

Emma Willard School

Coded Bias: How Racism Shows Up In Artificial Intelligence Algorithms

Part 2



Prepared by: Chiara Shah
January 17th, 2022

People Assume Algorithms Are Neutral

We think of computers as

Impartial

Random

Non-feeling

Unbiased

Secure

Automated

Last Year's Presentation...

Joy Buolamwini



Latanya Sweeney



Amazon's Recruiting Algorithm

COMPAS Algorithm

Princeton University's language Bot

Self-driving cars can't see people of color

Healthcare providers favoring white patients over black patients

Dr. Safiya Noble's Story



Internet Studies, Digital Media Scholar

UCLA Professor of Gender Studies and
African American Studies

Co-Founder and Co-Director of the
UCLA Center for Critical Internet Inquiry

2021 MacArthur Fellow

MacArthur Fellow Safiya Noble is highlighting the ways digital technologies and internet architectures magnify racism, sexism, and harmful stereotypes.



Dr. Safiya Noble's Story



Article from just yesterday...

The Biggest Danger of AI Isn't Skynet — It's Human Bias That Should Scare You

Artificial intelligence doesn't hold a candle to the human capacity for harm.



By John Loeffler

Jan 16, 2022



“...few people rail against the ways that AI systems are already harming humans.”



Article from just yesterday...

What if the majority of selfies that are fed into a facial recognition algorithm depict predominantly light-skinned, "white" faces? Well, then that algorithm will become very good at detecting those kinds of faces.

So, how do you think it would do when tasked to detect and identify darker-skinned, "black and brown" faces? In this respect, we could say that the algorithm has picked up a bias towards identifying lighter-skinned faces.



Article from just yesterday...

What about loan applications?

If you were to feed every loan application on record into a machine learning algorithm, along with whether or not that application was approved or rejected, then your machine learning algorithm would be very good at accepting the kinds of loan applications that have been previously accepted and rejecting those that have been previously rejected.



Article from just yesterday...

But what if the data you fed it consisted largely of

- 1) rejected loan applications from minority applicants with impeccable credit records and
- 2) accepted applications from white applicants with less than impeccable credit?

...then the algorithm would be inadvertently trained to hone in on the race of the applicants, rather than the credit scores, and assume that people from minority backgrounds or with darker skin tones should be rejected, since that seems to be the underlying pattern of the loan approval process.



Let's think about another example...

The last two examples led me to think about **college applications**.

If you were to feed every college application on record into a machine learning algorithm, along with whether or not that application was approved or rejected, then your machine learning algorithm would be very good at accepting the kinds of applications that have been previously accepted and rejecting those that have been previously rejected.



Let's think about another example...

But what if the data you fed it consisted largely of

- 1) rejected college applications from minority applicants with impeccable high school records
- 2) accepted applications from white applicants with less than perfect high school records?

...then the algorithm would be inadvertently trained to hone in on the race of the applicants, rather than their credentials, and assume that people from minority backgrounds or with darker skin tones should be rejected, since that seems to be the underlying pattern of the college acceptance process.



Let's think about another example...

What about _____?

If you were to feed every _____ on record into a machine learning algorithm, along with whether or not that _____ was approved or rejected, then your machine learning algorithm would be very good at accepting the kinds of _____ that have been previously accepted and rejecting those that have been previously rejected.



Let's think about another example...

But what if the data you fed it consisted largely of

- 1) _____ from minority applicants with _____
- 2) _____ from white applicants with _____

...then the algorithm would be inadvertently trained to hone in on the race of the applicants, rather than their credentials, and assume that people from minority backgrounds or with darker skin tones should be rejected, since that seems to be the underlying pattern of the _____ process.



Article from just yesterday...

“Since the 1990s, police departments the world over have relied on crime statistics to produce a “predictive policing” model for law enforcement, essentially to place police resources in the areas where the data says “most of the crime” takes place. But if most of your police resources are directed to a specific area, perhaps an area where minorities live, then you are also more likely to find crime in that area.”



Microsoft Chatbot

“In 2016, Microsoft released an AI-based conversational chatbot on Twitter that was supposed to interact with people through tweets and direct messages. However, it started replying with highly offensive and racist messages within a few hours of its release.

The chatbot was trained on anonymous public data and had a built-in internal learning feature, which led to a coordinated attack by a group of people to introduce racist bias in the system. Some users were able to inundate the bot with misogynistic, racist and anti-Semitic language.”

