AI4media
ARTIFICIAL INTELLIGENCE FOR
THE MEDIA AND SOCIETY

ROADMAP ON AI TECHNOLOGIES & APPLICATIONS FOR THE MEDIA INDUSTRY

SECTION: "AI FAIRNESS"

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 951911

info@ai4media.eu www.ai4media.eu

Author	**Samuel C. Hoffman** (IBM Yorktown Heights)

This report is part of the deliverable D2.3 - "*AI technologies and applications in media: State of Play, Foresight, and Research Directions*" of the AI4Media project.

You can site this report as follows:

F. Tsalakanidou et al., Deliverable 2.3 - AI technologies and applications in media: State of play, foresight, and research directions, AI4Media Project (Grant Agreement No 951911), 4 March 2022

This report was supported by European Union´s Horizon 2020 research and innovation programme under grant number 951911 - AI4Media (A European Excellence Centre for Media, Society and Democracy).

The information and views set out in this report are those of the author(s) and do not necessarily reflect the official opinion of the European Union. Neither the European Union institutions and bodies nor any person acting on their behalf.

Copyright

AI fairness

Current status

Machine learning has the potential to bring about enormous change for good in the world. Already we see mature applications in language translation, medicine, image and speech recognition, and more. These technologies have the chance to democratise their benefits and reach incredible scales. But, as with any other powerful tool, machine learning can also serve to exacerbate inequalities and make historically powerful groups more powerful. This is the crux of the field of ***algorithmic fairness***: how to ensure the fruits of AI are shared equitably and discrimination is prevented.

Besides the obvious ethical implications, there are many reasons for machine learning application designers to consider fairness. In many fields, there are legal protections for disadvantaged groups such as equal employment, housing, and banking. In other cases, user trust can be eroded if a company is shown to have discriminated, even after the problems are fixed. Conversely, transparency around fairness can be a competitive advantage.

In recent years, many advancements in AI have been shown to demonstrate ***biases***. Many facial recognition technologies have been shown to perform significantly worse on dark-skinned faces[1] (see Figure 1). Healthcare applications have been shown to result in poorer outcomes for minority patients[2]. AI tools for job recruiting could have discriminated against women applicants[3]. Criminal recidivism models led to harsher penalties for black defendants[4]. In all of these cases, discrimination was almost certainly not the explicit goal of the model's producer but rather, a side effect of inattention to the potential harms.

So how do these models come to espouse biases if they are not intentionally designed that way? Unlike early rule-based AI models, machine learning and especially deep learning study very large sets of data in order to discover patterns in the relationship between inputs and outputs. Therefore, if a certain class of people are not well represented in the data, either in total number or proportional outcomes, a machine learning model may generalise poorly to real users. Furthermore, if the data used to train a model contains historical instances of bias from the real world, these tendencies can easily transfer to the model's outputs. Human decision-making processes can be flawed in the same ways, however, even if the model happens to be no worse than humans, the speed, scale, and often lack of accountability (the removal of humans from the process) of AI means the overall effect will be more harmful.

[1] Buolamwini, J., & Gebru, T. (2018, January). Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency* (pp. 77-91). PMLR.

[2] Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Science, 366(6464), 447-453.

[3] Dastin, J. (2018, October 10). Amazon scraps secret AI recruiting tool that showed bias against women. Reuters. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

[4] Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016, May 23). Machine bias. ProPublica. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

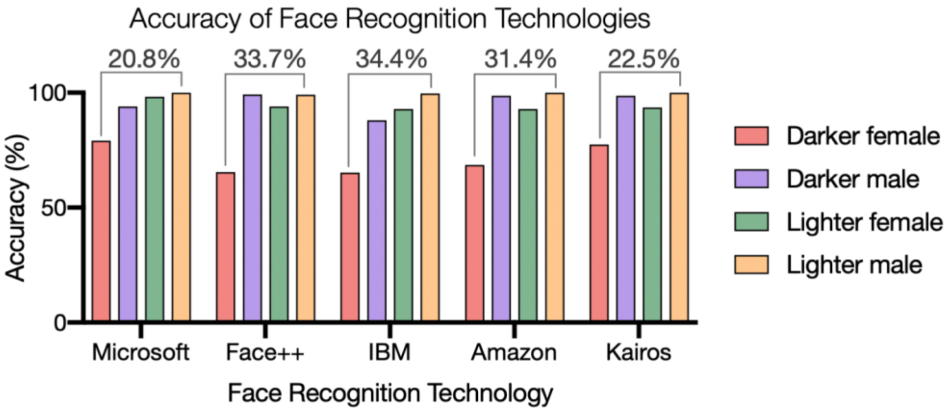

Figure 1: Results of the Gender Shades[5] project which audited various facial recognition models. This analysis prompted the targeted companies to improve their error rates on darker-skinned females by up to 30% within 7 months[6] (Image source: A. Najibi's blog post[7]).

