INTERDISCIPLINARY
RESEARCH CENTRE
for Technology, Work and Cultur

IFZ - Electronic Working Papers

Fair AI - (how) does it work?

English Version

Julian Anslinger

02 - 2021

RESEARCHING FOR CHANGE

IFZ - Electronic Working Papers
ISSN: 2077-3102

Edited by
IFZ - Interdisziplinäres Forschungszentrum für Technik, Arbeit und Kultur
Schlögelgasse 2
8010 Graz

Tel: +43/316/813909-0
E-mail: office@ifz.at;
Web: https://www.ifz.at

Funding

The paper is based on information gained in the project VEKIAA - Responsible Integration of AI Assistance Systems in the Workplace, funded by the Digitalisation Fund Work 4.0 of the Vienna Chamber of Labour and commissioned by Graz University of Technology.

The province of Styria and the city of Graz support the IFZ with funding.

The term Artificial Intelligence

The discourse on fair AI is characterized by a variety of different terms, ideas and approaches. This begins with the term "artificial intelligence", a buzzword that subsumes many different computer science methods (machine learning, deep learning, inductive programming, Bayesian estimation, search and optimization methods, etc.). The term artificial intelligence has little to do with the concept of intelligence that we associate with humans or animals. AI systems are created for very specific applications and then function only for that particular purpose; broad "Artificial General Intelligence" or "strong AI" remains in the realm of utopia. With this in mind, some scholars* advocate using other terms such as "algorithmic decision systems" (Schröter, 2019) or the exact method name.

What is fair AI?

Terms such as fair AI, ethical AI, trustworthy AI, or responsible AI are often used interchangeable. In a qualitative content analysis of 84 ethical guidelines, Anna Jobin and colleagues (2019) found the following frequently discussed aspects of fair AI: justice, transparency, harmlessness, responsibility and accountability, privacy, utility, freedom and autonomy, trust, sustainability, dignity, and solidarity. These are also reflected in a similar form in the Ethical Guidelines for Trustworthy AI of the expert group set up by the European Commission, in which the following core demands are made: Primacy of Human Action and Human Oversight, Technical Robustness and Security, Privacy and Data Quality Management, Transparency, Diversity, Non-Discrimination and Fairness, Social and Environmental Wellbeing, and Accountability. The European Commission's proposed regulation for AI systems, which was published in April 2021 and is expected to enter into force in an adapted form in 2024 at the earliest, addresses many of these aspects to varying degrees. In any case, the planned regulation is expected to have a similar impact on the entire world as the General Data Protection Regulation. At least until then, it is important to advance the discourse on fair AI and talk about the many ways in which AI can have a negative impact on society and the environment. Central to this is the topic of biases.

Biases

Three types of biases seem to play a particularly significant role: Data-based biases, algorithm-based biases, and culture-based biases.

1. Data-based biases: a central data-based bias is the so-called sample bias, which arises when the data underlying an algorithm do not reflect the (relevant) reality. For example, image recognition algorithms that had problems correctly identifying dark-skinned women as such while correctly classifying the faces of men or white people more often have become known (e.g., Buolamwini & Gebru, 2018; Zou & Schiebinger, 2018). In turn, another image recognition algorithm misidentified men pictured in kitchens as women (Zhao et al., 2017). In both cases, poor data is to blame. In the first example, the dataset used for training contained mostly photos of white people and men; in the second example, cooking men were less represented than cooking women in the underlying dataset.

2. Algorithm-based biases: Algorithm-based biases can occur, among other things, when users are not sufficiently involved in the development of the software. For example, it is relevant to pay attention to how the results of AI algorithms are presented in such a way that they can be usefully and correctly interpreted. Research has demonstrated, for example, that visualizations can be perceived more easily and quickly than data tables. On the other hand, however, it has been shown that visualizations can prevent people from engaging more deeply with the object being viewed. For example, in a study by the University of Michigan and Microsoft, it was shown that even data scientists used visualizations of AI results for interpretation, even if they were not fully comprehensible to them (Kaur, 2020). Moreover, the sometimes incomprehensible visualizations nevertheless increased the experts' confidence in the machine-learning algorithm. The fact that people mistakenly perceive results presented by computers as particularly objective and correct is a phenomenon described as automation bias (Mehrabi, 2021). In summary, careful consideration must also be given to programming algorithms and, at best, how the algorithms and their presentation of results might be affected.

