Bias in artificial intelligence

Grigorina BOCE,
Department of Informatics and Scientific Education, Mediterranean University of Albania, Tirana, Albania
grigorina.boce@umsh.edu.al

Abstract

As artificial intelligence entities and usage will continue to increase, as assumed at an exponential rate, the need to have proper studies that identify, fight and ameliorate its algorithms will be essential. According to many studies, bias has been identified in many AI(artificial intelligence, but from this point onwards we will refer to it with its acronym) entities, specifically in machine learning (MA), and it falls in different categories. The purpose of this paper is to provide an overall introduction of two categories of bias in AI and MA, concretely: racial bias and association bias(commonly known as gender bias), and then analyze the impact and risks that they have/might have in society and provide possible resolutions. The limitations of this paper are evident, as no empirical study has been conducted in itself but it is based on referential work conducted by other researchers.

Keywords: racial, gender, algorithm.

1. Introduction

"*Like all technologies before it, artificial intelligence will reflect the values of its creators.*" [1] The complexity of the world in which we live is increasing by the day and its acceleration is doing so as well. If we take a short journey back in time, how people lived in the 1880s didn't change *as drastically* with how people lived in 1910 *as* if we compare how people lived in the 1980s with how life was like in 2010. In just 30 years, the advent of the internet, personal computers and more recently the usage of artificial intelligence has made life back in the 1980s feel like centuries ago. When I observe today's middle school children as they watch a movie from the '80s, where there are fax machines and phones with cords, their reactions are impressive. I have heard children refer to Nokia 3310 as *ancient,* even though it was produced only 20 years ago. The semantics are interesting because they reveal our understanding of the world or lack of it thereof. If the near past due to technology feels like *forever* ago, it's even harder to imagine how much the future will change. The futurist Ray Kurzweil, in his essay 'The Law of Accelerating Returns', wrote: "We won't experience 100 years of progress in the 21st century — it will be more like 20,000 years of progress (at today's rate)" [2]. Thus, arguing our inability to comprehend how much the world will change in the future. Perhaps the challenge would not be only in imagining the future but rather accepting and internalizing all these upcoming changes and the most substantial one would be to take measures that it develops in a way that nurtures the fundamental values which we all agree upon, such as: fairness, equality and so on.

When people are questioned regarding their understanding of AI, according to a survey by Weber Shandwick and KRC Research, whose results were published in Harvard Business Review, which surveyed 2,100 consumers in an online survey encompassing five global markets (the U.S., Canada, the UK, China, and Brazil). According to their survey the knowledge that people had on AI varied a lot: " Two-thirds of those surveyed say they know something about AI, although only about two in 10 (18%) say that they know a lot. One-third acknowledged knowing nothing about AI. We found that by far the most common first impression of AI is "robots," as 22% of respondents said [3]."

The matter of fact is that AI is not that robot which you might have encountered in a Star Trek movie or The Matrix. It’s not something of the far future either. In reality, AI entities are being used in our daily lives without us being aware of it. Many AI entities are incorporated in the judicial system, for example, the software Correctional Offender Management Profiling For Alternative Sanctions (COMPAS) in the USA, utilizes AI in order to determine the release or not of an offender [4]. Furthermore, it’s used in facial recognition softwares and various scoring softwares to assist companies or states to make decisions in finance, jobs and insurance. However, the accuracy of AI doesn’t seem to reflect equally in these domains. By previous research it has been widely noted that bias in AI exists. Bias is defined as: *the action of supporting or opposing a particular person or thing in an unfair way, because of allowing personal opinions to influence your judgment.* (by Cambridge Online Dictionary) [5] Bias in AI is noted parcilucary in Machine Learning, consequently rendering in this manner the conclusion of the software as “unfair”[4]. *In the context of decision-making, for which AI is most predominantly used, fairness is the absence of any prejudice or favoritism toward an individual or a group based on their inherent or acquired characteristics* [6].

