IRIE
International Review of Information Ethics Vol. 32 (11/2022)

Author: Galit Wellner

Some Policy Recommendations to Fight Gender and Racial Biases in AI

Abstract:

Many solutions have been proposed to fight the problem of bias in AI. The paper arranges them into five categories: (a) "no gender or race" - ignoring and omitting any reference to gender and race from the dataset; (b) transparency - revealing the considerations that led the algorithm to reach a certain conclusion; (c) designing algorithms that are not biased; (d) "machine education" that complements "machine learning" by adding value sensitivity to the algorithm; or (e) involving humans in the process. The paper will selectively provide policy recommendations to promote the solutions of transparency (b) and human-in-the-loop (e). For transparency, the policy can be inspired by the measures implemented in the pharmaceutical industry for drug approval. To promote human-in-the-loop, the paper proposes an "ombudsman" mechanism that ensures the biases detected by the users are dealt with by the companies who develop and run the algorithms.

Keywords: Algorithms, Artificial Intelligence, Gender and Racial Bias, Machine Learning, Transparency

Agenda:

Authors:

Dr. Galit Wellner

- Tel Aviv University, Faculty of Humanities, Ramat Aviv, Tel Aviv, Israel, Email: galitwellner@tauex.tau.ac.il , Website: https://www.researchgate.net/profile/Galit-Wellner

ISSN 1614-1687

Introduction

The rise of AI algorithms has been accompanied by the belief that these systems are impartial and do not suffer from the biases that humans and previous technologies express. It becomes evident, however, that gender and racial biases can (and practically do) exist in AI algorithms. The question is where the bias is rooted – in the training dataset or the algorithm? Is it a linguistic issue (and hence related to the dataset) or a broader societal current (and hence can be also in the algorithm itself)?

Some computer and data scientists assert that the biases are rooted in the dataset (Caliskan, Bryson, and Narayanan 2017). In an article published in *Science* Magazine, Caliskan et al. demonstrate how the mechanism that identifies flowers as pleasant and insects as unpleasant, invokes the same dualism for European-American names vs. African-American names. In the context of gender bias, this type of algorithm matches male-related words and names to career, mathematics, and science, whereas female-related words and names correspond to family and arts. The authors claim that the

> *"substantive importance of [the] results raise[s] the possibility that all implicit human biases are reflected in the statistical properties of language" (p.185).*

An alternative explanation for the biases examines the algorithms. In her book *Weapons of Math Destruction*, computer scientist Cathy O'Neal claims that even the mere logic of an algorithm might provoke a bias, no matter what data are used (O'Neal 2016). Thus, data and algorithms are the two poles between which bias can be found.

Many solutions have been proposed to the problem of biases in AI. They can be mapped into five categories (Wellner and Rothman 2020; Wellner 2020; 2021): (a) "no gender or race" - ignoring and omitting any reference to gender and race from the dataset; (b) transparency - revealing the considerations that led the algorithm to reach a certain conclusion; (c) designing algorithms that are not biased; (d) "machine education" that complements "machine learning" by adding value sensitivity to the algorithm; or (e) involving humans in the process. These categories will be described in the first section of the article.

The second section is devoted to policy measures corresponding to the solutions. Due to space constraints, only two policy recommendations will be discussed, regarding (b) transparency and (e) human-in-the-loop. Although type (b) solutions (transparency) are complicated to implement due to the current regime of intellectual property as well as cultural and business considerations, such solutions can lead to policy guidelines similar to those of the pharmaceutical industry in the process of drug approval. Type (e) solutions ("human in the loop") can be implemented at several points in the lifecycle of an AI system: starting from the development stage, and ending at the usage stage in which an "ombudsman" mechanism can be applied to ensure that the biases detected by the users are dealt with and overcome by the companies who develop and run the algorithms.

Solutions to AI Bias

Many authors suggested many solutions, design tips, and recommendations on how to avoid or bypass the algorithmic gender bias. In this section, the various solutions are mapped into five groups, according to their target – the human user, the algorithm, or the dataset. The following review of solutions starts from the dataset, moves to the algorithm, and ends with the human user (Wellner and Rothman 2020; Wellner 2020; 2021).