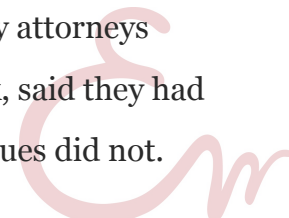


Remote Surveillance of workers during COVID

From the 9/24/2021 New York Times Article “**Keystroke tracking, screenshots, and facial recognition: The boss may be watching long after the pandemic ends**”

Attorneys required to use the new [face-scanning software](#) while working from home said they understood the need for security because reviewing sensitive documents is part of the job. But many felt the remote-work surveillance had gone too far. The facial recognition systems, they said, felt intrusive, dysfunctional or annoying, booting them out of their work software if they shifted in their seat, rested their eyes, adjusted their glasses, wore a headband or necklace, went to the bathroom or had a child walk through their room.

Even more problematically, some facial recognition systems have been shown in research to perform worse with people of color because the algorithms are less accurate at identifying people with darker skin tones. That leaves many attorneys fearful that they could be penalized because of the color of their skin. Three attorneys, all of whom are Black, said they had routinely struggled to be recognized by the face-scanning systems in a way that their lighter-skinned colleagues did not.



Potential Harm

POTENTIAL HARMS FROM ALGORITHMIC DECISION-MAKING

INDIVIDUAL HARMS			COLLECTIVE SOCIAL HARMS
ILLEGAL DISCRIMINATION	UNFAIR PRACTICES		
HIRING		LOSS OF OPPORTUNITY	
EMPLOYMENT			
INSURANCE & SOCIAL BENEFITS			
HOUSING			
EDUCATION			
CREDIT		ECONOMIC LOSS	
DIFFERENTIAL PRICES OF GOODS			
LOSS OF LIBERTY		SOCIAL STIGMATIZATION	
INCREASED SURVEILLANCE			
STEREOTYPE REINFORCEMENT			
DIGNATORY HARMS			

Where is the bias coming from?

1. Prejudiced assumptions made during the algorithm development process
2. Prejudices in the training data
3. Lack of complete training data
4. Too much training data that is skewed in a particular direction

Where is the bias coming from?

“...algorithms with too much data, or an over-representation, can skew the decision toward a particular result. Researchers at Georgetown Law School found that an estimated 117 million American adults are in facial recognition networks used by law enforcement, and that African-Americans were more likely to be singled out primarily because of their over-representation in mug-shot databases.”

Will AI ever be completely unbiased?

In theory...yes.

But practically, probably not.

What we can try to do is *minimize* the bias.

Detecting and Removing AI Bias

“...organizations need effective frameworks, toolkits, processes, and policies for recognizing and actively mitigating AI bias. Available open source tooling can assist in testing AI applications for specific biases, issues, and blind spots in data.”

Frameworks

Here are some examples:

[The Aletheia Framework from Rolls Royce](#) provides a 32-step process for designing accurate and carefully managed AI applications.

[Deloitte's AI framework](#) highlights six essential dimensions for implementing AI safeguards and ethical practices.

A [framework from Naveen Joshi](#) details cornerstone practices for developing trustworthy AI. It focuses on the need for explainability, machine learning integrity, conscious development, reproducibility, and smart regulations.

Toolkits

Here are some particularly useful toolkits:

[AI Fairness 360](#) from IBM is an extensible (and open source) toolkit that enables examination, reporting, and mitigation of discrimination and bias in machine learning models.

[IBM Watson OpenScale](#) provides real-time bias detection and mitigation and enables detailed explainability to make AI predictions trusted and transparent.

[Google's What-If Tool](#) offers visualization of machine learning model behavior, making it simple to test trained models against machine learning fairness metrics to root out bias

We need people doing this job

Public Interest Technologist.

- Technology practitioners who focus on social justice, the common good, and/or the public interest.
- Should have a background in law, technology, and policy, and not necessarily a computer science degree.