Of course, machines differently from people do not get bored or tired [7], so we would naturally expect them to be the most objective when it comes to decision-making. So, why is AI unfair and biased? What happens exactly and what categories of bias are shown? Facial recognition software which are embedded in most smart phones but also in other devices, have demonstrated to be more accurate on male and white people, [4] clearly demonstrating a bias in terms of judgement towards women and especially women of color. If you search for images in different search engines, such as “doctor” the results most likely would show a man and if you would alternatively search for a “nurse”, the results have a tendency to show women. Of course, women can be, and are, doctors and men are/can be nurses but the machine learning algorithm, fed by historical data, reinforces a cultural bias, which is named association bias or more commonly known as gender bias which we will see in part III.Furthermore, in the COMPAS software mentioned above, studies have shown that there is a higher score in terms of risk assessment for Afro-American offenders as compared to Caucasians (if other variables fall within the same profile) [4]. COMPAS assists judges to take their decisions on who to release, therefore, the impact of bias in such a software can be tremendous in a person’s life. This categorizes as *racial bias,* which we will see in section II.

2. Racial Bias in AI

We are all aware of the definition of racism as the belief that one race is more superior than another/others. Even after centuries of struggle, racism still takes place in our everyday lives, especially when it is institutionalized racism. This belief system leads people or institutions to make unfair and unequal decisions, hence to be biased. Another term, with which we may not be very familiar with, is *implicit racial bias*. “*It is important to distinguish implicit racial bias from racism or discrimination. Implicit biases are associations made by individuals in the unconscious state of mind. This means that the individual is likely not aware of the biased association*.” [9] These implicit racial biases that people have, very often get “fed” to Machine Learning algorithms under supervised learning and can have detrimental outcomes to society as a whole.

As in the example mentioned in the introduction, COMPAS, has racial bias embedded in itself, hence giving a non-objective result to the judge [4]. Another example would be Google's photo app, which gives automatic labels (inferring words, usually either adjectives or nouns) to images in digital photo albums, was actually labelling photos of black people as gorillas. Google made a public apology, claiming it unintentional. However, similar errors appeared in Hewlett-Packard's web camera software, which had a lot of trouble recognizing images of people with a darker skin tone and on Nikon's camera software,that labeled images/photos of Asian people as "blinking" [1].

From a technical perspective, when we consider the population (or more commonly the sample) in our statistical/machine-learning algorithm as homogenous and we do not take into account how heterogenous it might be in reality, then we, as scientists, might make fatal mistakes, when we advise people on their health, finances and so on. Let's take an example that actually demonstrates what bias might look like in AI.

The paper, A Survey on Bias and Fairness in Machine Learning, has taken a rning, consider a hypothetical nutrition study to demonstrate how the heterogeneities can bias data. This hypothetical study looks at how consuming pasta on a daily basis might impact body mass index (BMI) [6].

"*Regression analysis (solid red line) demonstrates a positive relationship in the population between the consumption of pasta and BMI. The positive trend suggests that increased pasta consumption is associated with higher BMI. However, unbeknown to researchers, the study population was heterogeneous, composed of subgroups that vary in their fitness level—people who did not exercise, people with normal activity levels, and athletes. When data is disaggregated by fitness level, the trends within each subgroup are negative (dashed green lines), leading to the conclusion that increased pasta consumption is, in fact, associated with a lower BMI. Recommendations for pasta consumption that come from the naive analysis are opposite to those coming from a more careful analysis that accounts for the difference between subgroups*". [6]

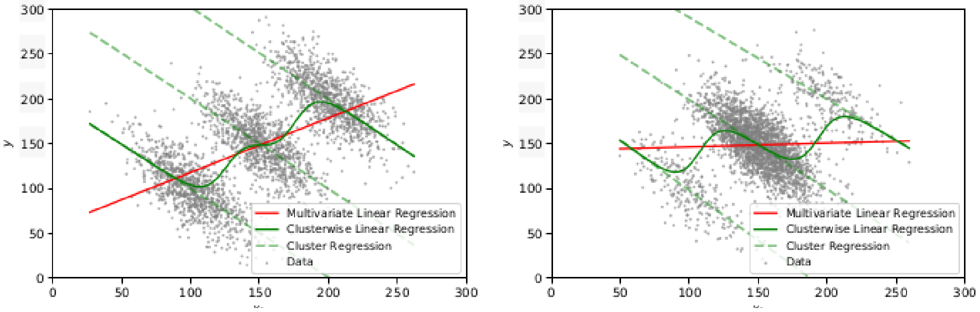


Fig. 1.
Source: Mehrabi et al.