But how can we tell if a machine learning model will have bias or compare the bias of different models? There is no one answer to this question since the relative value of tradeoffs varies based on the application. From an *individual* perspective, similar people should be treated similarly (measuring similarity between people is yet another tricky task, however)[8]. Likewise, changing uninformative aspects such as race or religion (a *counterfactual* example) should not result in different outcomes. From a *group* perspective, people from different sections of the population – grouped by a sensitive attribute such as gender, race, disability, belief, or age – should, on average, receive equal results. In some cases, the *procedure* or treatment of all individuals must be equivalent. These are all compounded by the fact that the real world often contains institutional and systematic disadvantages so in order to correct for these, measuring performance (e.g. accuracy) itself must take fairness into account.

It is easy to see how some of these definitions can be in conflict with each other. For instance, in a medical setting, following the same procedure for men and women would be inadvisable in many situations and yet we should still expect tests and surgeons to perform equally well on both. Less obvious, though, is the fact that even different group-based metrics are often impossible to satisfy simultaneously[9,10]. Furthermore, Simpson's paradox[11] tells us we must carefully choose the groups of interest since bias can be hidden by grouping data differently.

[5] Gender Shades: http://gendershades.org/overview.html

[6] Raji, I. D., & Buolamwini, J. (2019). Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial AI products. In Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (pp. 429-435).

[7] A. Najibi, Racial Discrimination in Face Recognition Technology (2020): https://sitn.hms.harvard.edu/flash/2020/racial-discrimination-in-face-recognition-technology/

[8] Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012, January). Fairness through awareness. In Proceedings of the 3rd innovations in theoretical computer science conference (pp. 214-226).

[9] Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. Big data, 5(2), 153-163.

[10] Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016). Inherent trade-offs in the fair determination of risk scores. arXiv preprint arXiv:1609.05807.

[11] Wikipedia, Simpson's paradox: https://en.wikipedia.org/wiki/Simpson%27s_paradox

Thoughtful attention must be paid to which priorities are most important to various stakeholders in order to define fairness in a given situation.

Research challenges

Much research in the past few years has focused on mitigating biased outcomes in predictive AI models. These can generally be broken down into three categories: ***data mitigation or pre-processing*** which attempts to modify the training data so that future algorithms that use it will be less biased, ***fair estimators or in-processing*** methods which optimise for fairness and predictive performance concurrently, and ***prediction or post-processing*** which intervenes after a model makes a decision and changes it if bias is detected (Figure 2). Many proposed methods make theoretical guarantees about improving certain fairness metrics. It is important to remember, however, that improving the quality, size, and makeup of the original training data is usually preferred since mitigation generally comes with a tradeoff between accuracy and fairness.

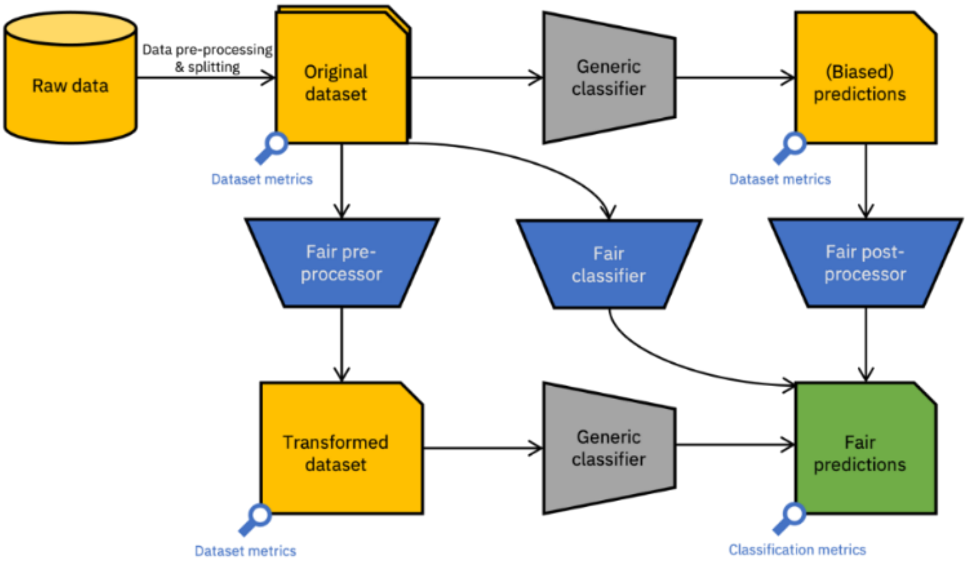

Figure 2: Various ways to mitigate bias in AI models. Bias should also be measured at multiple points in the AI lifecycle, including in deployment.