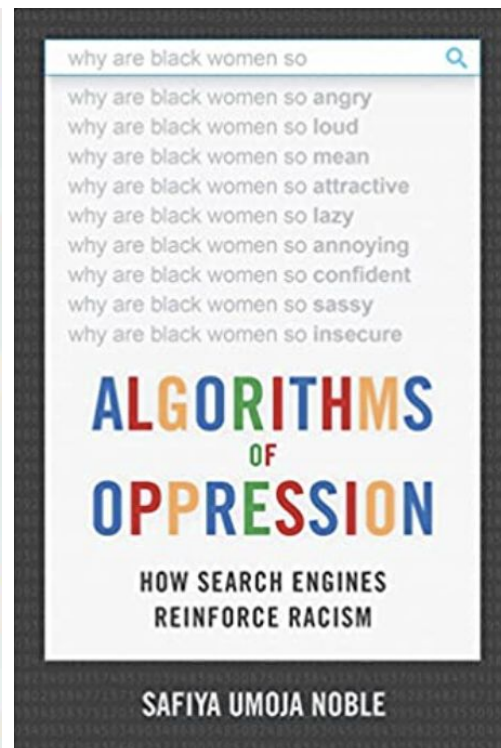
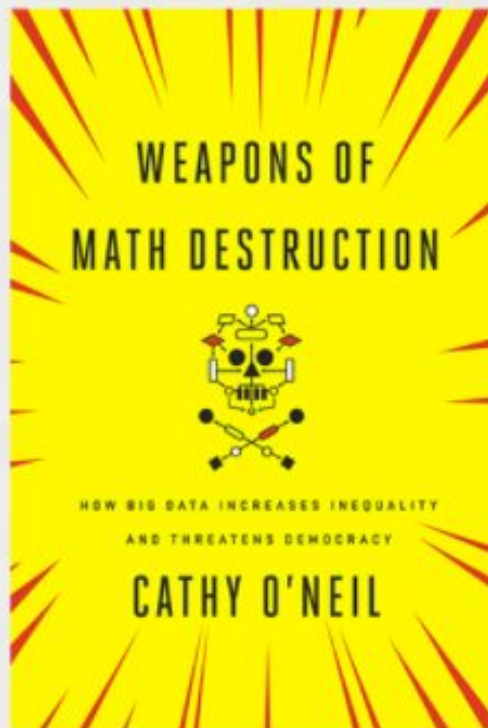
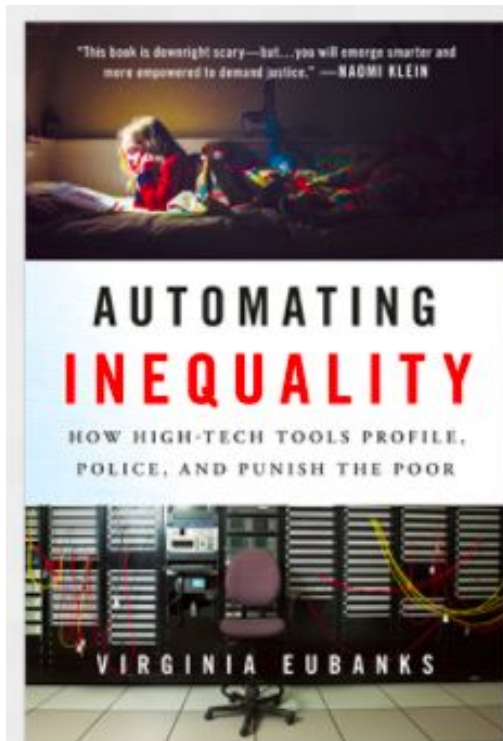


“It’s important for algorithm operators and developers to always be asking themselves: Will we leave some groups of people worse off as a result of the algorithm’s design or its unintended consequences?”

“Algorithmic Bias Detection and Mitigation: Best Practices and Policies to Reduce Consumer Harms”



For More Information...



For More Information...

Algorithm Bias: <https://guides.lib.fsu.edu/algorithm>

A website with definitions, videos, examples, and links.

The section “Understanding Algorithms” walks you through how algorithms work so you can get a foundational understanding before delving more deeply into the topic.

Other sections include:

- Further Reading - Books

- Further Reading - Research and Articles

- Examples of Algorithm Bias

- Organizations fighting Algorithm Bias: “Black in AI” , “LatinX in AI”, etc



For More Information...



Coded Bias
movie.

Now on
Netflix!!!! I urge
you to watch it.



Questions?



Sources (includes sources from 2021 Presentation)

Abril, Danielle, and Harwell, Drew. "Keystroke tracking, screenshots, and facial recognition: The boss may be watching long after the pandemic ends." *The Washington Post*. 24 Sept. 2021.

<https://www.washingtonpost.com/technology/2021/09/24/remote-work-from-home-surveillance/> Accessed 16 Jan. 2022.

"AI, Ain't I Woman?" *YouTube*, uploaded by Joy Buolamwini, Google, 28 June 2018, www.youtube.com/watch?v=QxuyfWoVV98. Accessed 18 Jan. 2021.

Buolamwini, Joy, and Timnit Gebru. "Gender Shades." *Gender Shades*, MIT Media Lab, 2018, gendershades.org/. Accessed 18 Jan. 2021.

"Dr. King Said It: I'm Black and I'm Proud." *YouTube*, www.youtube.com/watch?v=Suw_CQ3zfTY&feature=youtu.be. Accessed 18 Jan. 2021.

Hadhazy, Adam. "Biased Bots: Artificial-intelligence Systems Echo Human Prejudices." *Princeton University*, 18 Apr. 2017, www.princeton.edu/news/2017/04/18/biased-bots-artificial-intelligence-systems-echo-human-prejudices. Accessed 18 Jan. 2021.

