CMU.EDU
Anupam Datta DANUPAM@CMU.EDU
Carnegie Mellon University

Jael Makagon JAEL@BERKELEY.EDU
Deirdre K. Mulligan DMULLIGAN@BERKELEY.EDU
University of California, Berkeley

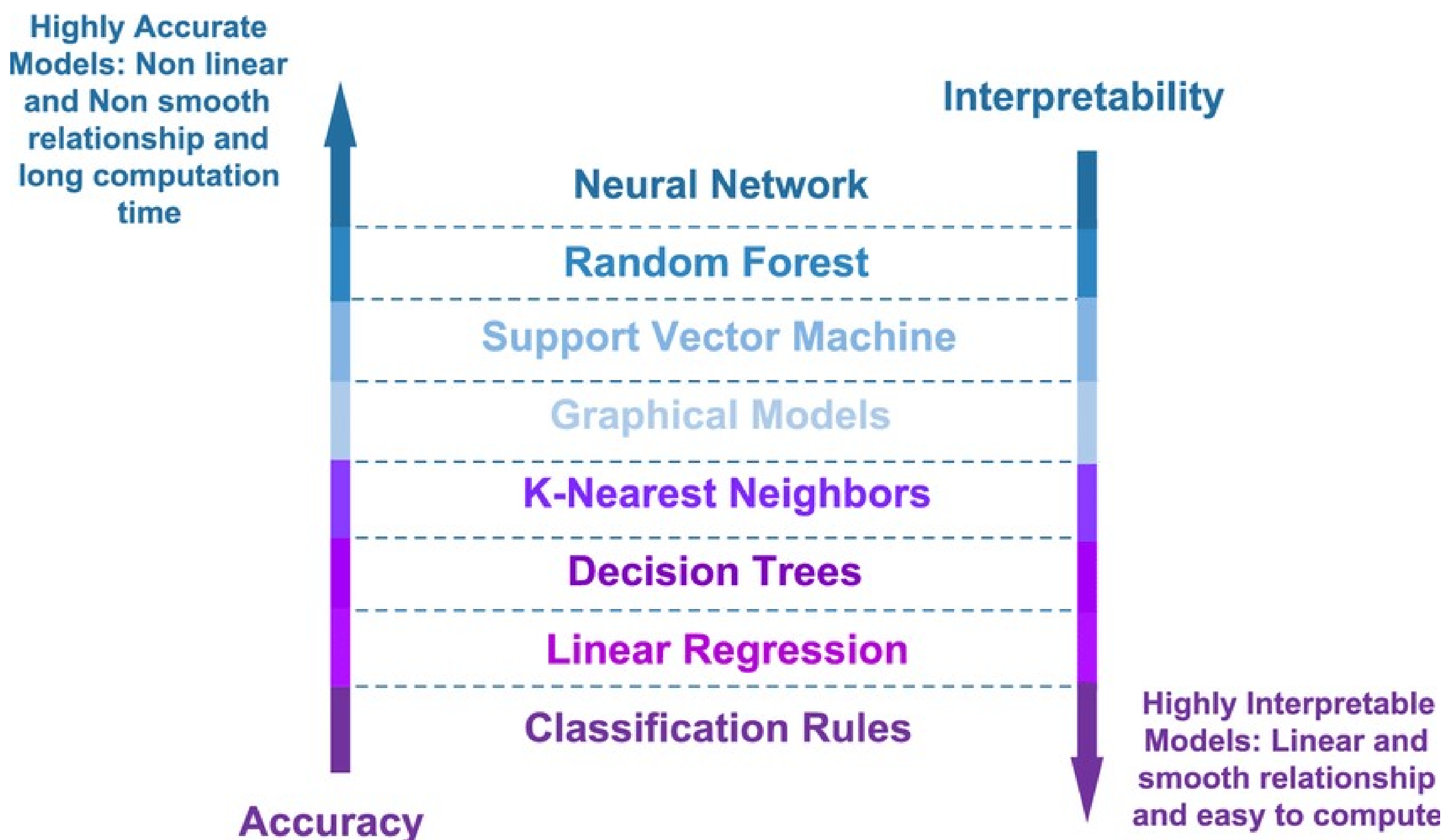
Michael Carl Tschantz MCT@ICSI.BERKELEY.EDU
International Computer Science Institute

