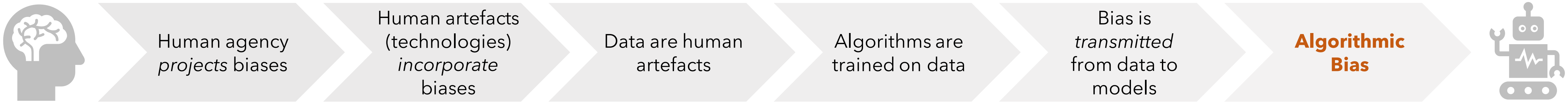
THE AI (IN)JUSTICE LEAGUE

Why Algorithmic Bias is problematic and how to handle it.

Diletta Goglia

Introduction to Algorithmic Bias

Definition



Human agency *projects* biases

Human artefacts (technologies) *incorporate* biases

Data are human artefacts

Algorithms are trained on data

Bias is *transmitted* from data to models

Algorithmic Bias

Different kind of Bias

Data-driven Bias: incomplete / incorrect / poorly labelled datasets (ex. not enough example of each class)

Bias through interactions: learned from interactions with other users or agents (ex. Microsoft Tay twitterbot)

Similarity Bias: filter bubbles, set of similar informations that tend to confirm each other (ex. recommending algorithms).

Conflicting goals Bias: narrow AI creates negative consequences in lateral or secondary applications (ex. different jobs for men and women).

Emergent Bias: algorithmic version of "confirmation bias": it reinforce prejudices and questionable behaviour that are already in society.

Methodology

Social analysis comes first.

- Not all biases are problematic, it depends on the culture. Whether bias is problematic or neutral is an ethical and social issue. The wrongness of bias is always relative to the cultural context in which it is active: framing an algorithmic bias as problematic cannot precede but only follow its ethical and social determination.
- For this reason, in this analysis, social and ethical reflections are put first: we will see some examples of algorithmic bias and we will frame it into our cultural background to understand whether or not they constitute an ethical problem.

Then… action!

- If the previous social analysis will report that human bias is imitated, transmitted, and even amplified from AI systems with harmful consequences (and we will see that it is), then this is a problem to face seriously and immediately.
- In fact, bias does not represent a problem unless algorithms start discriminating certain groups or categories of people: we will see clearly harmful examples of bias in predictive algorithms that led to exclusionary experiences, and we will also see how the problem is addressed and how it should be.
- Even if the issue is serious, efforts to fix it are still too few: for this reason we will explore appropriate technical practices to handle it.

Algorithmic Bias: why it is problematic

“There is a wealth of empirical evidence showing that the use of AI systems can often replicate historical and contemporary conditions of injustice, rather than alleviate them”. [4]

“Algorithms that may conceal hidden biases are already routinely used to make vital, financial and legal decisions. Proprietary algorithms are used to decide, for instance, who gets a job interview, who gets granted parole, and who gets a loan”. [5]

“The tech is already making important decisions about your life and potentially ruling over which political advertisements you see, how police officers are deployed in your neighbourhood, and even predicting your home’s risk of fire”. [8]

“There’s no guarantee companies building or using this tech will make sure it’s not discriminatory, especially without a legal mandate to do so”. [8]

Impact

The Georgia Institute of Technology indicated that autonomous vehicles might have a more difficult time detecting pedestrians with darker skin. Those cars were programmed mostly by young white male engineers, so the systems were 5% less accurate when recognizing people who have darker skin than those with lighter skin: fewer images of people with darker skin tones were used during programming and testing. [6] → **Data-driven bias**

Amazon tried to use AI to build a résumé-screening tool to make the process of sorting through job applications more efficient. It built a screening algorithm using résumés the company had collected for a decade, but those résumés tended to come from men. That meant the system, in the end, learned to discriminate against women. [8][15] → **Conflicting goals bias**

A recent research article on *Science* showed that an algorithm, widely used for population health management to predicts which patients will benefit from extra medical care, has a significant racial bias: it dramatically underestimates the health needs of the sickest black patients. The magnitude of bias is large: removing it would more than double the number of black patients eligible for a program that gives extra medical help. [11][16]

In 2016, the professional networking site LinkedIn was discovered to recommend male variations of women's names in response to search queries. The site did not make similar recommendations in searches for male names. The company said this was the result of an analysis of users' interactions with the site. [14] → **Bias through interaction**

PredPol's drug crime prediction algorithm (still in use since 2012) was trained on data that was heavily biased by past housing segregation so it would more frequently send police to certain neighbourhoods where a lot of racial minority lived. Arrests increased there and arrest data was fed back into the algorithm and it would predict more future drug arrest in those areas and send the police there again. [1] → **Emergent bias**

The struggle is real!

In USA predictive algorithms are routinely used to perform risk assessments in criminal sentencing.