"How to Keep Human Bias Out of AI." *YouTube*, uploaded by TED, Google, 12 Apr. 2019, www.youtube.com/watch?v=BRRNeBKwvNM&feature=youtu.be. Accessed 18 Jan. 2021.



Jacob, Shomrom. "AI Bias is prevalent but preventable – here's how to root it out." *venturebeat.com*, 8 Aug. 2021.

<https://venturebeat.com/2021/08/08/ai-bias-is-prevalent-but-preventable-heres-how-to-root-it-out/>. Accessed 16 Jan. 2022.

Kantarci, Atakan. "Bias in AI: What It Is, Types & Examples, How & Tools to Fix It." *AIMultiple*, 9 Jan. 2021, research.aimultiple.com/ai-bias/. Accessed 18 Jan. 2021.

Larson, Jeff, et al. "How We Analyzed the COMPAS Recidivism Algorithm." *ProPublica*, 23 May 2016, www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm. Accessed 18 Jan. 2021.

Lee, Nicol Turner, et al. *Algorithmic Bias Detection and Mitigation: Best Practices and Policies to Reduce Consumer Harms*. Brookings Institute, 22 May 2019. *Brookings*, www.brookings.edu/research/algorithmic-bias-detection-and-mitigation-best-practices-and-policies-to-reduce-consumer-harms/. Accessed 18 Jan. 2021.

Loeffler, John. "The Biggest Danger of AI Isn't Skynet – It's Human Bias That Should Scare You." *Interesting Engineering*, 16 Jan. 2022. <https://interestingengineering.com/the-biggest-danger-of-ai-isnt-skynet-its-human-bias> Accessed 16 Jan. 2022.

MacArthur Foundation. "Safiya Noble, Internet Studies and Digital Media Scholar | 2021 MacArthur Fellow (Extended)". *YouTube*, 7 Dec. 2021. <https://www.youtube.com/watch?v=WFzuMCUIah8>. Accessed 16 Jan. 2022.



Nouri, Steve. "The Role of Bias in Artificial Intelligence." *Forbes*, 4 Feb. 2021.

<https://www.forbes.com/sites/forbestechcouncil/2021/02/04/the-role-of-bias-in-artificial-intelligence/?sh=66cd016a579d>. Accessed 16 Jan. 2022.

Pangburn, DJ. "Schools Are Using Software to Help Pick Who Gets In. What Could Go Wrong?" *Fast Company*, 17 May 2019, www.fastcompany.com/90342596/schools-are-quietly-turning-to-ai-to-help-pick-who-gets-in-what-could-go-wrong. Accessed 18 Jan. 2021.

Perry, Andre M., and Nicol Turner Lee. "AI Is Coming to Schools, and If We're Not Careful, so Will Its Biases." *The Avenue*, Brookings Institute, 26 Sept. 2019, www.brookings.edu/blog/the-avenue/2019/09/26/ai-is-coming-to-schools-and-if-were-not-careful-so-will-its-biases/. Accessed 18 Jan. 2021.

"Race, Technology, and Algorithmic Bias." *YouTube*, uploaded by Harvard University, Radcliffe Institute, Google, 7 May 2019, www.youtube.com/watch?v=Y6fUc5_whX8. Accessed 18 Jan. 2021.

Schneier, Bruce, editor. "Public-Interest Technology Resources." *Public Interest Technology*, WordPress, 16 Jan. 2021, public-interest-tech.com/. Accessed 18 Jan. 2021.



"Sojourner Truth: Ain't I a Woman." *National Park Service*, 17 Nov. 2017, www.nps.gov/articles/sojourner-truth.htm. Accessed 18 Jan. 2021.

"Sojourner Truth's 'Ain't I a Woman': Nkechi at TEDxFiDiWomen." *YouTube*, uploaded by TEDx Talks, Google, 7 Feb. 2013, www.youtube.com/watch?v=eUdxsQ0Qsrc&feature=youtu.be. Accessed 18 Jan. 2021.

Sweeney, Latanya. "Discrimination in Online Ad Delivery." *Data Privacy Lab*, President and Fellows Harvard University, 2013, dataprivacylab.org/projects/onlineads/index.html. Accessed 18 Jan. 2021.

Truth, Sojourner. "Ain't I a Woman?" May 1851. *Andrew Jackson's Hermitage*, Andrew Jackson Foundation, thehermitage.com/wp-content/uploads/2016/02/Sojourner-Truth_Aint-I-a-Woman_1851.pdf. Accessed 18 Jan. 2021. Speech.