Another example of this has been noted by ProPublica, regarding the AI algorithm in the judicial system, which estimates the probability of a criminal to commit another crime if they get out of prison. In the case study, there were the two cases of Brisha Borden and

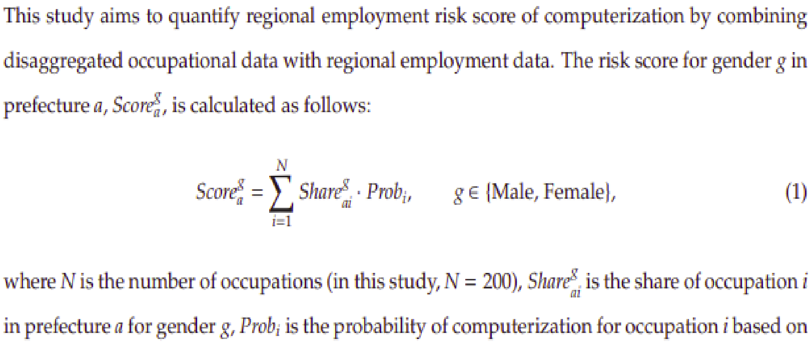
Vernon Prater. Broden stole a bike as she was waiting for her sister to come out of school and got arrested and charged with burglary and petty theft for the bike she stole, valued at an amount of $80. On the other hand, Prater (who was a 42 year old man with prior record) stole about the same value of items, so 80$, from a construction store. Even though Prater was a lifelong criminal, charged with armed robbery and many other crimes, when both individuals were booked in the prison system, something strange occured: Broden - who is black - was given a high score in terms of the risk of commiting a crime again whereas Prater - who is white - was considered by the algorithm as low risk. Two years later, Broden is not charged with any new crimes but Prater on the other hand got another 8-year long sentence. *The algorithm was wrong and racially biased* [4].

2.1. Gender Bias in AI

If automation and AI are not developed and applied in a gender-conscious manner, they are likely to reproduce or even reinforce already existing gender stereotypes and social norms which are discriminatory.

Examples:

- Virtual personal assistants such as Alexa, Siri and Cortana have female names and a default female voice. Companies which are behind these virtual assistants are reinforcing the social reality in which the majority of secretaries & personal assistants are women.
- Gender bias pervades AI algorithms as well. With close to 78% of AI professionals being men, the algorithm creation is informed and dominated by male experiences. This gender bias might have considerable adverse implications for women. Algorithms could affect women's access to different jobs or loans by automatically vetting out their applications or giving women an unfavourable rating. In addition, the algorithm-based risk assessment in the criminal justice systems can work against women if for example doesn't consider the fact that women are less likely than men to reoffend [11].
- The Robotization and automation of jobs will impact both men and women. But gender bias is likely to cause that automation impacts women disproportionately. For example, if more than 70% of workers in apparel manufacturing are women, automation will undoubtedly affect women more than men [11].



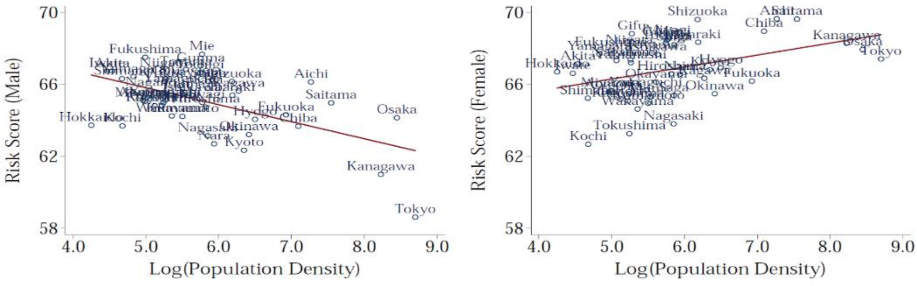
This study aims to quantify regional employment risk score of computerization by combining disaggregated occupational data with regional employment data. The risk score for gender g in prefecture a, $Score_a^g$, is calculated as follows:

$$Score_a^g = \sum_{i=1}^{N} Share_{ai}^g \cdot Prob_i, \qquad g \in \{\text{Male, Female}\}, \qquad (1)$$

where N is the number of occupations (in this study, $N = 200$), $Share_{ai}^g$ is the share of occupation i in prefecture a for gender g, $Prob_i$ is the probability of computerization for occupation i based on

Fig. 2.

Soure: Frey and Osborne (2017).