These algorithms output a risk factor to quantify the likelihood that a person would commit a crime in the future.

Today they are used in Arizona, Colorado, Delaware, Kentucky, Louisiana, Oklahoma, Virginia, Washington, and Wisconsin.

In 2016 investigative journalist from ProPublica found that one of the most popular algorithm was significantly disadvantageous to black defendants.

They in fact were nearly twice likely to be predicted as reoffending when they did not. At the same time, a white defendant is more likely to be predicted as not reoffending, when in fact he/she did.

Scores were disproportionately skewed to suggest blacks to be at risk of relapse, 77% more often than whites.

A person subject to this algorithm is more likely to be incarcerated if he/she is black.

As computer scientists, fighting these unethical algorithms is compulsory, not optional.

Predicted to reoffend, but don't	Predicted to not reoffend, but do
False positive	**False negative**
Black: 44.9%	Black: 28.1%
White: 23.5%	White: 47.7%

A ProPublica's study shows a bias against black people in a particularly widespread algorithm used in criminal justice.

References: [2][12]

Current state in addressing the issue:

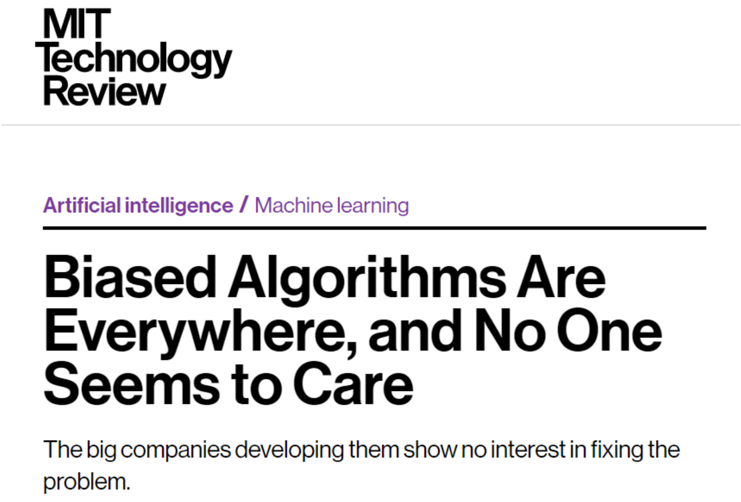
→ Too little interest and no professionals training.

→ The fact that bias is a necessary and non-avoidable condition in data should not justify an eventually bad output or a careless behaviour from companies and governments.

A whole picture of the situation.

Yaël D. Eisenstat, «*The Real Reason Tech Struggles With Algorithmic Bias*,» on *Wired,* 02 12 2019.

- " […] If the question is whether the tech industry doing enough to address these biases, the straightforward response is no."
- "Humans cannot wholly avoid bias, as countless studies and publications have shown. Insisting otherwise is an intellectually dishonest and lazy response to a very real problem."
- "I did not know anyone who intentionally wanted to incorporate bias into their work. I also did not find anyone who actually knew what it meant to counter bias in any true and methodical way."
- "While tech companies often have mandatory "managing bias" training to help with diversity and inclusion issues, I did not see any such training on the field of cognitive bias and decision making, particularly as it relates to how products and processes are built and secured. […] They demonstrate why tasking untrained engineers and data scientists with correcting bias is, at the broader level, naïve, and at a leadership level insincere."



MIT Technology Review

Artificial intelligence / Machine learning

Biased Algorithms Are Everywhere, and No One Seems to Care

The big companies developing them show no interest in fixing the problem.

> […] "Stakeholders, including the companies that develop and apply machine learning systems and government regulators, show little interest in monitoring and limiting algorithmic bias."

Solutions

New professionals.

Investing in training computers engineers and computer scientists to address appropriately this issue

Raise awareness among users.

Raise awareness among people about how these technologies work and how to be careful in fields where they are used

Legislation.

The Toronto Declaration calls for applying a human rights framework to harms caused by algorithmic bias [23]

Legislation (II).

The Algorithmic Accountability Act (April 2019) requires companies to evaluate and fix biased computer algorithms that result in inaccurate, unfair, or discriminatory decisions. It is the first legislative effort to regulate AI systems across industries in the US [7]

Cooperation with human.

It is important to be critical about AI responses and recommendations (active role and collaboration between humans and machine) [1]

Modular AI architectures

in which implicit learning of statistical regularities can be compartmentalized and augmented with explicit instructions or rules of appropriate conduct (a way to ensure that what is taking into consideration from the machine is socially acceptable) [24]

Technical solution.

Mathematical formulations of non-discrimination in decision-making. [24] Also, building NNs can help *detecting* biases [24]

Transparency

(i.e. the ability to examine inputs and outputs to examine why an algorithm is giving certain recommendations) is very important (but still difficult for example in deep learning methods, because of hidden layers) [1]

Full spectrum inclusion of data.