3. Culture-based biases: imagine an AI algorithm in which the data is selected to perfectly represent reality and, furthermore, the algorithm is programmed without any algorithmic biases. A perfect non-biased AI, right? No, not necessarily. In some cases of biased AI systems, it may happen that the bias does not result from data or algorithms, but exists in society and has been reproduced in data or algorithm

   An example of this is an AI algorithm used by Amazon from 2014 to 2015 to evaluate resumes of job applicants* (Dastin, 2018). To train the algorithm, historical recorded cv data and their success rates of previous applicants were used. To rule out discriminatory effects from the beginning, the algorithm was set to exclude gender as a feature. However, the machine learning program found an indirect and unanticipated way to distinguish between male and female applicants. Specifically, it found that successful applications for tech jobs were mostly submitted by people who did not have attributes associated with femininity. For example, the mention of the word woman (as in women's sports club) automatically resulted in a lower score for the respective resumes. The AI thus perpetuated a phenomenon anchored in Western society, the fact that comparatively few women are employed in technical professions. Women applicants may not have been given the chance to prove their technical competence and were excluded from the outset. In other words, they were discriminated against.

   Another example of the perpetuation of culturally existing inequalities among different groups is an algorithm programmed by the Austrian Public Employment Service (AMS). The algorithm, which the AMS calls the "Labor Market Opportunities Assistance System" (AMAS), is intended to evaluate job opportunities and support the assessment of job seekers' eligibility for assistance by employment agencies. Persons with high and low job chances, calculated by the algorithm, are to be supported less

than persons with medium job chances. Although the algorithm is not based on procedures that would allow it to be called artificial intelligence, it is suitable as an illustrative example of culture-based biases of algorithms, also because of its potentially immense influence. A study commissioned by AK Upper Austria and conducted by the Institute of Technology Assessment examined AMAS (Allhutter, 2020). The study was able to show that AMAS rates individuals differently in their job opportunities based on various person characteristics such as gender, age, immigrant background, and disability. The assessment is based on historically recorded data showing that individuals with certain characteristics or combinations of characteristics are more likely to have found a job than others. For example, AMAS rates the job opportunities of women with the characteristic child care responsibilities lower than those of men with child care responsibilities. This difference in job opportunities may indeed be so due to various realities in society (e.g., due to discriminatory hiring practices; see, e.g., Hipp, 2019), but it should not actually have a negative impact on job seekers' promotion recommendations. A similar downgrading in the AMAS rating system is found for women with an immigration background. Instead of providing special support to these groups of people in order to better integrate them into the labor market and achieve long-term equality, the AMS chooses to discriminate against these groups and perpetuate socially existing inequalities.

Preventing biases

How can biases and other negative effects of AI systems in companies be prevented? Here is a brief and naturally incomplete overview of important measures.

- The development of an AI system should never be completed. Possible biases and further effects must be considered from the beginning and taken into account during the process. This begins with business interest and the question of what AI systems and data are actually needed; continues through the technology development process (i.e., algorithm development, training, and testing, implementation), to implementation and long-term monitoring.
- Data underlying training should reflect relevant reality as much as possible. If circumstances change or the relevance of previously unconsidered variables becomes apparent, an algorithm should be retrained.
- If there is a risk of perpetuating socially existing inequalities, algorithmic corrective measures should be taken. For example, different weights could be applied to data to correct for existing biases in the training data.
- AI systems should be designed to be as transparent as possible in order to detect possible biases during operation or to be able to reconstruct them in retrospect. The annotation of the training data should follow a precisely defined standard and, if possible, be carried out in parallel by several people so that the agreement of the annotating persons can be calculated.
- Ethical and gender and diversity issues should be considered during AI programming. Good tips for an approach in various scientific fields can be found, for example, at: https://genderedinnovations.stanford.edu