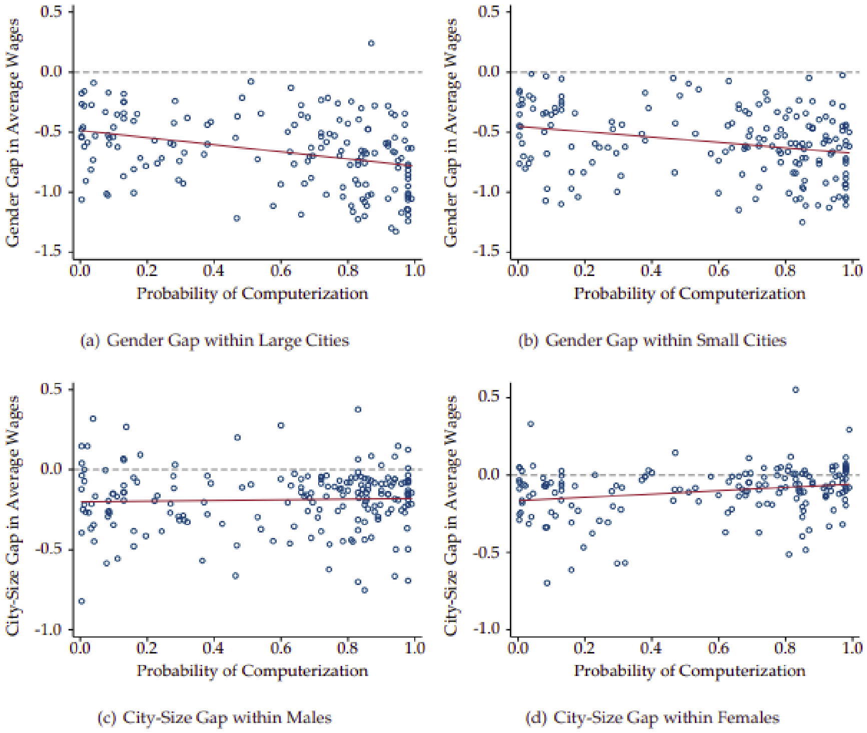
(a) Risk Score (Male)

(b) Risk Score (Female)

Fig. 3.

Reproduced from Figure 3 in Hamaguchi and Kondo (2018), Employment risk score of computerisation and city size [12].

As we can note from the formula above which is used to measure how employment will change with introduction of AI in Japan as a case study and its visualization in graphs it can be clearly noted that it impacts women and men very differently, by impacting women more especially in areas of large density of population such as Tokyo and Osaka.



(a) Gender Gap within Large Cities

(b) Gender Gap within Small Cities

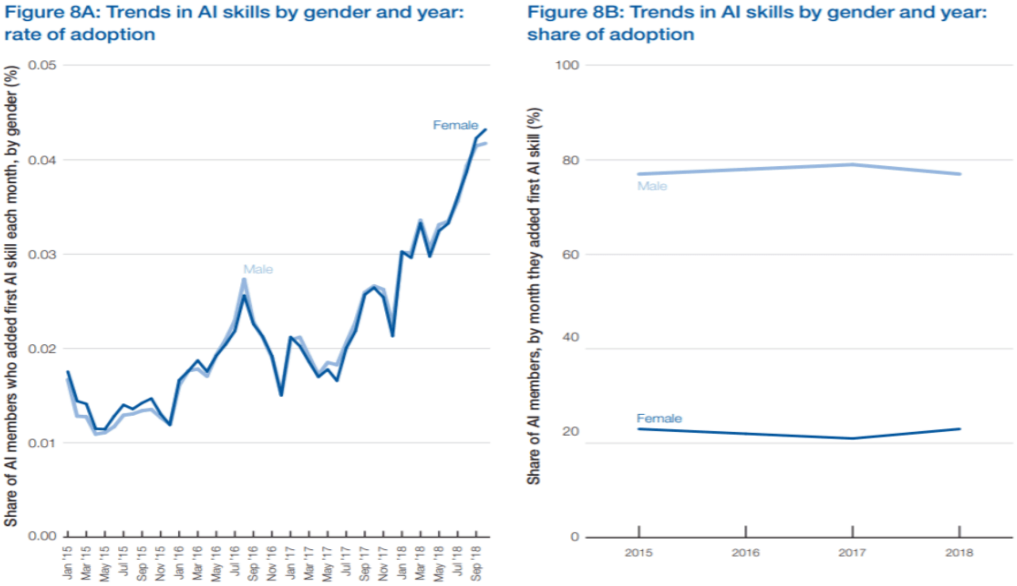
(c) City-Size Gap within Males

(d) City-Size Gap within Females

Fig. 4.