Source: Michelle Carney

CUNY School of Professional Studies

10

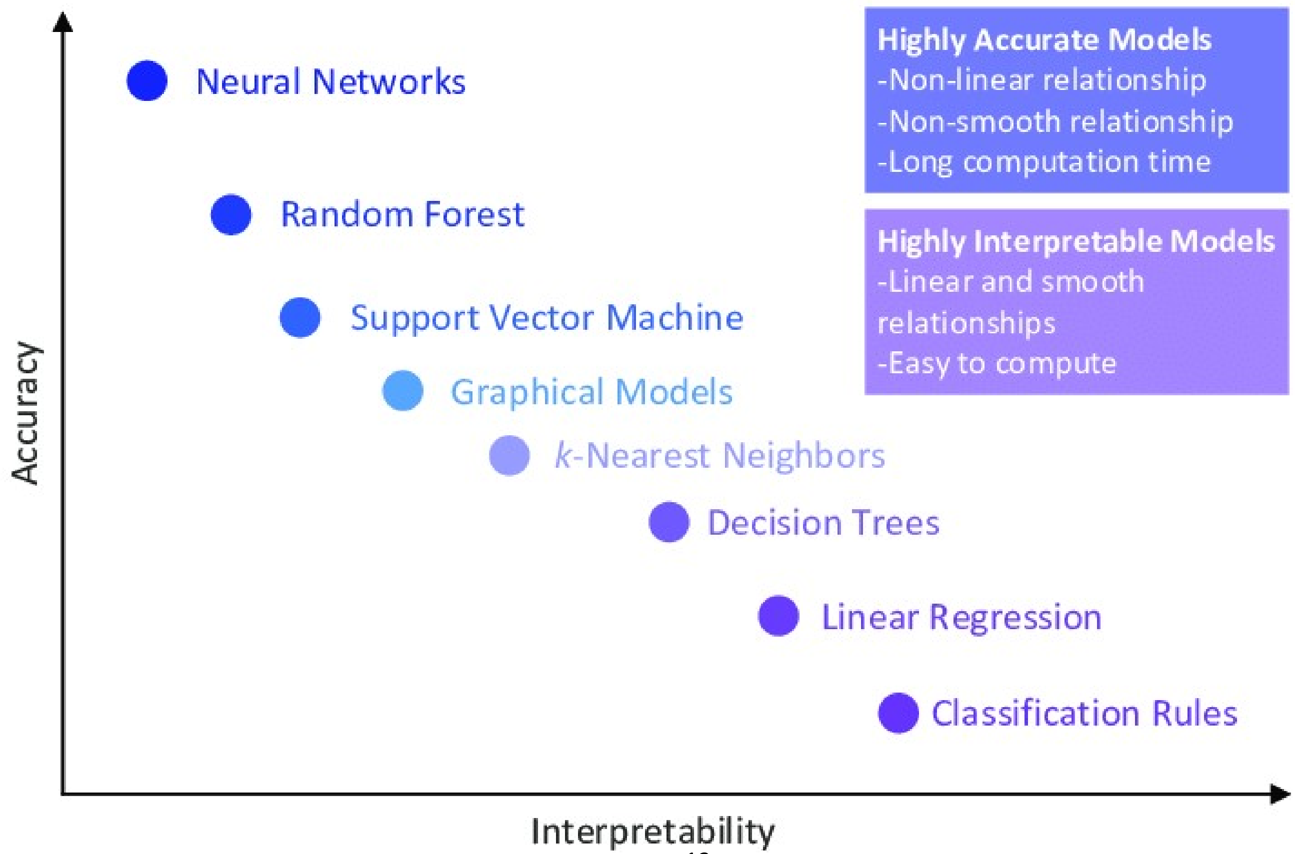
Interpretability



CUNY School of Professional Studies

11

Interpretability



CUNY School of Professional Studies

Mitigating Bias

Sources of bias in AI

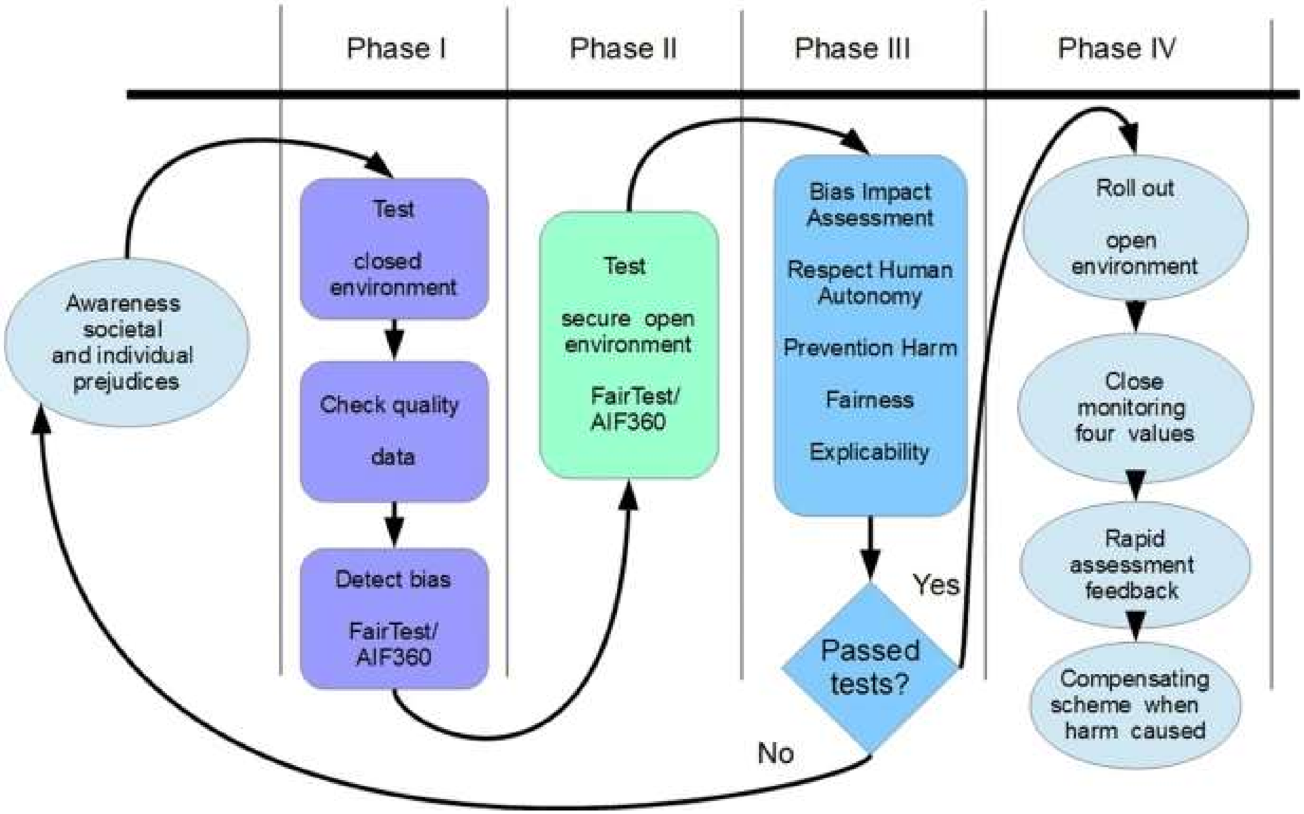
1. Sampling Bias
   Occurs when the training data is not representative of the population it serves, leading to poor performance and biased predictions for certain groups.
2. Algorithmic Bias
   Results from the design and implementation of the algorithm, which may prioritize certain attributes and lead to unfair outcomes.
3. Representation Bias
   Happens when a dataset does not accurately represent the population it is meant to model, leading to inaccurate predictions.
4. Confirmation Bias
   Materializes when an AI system is used to confirm pre-existing biases or beliefs held by its creators or users.
5. Measurement Bias
   Emerges when data collection or measurement systematically over- or underrepresents certain groups.
6. Interaction Bias
   Occurs when an AI system interacts with humans in a biased manner, resulting in unfair treatment.
7. Generative Bias
   Occurs in generative AI models, like those used for creating synthetic data, images, or text

Types of AI Bias

1. **Algorithm**
   - Systematic
   - Consistent
2. **Cognitive**
   - Human input
3. **Confirmation**
   - Pre-existing
4. **Learning models & data**
   - Supervised: Diversity of stakeholders
   - Unsupervised
5. **Balanced Team**
   - Varied AI team: Racially, Economically, Gender, Innovators, Creators, Consumers
6. **Emerging Risks**
   - Typically GenAI e.g. copyright infringement

How to avoid bias

1. **Data processing**
   - Mindful of each step
   - Pre-processing
   - In-processing
   - Post-processing
2. **Continuous Monitoring**
   - Real-world data
   - Third party
3. **Confirmation**
   - Pre-existing
4. **Out-group homogeneity**
   - Assumption about group
5. **Exclusion**
   - Data left out



Source: Springer

ways to build Ethics into AI)