- "I-Methodology" (i.e. the inscription of almost exclusively own experiences by technology developers*) should be prevented. For this, in addition to user-centered design, an inter- and transdisciplinary way of working in teams that are as diverse as possible is worthwhile. Diversity, inter- and transdisciplinarity increases the robustness of technologies and scientific knowledge.
- The needs and characteristics of all potential stakeholders should always be kept in mind. For this purpose, it is worthwhile to bring those affected to the same table right from the start and weigh interests and needs in workshops, for example. This process is wonderfully described in the WeBuildAi framework by Min Kyung Lee (2019).
- In companies, employee representatives should always be in direct dialog with corporate IT in order to make the needs of employees clearly heard. One possible approach is described in the publication How to make it fAIr - Methods of participatory technology design for the application field of artificial intelligence by the project dAIalog.at. In addition, a handbook will be published in the VEKIAA project (Responsible Integration of AI Assistance Systems in the Workplace) in summer 2022, which will go into this process in greater depth.

Literatur

Doris Allhutter et al. (2020). Der AMS-Algorithmus. Eine Soziotechnische Analyse des Arbeitsmarktchancen-Assistenz-Systems (AMAS). url: http://epub.oeaw.ac.at/0xc1aa5576_0x003bfdf3.pdf.

Joy Buolamwini & Timnit Gebru (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In: Conference on fairness, accountability and transparency. PMLR. S. 77-91.

Jeffrey Dastin (2018). Amazon scraps secret AI recruiting tool that showed bias against women. url: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/a...;

Lena Hipp (2019). Do Hiring Practices Penalize Women and Benefit Men for Having Children? Experimental Evidence from Germany. In: European Sociological Review 36.2 (Nov. 2019), S. 25-{264. issn: 0266-7215. doi: 10.1093/esr/jcz056. eprint: https://academic.oup.com/esr/article-pdf/36/2/250/33018077/jcz056.pdf. url: https://doi.org/10.1093/esr/jcz056.

Anna Jobin, Marcello Ienca, & Effy Vayena (2019). The global landscape of AI ethics guidelines. In: Nature Machine Intelligence 1, Nr. 9: 389–99, https://doi.org/10.1038/s42256-019-0088-2.

Harmanpreet Kaur et al. (2020). Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning. In: CHI Conference on Human Factors in Computing Systems Proceedings, 1–14, https://doi.org/10.1145/3313831.3376219.

Min Kyung Lee et al. (2019). ACM Hum.-Comput. Interact. 3, CSCW, Article 181 (November 2019), 35 pages. DOI:https://doi.org/10.1145/3359283

Ninareh Mehrabi et al. (2021). A survey on bias and fairness in machine learning. In: ACM Computing Surveys (CSUR) 54.6, S. 1-35.

Ulf Mellström (2009). The intersection of gender, race and cultural boundaries, or why is computer science in Malaysia dominated by women? In: Social studies of science 39.6, S. 885-907.

David Poole & Alan Mackworth & Randy Goebel. (1998). Computational Intelligence: A Logical Approach.

Welf Schröter (2019). Der mitbestimmte Algorithmus. Hg.: Welf, Schröter. Talheimer Verlag.

Anita Thaler et al. (2021). How to make it fAIr – Methoden partizipativer Technikgestaltung für das Anwendungsfeld der Künstlichen Intelligenz. https://daialog.at/wp-content/uploads/2021/09/dAIalog.at_-Methoden-part...;

Jieyu Zhao et al. (2017). Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. doi: 10.18653/v1/D17-1323.

James Zou & Londa Schiebinger (2018). AI can be sexist and racist | it's time to make it fair. In: Nature 559.7714, S. 324-326.