2.2. Potential causes:

1. Features & modeling techniques: The measurements used as inputs for models of machine-learning , or even the actual model training in itself, may introduce bias [10].
2. A skewed or incomplete training dataset: This happens when demographic categories are missing from the training data. Models developed with this data can then fail to scale properly when applied to new data containing those missing categories. For example, if female speakers make up a low percentage of your training data, let's say 12 percent, then when a trained machine learning model is applied to females, it will potentially produce a higher degree of errors [10].
3. The labels used in training: AI applications are generally trained using data that are generated by humans, and humans are inherently biased. Most commercial AI systems use supervised machine learning, labeling the training data in order to teach the selected model how to behave. Oftentimes humans create these labels and considering that frequently people manifest conscious & unconscious bias and the machine-learning models are trained to estimate these labels, the misclassification and unfairness towards a specific gender category will be encoded into the model, leading to bias [10].



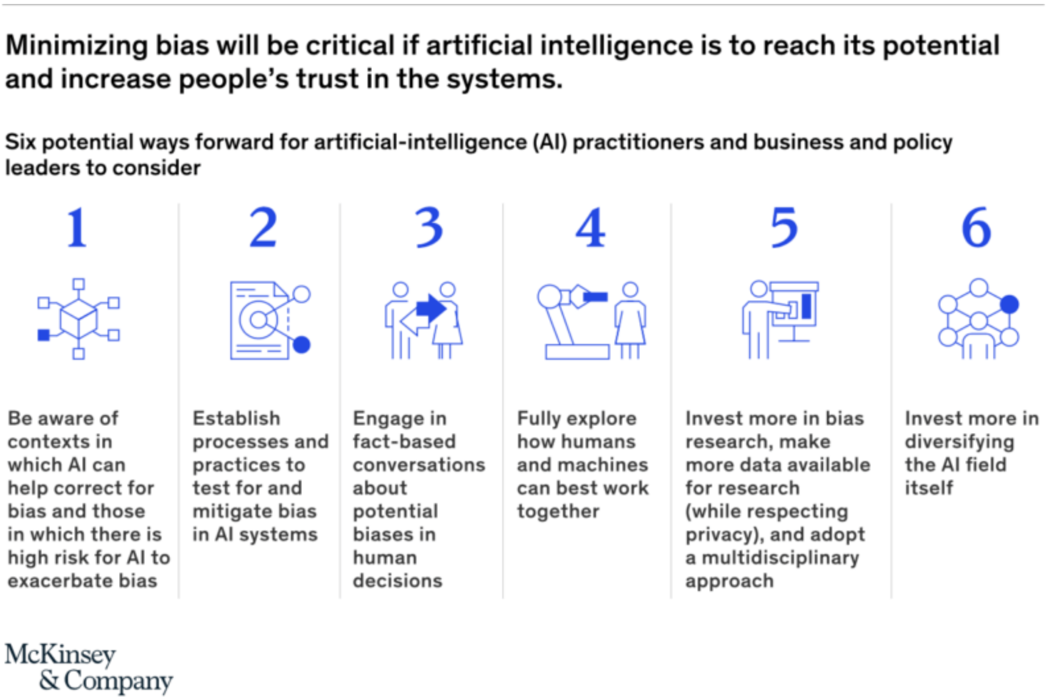
Source: LinkedIn.
Note: Adoption trends show how this pool of AI talent has grown over time, based on when members first indicated having an AI skill. To generate these trends, we first looked at the total number of members with AI skills and segmented this group by gender. We then identified the date when each member added the first AI skill to their profile and calculated this as a proportion of all members, by month, for each segment.

Fig. 5. The Global Gender Gap Report 2018 by World Economic Forum

3. Conclusions:

Like all new technologies, it takes time to come up with precision and accuracy. The matter of fact is, that unlike other technologies, AI has a very wide scope and can impact directly people's lives and the society's as a whole. In order to make sure that fairness is applied in the algorithms incorporated in AI entities, a few strategies have emerged [14]: Preprocessing data to make sure that the algorithm learns from a "clean" dataset. "*Counterfactual fairness"* has been coined as a term which tackles sensitive attributes such as gender and race. Silvia Chiappa's path-specific counterfactual method takes into consideration different ways that sensitive attributes impact outcomes—certain influences might be considered fair and could be retained by the algorithm, whereas other influences might be considered unfair and hence they should be discarded [15].