01

Create an Ethical Culture

- *Build diverse teams*
- *Cultivate an ethical mindset*
- *Conduct social systems analysis*

02

Be Transparent

- *Understand your values*
- *Give users control over their data*
- *Take feedback*

03

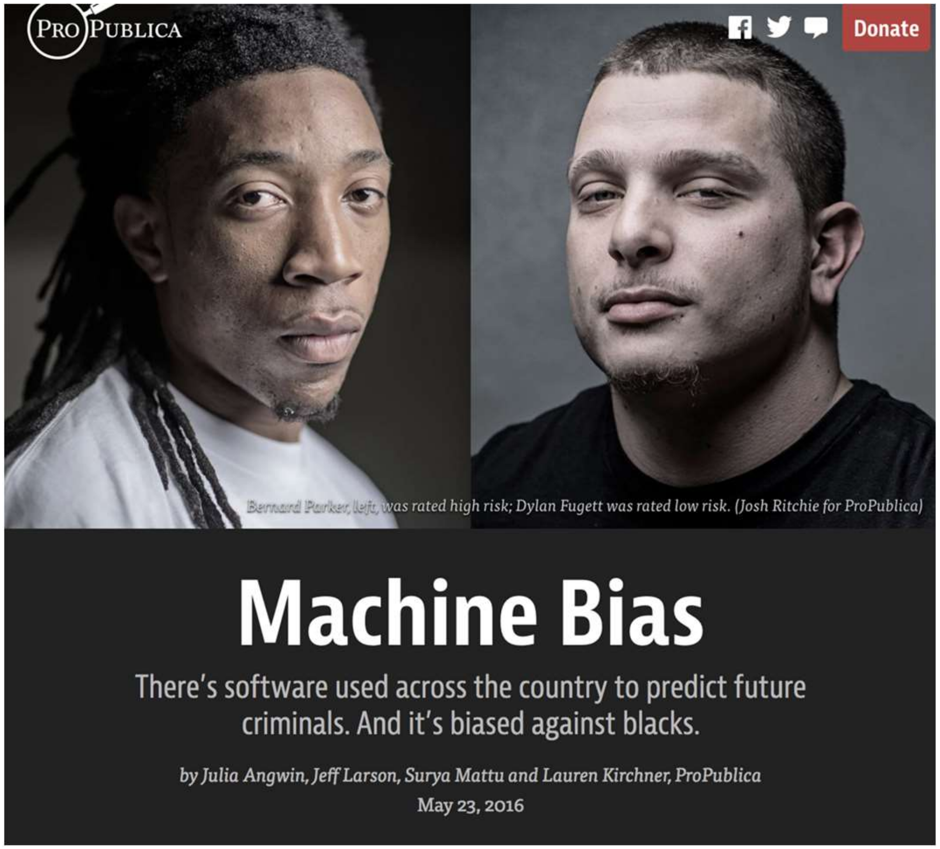
Remove Exclusion

- *Understand the factors involved*
- *Prevent data set bias*
- *Prevent association bias*
- *Prevent confirmation bias*
- *Prevent automation bias*
- *Mitigate interaction bias*

Source: Kathy Baxter (AI Ethics/SalesForce)