Dataset: No gender, please!

The first solution is simply avoiding a reference to the gender in the dataset. It is termed "extrinsic bias" because the data are considered external to the AI system (Zerilli et al. 2019). However, it was already shown

that AI algorithms can deduce gender and race from other data items. For example, omitting the race led the algorithm to consider the address, so that neighbourhoods became signifiers of the race of most inhabitants.

A variation of this type of solution is offered by Luca and Fisman (2016) to avoid gender and racial biases in platforms like Uber or AirBnB in which female service providers receive fewer orders and lower income. They propose to target gender and race issues in the design phase and avoid providing "too much information." They suggest that service platforms should show pictures of hosts only after the transaction is confirmed. Although the dataset includes pictures that may reflect gender and race, the data are provided when it does not affect the potentially biased decision thereby mitigating the risk of bias. This solution does not eliminate the indirect gender reference and instead exposes this risk only at later stages.

Algorithms: Transparency

From datasets to algorithms, a commonly suggested solution calls for algorithms' transparency. Although frequently the demand for transparency is translated into a demand to reveal the code, this is not a must. On the contrary, the code is unlikely to reveal biases and attention should be paid to the procedures before and after the development. The underlying assumption is that if we know how the algorithm concluded then we can detect the bias (Zerilli et al. 2019). For instance, IBM recommends that AI systems display which factors weighted the decision in one direction vs. another; the confidence in the recommendation; and the factors behind that confidence (Lomas 2018).

Another aspect of transparency relates to how the data were collected and annotated. This solution requires a new profession, "data curators" who would supply "nutrition labels" for AI training datasets (Zou and Schiebinger 2018, p. 325–26); see also (D'Ignazio and Klein 2020)).

O'Neil calls for more radical transparency. She advises that the developers should reveal their guiding logic and expose the choices they made from the beginning of the research like, what were the parameters chosen to decide who the most underperforming teachers are in public education systems. Pupils' grades, she shows, cannot be taken as the sole parameter, as it leads to constant laying off of teachers working in the lowest ranking schools, even if the teachers did a great job. Those teachers are eventually hired by private schools, and the public systems remain weak.

Algorithms: Anti-bias Algorithms

The third kind of solution also deals with algorithms. It focuses on how the algorithms calculate and decide what Zerilli et al call "intrinsic bias." These solutions call upon developers to improve their algorithms so that the system can bypass the "built-in" biases of the datasets (see Zou and Schiebinger 2018). For example, a few years ago Gmail's then-new auto-complete feature turned to be biased. When a user typed "I am meeting an investor next week," the algorithm suggested a possible follow-up question: "Do you want to meet him?" and neglected to suggest "her" (Dave 2018). Eventually, Gmail has decided to "mute" this feature thereby bypassing the problem. Now the algorithm does not make any suggestions. Gmail's solution does not deal directly with the way the algorithm treats gender but rather tweaks the output side. Alternatives might be a modification of the algorithm so that it suggests also "her" or implementing a transparency tactic (see section 2.2 above) according to which the system presents several options.

One variation of this kind of solution requires that the developers define in advance which bias they wish to avoid. Another variation recommends the developers identify "sensitive attributes" and avoid dependence on such attributes (Nallur 2020). In this spirit, Fisman and Luca suggest avoiding discrimination in online platforms by asking: "Should your algorithms be discrimination-aware?" (Fisman and Luca 2016, p.95) They call upon system designers to think of the user experience through the lenses of possible biases, because:

> *"Thus far many algorithm designers have ignored factors such as race and gender and just hoped for the best. But in many cases the probability that an algorithm will unintentionally achieve equality is ssentially zero".*

This strategy should involve self-scrutiny practices of the companies so that they "proactively monitor and respond to such problems" (ibid).

These proposed variations mostly work *ex-post factum*, after the systems are running, and after some damages are caused. Put differently, biases and sensitivities are usually revealed after the system is operative.

Algorithms: Machine Education

The concept of "machine learning" is based on a certain mode of teaching, where a human developer (or a data scientist) serves as a "teacher" to the algorithm by warning the algorithms during the training stage from sensitive parameters. The idea of education builds on a different logic that is based on self-reflection and forecasting several possible scenarios for different societal settings.