If we want to have less biased algorithms and less risk of discrimination, we need more training data on protected categories or less represented classes, like race, gender or age [1][17] . It is important to expose algorithms to well balanced training samples. [7]

Full spectrum inclusion of developers.

The design of AI systems is primarily the domain of white, male engineers: algorithmic bias may be minimized by expanding inclusion in the ranks of those designing AI systems [20][21][22]

Keep attention to correlation.

By not specifying protected informations like race or gender, maybe algorithms cannot discriminate. But these classes may emerge as correlated features (ex. because of segregation in US, zip code can be strongly correlated to race) [1]

Collect users' feedback.

Building platform that can detect bias by collecting people experience [19]

Thanks for your attention!

UNIVERSITÀ DI PISA

Social and Ethical issues in Information Technology
A.Y. 2019/20
Master Degree in Computer Science
Artificial Intelligence curriculum

Related topics

Social and ethical issues related to Bias.

Here:

- NN & Machine Learning
- Big Data

But also:

- Singularity and Superintelligence
- Responsibility allocation
- Anthropomorphism

References

[1] J. Ashe, «Algorithmic Bias and Fairness» 13 12 2019. [Online]. Available: https://www.patreon.com/crashcourse

[2] H. Farid, «The danger of predictive algorithms in criminal justice» 02 10 2018. [Online]. Available: https://www.youtube.com/watch?v=p-82YeUPQh0

[3] Y. D. Eisenstat, «The Real Reason Tech Struggles With Algorithmic Bias» *WIred*, 02 12 2019.

[4] A. Zimmermann, E. Di Rosa e K. Hochan, «Technology Can't Fix Algorithmic Injustice» *Boston Rewiew*, 09 01 2020.

[5] W. Knight, «Biased Algorithms Are Everywhere, and No One Seems to Care» *MIT Technology Review*, 12 07 2017.

[6] M. Fischer, «Machine learning can't fix algorithmic bias. But humans can» *Quartz at work*, 20 02 2020.

[7] S. Schellenberg, «How biased algorithms perpetuate inequality» *NewStatesman*, 29 04 2020.

[8] R. Heilweil, «Why algorithms can be racist and sexist» *Vox*, 18 02 2020.

[9] K. Haoarchive, «This is how AI bias really happens—and why it's so hard to fix» *MIT Technology Overview*, 02 02 2019.

[10] R. Hauser, «Can we protect AI from our biases?» *TED Institute*, 12 02 2018. [Online] Available: https://www.youtube.com/watch?v=eV_tx4ngVT0

[11] Z. Obermeyer, B. Powers, C. Vogeli, S. Mullainathan, «Dissecting racial bias in an algorithm used to manage the health of populations» *Science,* 25 Oct 2019, Vol. 366, Issue 6464, pp. 447-453

[12] J. Angwin, J. Larson, S. Mattu and L. Kirchner, «Machine Bias» *ProPublica*, 23 05 2016.

[13] K. Hammond, «5 unexpected sources of bias in artificial intelligence», *Techcrunch,* 11 12 2016.

[14] M. Day, «How LinkedIn's search engine may reflect a gender bias», *The Seattle Times*, 31 08 2016.

[15] J. Dastin, «Amazon scraps secret AI recruiting tool that showed bias against women», *Reuters*, 09 10 2018.

[16] C. Y. Johnson, «Racial bias in a medical algorithm favors white patients over sicker black patients», *The Washington Post,* 24 10 2019

[17] M. Veale, R. Binns, «Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data», *Big Data & Society*, 2017

[18] R. Thomas, «Getting Specific About Algorithmic Bias», *YouTube,* 01 10 2019. [Online] Available: https://www.youtube.com/watch?v=S-6YGPrmtYc

[19] J. Buolamwini, «How I'm fighting bias in algorithms», *YouTube*, 29 03 2017. [Online] Available: https://www.youtube.com/watch?v=UG_X_7g63rY

[20] K. Crawford, «Artificial Intelligence's White Guy Problem», *The New York Times*, 25 06 2016

[21] «How to Prevent Discriminatory Outcomes in Machine Learning», *World Economic Forum*, 12 03 2018

[22] A. Jobin, M. Ienca, E. Vayena, « The global landscape of AI ethics guidelines». *Nature Machine Intelligence*. Vol. 1 Issue 9, 02 09 2019 389–399

[23] «The Toronto Declaration: Protecting the rights to equality and non-discrimination in machine learning systems», *Human Rights Watch,* 03 07 2018

[24] A. Caliskan, J. J. Bryson, A. Narayanan, «Semantics derived automatically from language corpora contain human-like biases», *Science,* vol. 365, 14 04 2017