More case studies

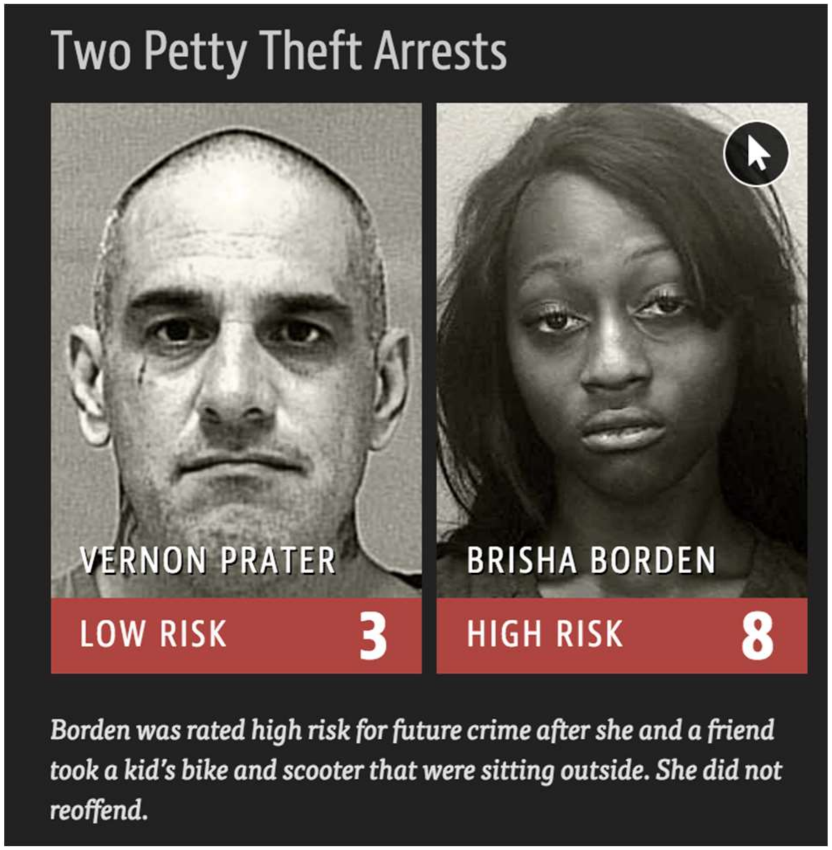
Parole recidivism

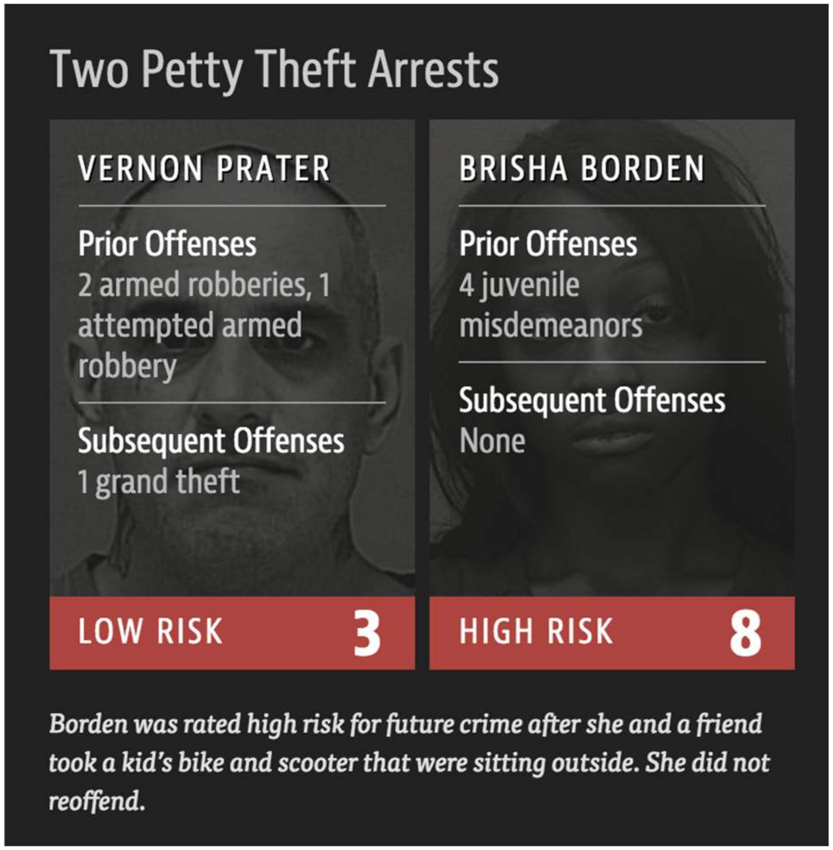


Source: Michelle Carney

19

Parole recidivism





Source: Michelle Carney

CUNY School of Professional Studies

20

Gender Classifier	Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap
Microsoft	94.0%	79.2%	100%	98.3%	20.8%
FACE++	99.3%	65.5%	99.2%	94.0%	33.8%
IBM	88.0%	65.3%	99.7%	92.9%	34.4%

Source: Buolamwini and Gebru (2018)

CUNY School of Professional Studies

21

ACLU

Email address | ZIP code | GET UPDATES

BECOME A MEMBER / RENEW / TAKE ACTION / DONATE

ISSUES KNOW YOUR RIGHTS DEFENDING OUR RIGHTS BLOGS ABOUT SHOP

FREE FUTURE

FREE FUTURE

Amazon's Face Recognition Falsely Matched 28 Members of Congress With Mugshots

By Jacob Snow, Technology & Civi
JULY 26, 2018 | 8:00 AM

TAGS: Face Recognition Technology, Surve

Amazon's face surveillance technology is the target of growing opposition nationwide, and today, there are 28 more causes for concern. In a test the

Source: Michelle Carney

CUNY School of Professional Studies