In the context of algorithms, education can be implemented in supervising software that can identify biases and then alter the dataset (Zou and Schiebinger), thereby serving as a "moral compass" to the system. Such an algorithm can be implemented in GAN architecture, as a feature in an existing algorithm, or as an independent system. The challenge is to identify the bias, and like any ethical issue, it is hard to code it into a binary set of algorithmic decisions (Wellner and Rothman 2020); (Nallur 2020).

Humans: Human in the loop

After dealing with datasets and algorithms, we move on to the human element. The last kind of solution focuses on the cooperation between humans and algorithms. Human involvement is necessary because algorithms cannot refer to abstract ideas like gender or race (Marcus 2018). A technique like deep learning can recognize patterns in huge amounts of data, but the reasoning and meaning extraction can be done only by humans. Therefore, the practical solution should combine both – algorithms to identify the pattern and humans to understand its meaning.

Human involvement is of utmost importance in the development stages when the problem is defined and data sources are selected. But human involvement can be no less crucial in the usage phase, for the detection of bias in real-life scenarios. Thus, users should develop certain skills for the usage of AI, similarly to the "digital literacy" campaigns of the early personal computer days (and later the Internet age). The "AI literacy" can include awareness of potential biases and methods to reveal them, as well as finding ways to select other options, report, or bypass them.

Policy Recommendations

The various solutions as mapped into the five categories above can be implemented by developers as well as by regulators. In both cases, creativity is required to translate the principles into real-life applicable frameworks. Due to limited space, the recommendations for policymakers shall be focused on two promising categories – (b) transparency and (e) human-in-the-loop. The first represents a solution that has already been successfully implemented in other technological domains and the latter represents an updated version of an old solution to a similar problem, though outside the technology domain.

Policy and transparency

Type (b) solutions that promote transparency are complicated to implement due to the current regime of intellectual property according to which trade secrets must be kept secret. This line of thinking is frequently invoked in response to calls for transparency that are focused on revealing the source code or any part of the development process.

An EU-funded project named SHERPA[1] suggested solving the barrier of transparency by "borrowing" from the health sector an FDA-like mechanism. When a pharmaceutical company wishes to receive regulatory approvals for a drug it developed, the company discloses to the FDA how the drug was developed and any related information. The company maintains this information under strict confidentiality and so does the FDA. In practice, the FDA acts as a proxy of the patients and assesses the information provided by the company to ensure that the drug is not toxic and that it can cure a certain disease. Likewise, a future AI regulator can examine algorithms and datasets under strict confidentiality to ensure that they are not biased. Implementing an FDA-like regulation implies that we treat algorithms like drugs – they can cure but if unregulated can harm. This mechanism ensures some form of transparency and serves as a major building block in a future approval regime for AI algorithms.

Policy and human involvement

Type (e) solutions ("human in the loop") can be implemented throughout the lifecycle of an AI system starting from the early stages in which the algorithm is developed and trained, and ending at the usage stage. In this article, I focus on the latter and examine the frustrating situation in which users identify a bias, but their complaints are not heard. Often large companies operating AI algorithms provide no more than automatic answers according to predefined scenarios. Identifying a malfunctioning, such as gender or racial bias, is not among the scenarios those companies are handling. Today users can tell their stories on media – from traditional media outlets to social media, hoping to be heard by the public thereby creating some pressure on the company responsible for the biased algorithm. They can also go to court, but this procedure is expensive and time-consuming.

The following policy suggestion relies on the affinity between AI algorithms and bureaucratic structures. Adam Clair explains:

> *"Like algorithms, bureaucratic processes are built on the assumption that individual human judgment is too limited, subjective, and unreliable, deficiencies that lead to nepotism, prejudice, and inefficiency." (Clair 2017).*

To solve these drawbacks of bureaucracy, democratic regimes have promoted concepts of transparency, as well as the "ombudsman" mechanism. An ombudsman is an independent official who can investigate complaints against the company that operates the algorithm. It can be a governmental entity that can demand answers in case of injustice, and sometimes can even initiate an investigation. Alternatively, it can be a position in every company above a certain size. Whereas lack of transparency might end up in court, the "ombudsman" is more accessible and frequently faster.