"*Innovative training techniques such as using transfer learning or decoupled classifiers for different groups have proven useful for reducing discrepancies in facial analysis technologies."* [14]



Minimizing bias will be critical if artificial intelligence is to reach its potential and increase people's trust in the systems.

Six potential ways forward for artificial-intelligence (AI) practitioners and business and policy leaders to consider

1 Be aware of contexts in which AI can help correct for bias and those in which there is high risk for AI to exacerbate bias

2 Establish processes and practices to test for and mitigate bias in AI systems

3 Engage in fact-based conversations about potential biases in human decisions

4 Fully explore how humans and machines can best work together

5 Invest more in bias research, make more data available for research (while respecting privacy), and adopt a multidisciplinary approach

6 Invest more in diversifying the AI field itself

McKinsey & Company

Fig. 6.
Source: MCKINSEY

References

[1] Crawford, K. C. (2016), Artificial intelligence's white guy problem, The New York Times, https://www.nytimes.com/2016/06/26/opinion/sunday/artificial-intelligences-white-guy-problem.html

[2] Kurzweil, R. K. (2001, March 7), *The Law of Accelerating Returns*, Https://Www.Kurzweilai.Net/. https://www.kurzweilai.net/the-law-of-accelerating-returns

[3] Gaines-Ross, L. G. R. (2016), What Do People — Not Techies, Not Companies — Think About Artificial Intelligence? Harvard Business Review, https://hbr.org/2016/10/what-do-people-not-techies-not-companies-think-about-artificial-intelligence

[4] Angwin, Larson, Mattu, Kirchner, J. A, J. L, S. M, L. K. (2016, May 23), Machine Bias. Https://Www.Propublica.Org/. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[5] Cambridge Dictionary. (2021), Https://Dictionary.Cambridge.Org/. https://dictionary.cambridge.org/dictionary/english/bias

[6] Mehrabi, Ninareh & Morstatter, Fred & Saxena, Nripsuta & Lerman, Kristina & Galstyan, Aram (2019), A Survey on Bias and Fairness in Machine Learning.

[7] Shai Danziger, Jonathan Levav, and Liora Avnaim-Pesso (2011), Extraneous factors in judicial decisions, Proceedings of the National Academy of Sciences 108, 17 (2011), 6889–6892.

[8] BUOLAMWINI, J. B. (2019), Artificial Intelligence Has a Problem With Gender and Racial Bias. Here's How to Solve It, TIME, https://time.com/5520558/artificial-intelligence-racial-gender-bias/

[9] Maryfield, B. M. (2018, December), Implicit Racial Bias. Justice Research and Statistics Association. https://www.jrsa.org/pubs/factsheets/jrsa-factsheet-implicit-racial-bias.pdf

[10] Josh Feast (2019), 4 Ways to Address Gender Bias in AI, Harvard Business Review 2009, https://hbr.org/2019/11/4-ways-to-address-gender-bias-in-ai

[11] Surya Deva (2020), Addressing the gender bias in artificial intelligence and automation, Open Global Rights, https://www.openglobalrights.org/addressing-gender-bias-in-artificial-intelligence-and-automation/

[12] Hamaguchi, N and K Kondo (2018), Regional employment and artificial intelligence in Japan, RIETI discussion paper 18-E-032.

[13] Lim, H. L. (2020, July 20), 7 Types of Data Bias in Machine Learning, Lionbridge.Ai. https://lionbridge.ai/articles/7-types-of-data-bias-in-machine-learning/

[14] Silberg, Manyik, J. S. J. M. (2019, June), Notes from the AI frontier: Tackling bias in AI (and in humans). Mckinsey. https://www.mckinsey.com/~/media/mckinsey/featured%20insights/artificial%20intelligence/tackling%20bias%20in%20artificial%20intelligence%20and%20in%20humans/mgi-tackling-bias-in-ai-june-2019.pdf

[15] Chiappa, S. (2019), Path-Specific Counterfactual Fairness, Proceedings of the AAAI Conference on Artificial Intelligence, 33, 7801–7808. https://doi.org/10.1609/aaai.v33i01.33017801

[16] Barocas, S., & Selbst, A. D. (2016), Big Data's Disparate Impact, SSRN Electronic Journal, 62. https://doi.org/10.2139/ssrn.2477899