This proposed policy is intended to empower the users by delegating some investigative power to an entity to which the companies must reply coherently and within a reasonable time frame.

[1] The SHERPA Project (https://www.project-sherpa.eu) was funded by the EU to analyse how AI and big data analytics impact ethics and human rights. The project involved dialogues with various stakeholders in order to develop novel ways to understand and address these challenges and to find desirable and sustainable solutions that can benefit both innovators and society.

To sum up, the proposed regulatory solutions offered in this article attempt to re-use existing mechanisms and adjust them to the fast-changing landscape of AI. In both suggestions, cooperation is required from the companies who develop and operate these algorithms. Such cooperation can be based on their wish to promote ethical algorithms ("soft regulation") and can be dictated by states ("hard regulation"). If the former does not come to fruition, the latter should be sought and promoted.

References

Caliskan, Aylin, Joanna J. Bryson, and Arvind Narayanan. 2017. "Semantics Derived Automatically from Language Corpora Contain Human-like Biases." *Science* 356 (6334): 183–86. https://doi.org/10.1126/science.aal4230.

Clair, Adam. 2017. "Rule by Nobody: Algorithms Update Bureaucracy's Long-Standing Strategy for Evasion." *Real Life*, 2017. https://reallifemag.com/rule-by-nobody/.

D'Ignazio, Catherine, and Lauren F. Klein. 2020. *Data Feminism*. *Data Feminism*. The MIT Press. https://doi.org/10.7551/mitpress/11805.001.0001.

Dave, Paresh. 2018. "Fearful of Bias, Google Blocks Gender-Based Pronouns from New AI Tool | Reuters." Reuters. 2018. https://www.reuters.com/article/us-alphabet-google-ai-gender/fearful-of-bias-google-blocks-gender-based-pronouns-from-new-ai-tool-idUSKCN1NW0EF.

Fisman, Ray, and Michael Luca. 2016. "Fixing Discrimination in Online Marketplaces." *Harvard Business Review* 2016 (December): 89–96.

Lomas, Natasha. 2018. "IBM Launches Cloud Tool to Detect AI Bias and Explain Automated Decisions." *TechCrunch*, 2018. https://techcrunch.com/2018/09/19/ibm-launches-cloud-tool-to-detect-ai-bias-and-explain-automated-decisions/.

Marcus, Gary. 2018. "The Deepest Problem with Deep Learning." Medium. 2018. https://medium.com/@GaryMarcus/the-deepest-problem-with-deep-learning-91c5991f5695.

Nallur, Vivek. 2020. "Landscape of Machine Implemented Ethics." *Science and Engineering Ethics* 26 (5): 2381–99. https://doi.org/10.1007/s11948-020-00236-y.

O'Neal, Cathy. 2016. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. New York: Crown Publishing Group.

Wellner, Galit. 2020. "When AI Is Gender-Biased: The Effects of Biased AI on the Everyday Experiences of Women." *Humana.Mente* 13 (37): 127–50.

———. 2021. "I-Algorithm-Dataset: Mapping the Solutions to Gender Bias in AI." In *Gendered Configurations of Humans and Machines: Interdisciplinary Contributions*, edited by Jan Büssers, Anja Faulhaber, Myriam Raboldt, and Rebecca Wiesner, 79–97. Verlag Barbara Budrih.

Wellner, Galit, and Tiran Rothman. 2020. "Feminist AI: Can We Expect Our AI Systems to Become Feminist?" *Philosophy and Technology* 33 (2): 191–205. https://doi.org/10.1007/s13347-019-00352-z.

Zerilli, John, Alistair Knott, James Maclaurin, and Colin Gavaghan. 2019. "Transparency in Algorithmic and Human Decision-Making: Is There a Double Standard?" *Philosophy and Technology* 32 (4): 661–83. https://doi.org/10.1007/s13347-018-0330-6.

Zou, James, and Londa Schiebinger. 2018. "Design AI so That Its Fair." *Nature* 559 (7714): 324–26. https://doi.org/10.1038/d41586-018-05707-8.