Data plays a key part in not only algorithmic fairness but machine learning at large. In fact, data is so valuable that transparency is often explicitly avoided as a business priority. Furthermore, ***data privacy*** continues to be of greater concern to the public (see section on "*Privacy-preserving AI*"). However, the ***opaqueness of data*** also makes it more difficult to reproduce and assess AI models and requires trust in the institutions designing them. ***Data and model transparency*** is an area of focus for researchers which aims to build trust in algorithms by divulging performance statistics (e.g. accuracy, fairness, robustness, etc.) and details about the development lifecycle.

To this end, various forms of "factsheets" have been proposed for data and models to document this information in a consistent and comprehensive manner[12,13].

Much research tends to focus on model training but what happens after a model is deployed is at least as important. Many high-stakes decisions are not made by algorithms alone and so investigating the ***interaction between humans and the machines*** that advise them is an important research objective. In many cases, human decision makers such as doctors and judges consider algorithmic predictions but must make the final decision themselves and understanding when and why they disagree is critical. This could improve fairness if the algorithm is flawed or not if the algorithm is correct but not trusted[14,15]. Furthermore, decision-making algorithms can create feedback loops and downstream effects which must be monitored and accounted for. In predictive policing, for example, the very act of increasing police presence in an area can cause the algorithm to continue suggesting crimes will be committed there[16].

Another big area of research is in assessing machine learning models which are not directly involved in decision-making but form a representation of the world which can be influenced by the same biases we have discussed. Large language models are an example of this. For instance, translation models have long struggled with assuming gender from context when translating between different languages. Models trained on large, unfiltered text can also learn to associate stereotypes which can amplify biases[17]. This in turn can affect downstream tasks such as sentiment analysis[18]. ***Representational bias*** can also happen in search results or algorithmic recommendations such as seeing mostly men when searching for images of CEOs[19].

Finally, the field of algorithmic fairness is deeply linked with ***explainable AI and causality research***. Models may feel unfair if there is no insight into how a decision was made or what aspects of the input led to a certain decision. Neural networks are often derided for being black boxes and navigating the competing desires for better accuracy and easier interpretability is a topic of much interest. Causal models could also play a role in ensuring fairness since they attempt to relate features which reliably cause an outcome instead of just recognising patterns of correlation (see section on "*Causality and machine leaning*"). These models tend to be more

[12] Arnold, M., et al. (2019). FactSheets: Increasing trust in AI services through supplier's declarations of conformity. IBM Journal of Research and Development, 63(4/5), 6-1.

[13] Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Iii, H. D., & Crawford, K. (2021). Datasheets for datasets. Communications of the ACM, 64(12), 86-92.

[14] Green, B., & Chen, Y. (2019, January). Disparate interactions: An algorithm-in-the-loop analysis of fairness in risk assessments. In Proceedings of the conference on fairness, accountability, and transparency (pp. 90-99).

[15] Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., & Mullainathan, S. (2018). Human decisions and machine predictions. The quarterly journal of economics, 133(1), 237-293.

[16] Ensign, D., Friedler, S. A., Neville, S., Scheidegger, C., & Venkatasubramanian, S. (2018, January). Runaway feedback loops in predictive policing. In Conference on Fairness, Accountability and Transparency (pp. 160-171). PMLR.

[17] Bolukbasi, T., Chang, K. W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. Advances in neural information processing systems, 29, 4349-4357.

[18] Thompson, A. (2017, October 25). Google's Sentiment Analyzer Thinks Being Gay Is Bad. Vice. https://www.vice.com/en/article/j5jmj8/google-artificial-intelligence-bias

[19] Kay, M., Matuszek, C., & Munson, S. A. (2015, April). Unequal representation and gender stereotypes in image search results for occupations. In Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (pp. 3819-3828).

easily explained since they work on logical principles and therefore have a great potential to be unbiased but they have yet to mature enough to be adopted at large scale.

Societal and media industry drivers

Vignette1: Detecting racial bias in targeted online advertising

Mark is the editor at an online magazine. One of the reporters has just published an article on a recent court case and Mark noticed an advertisement being shown on the article page was for a bail bond agency. Mark is African American. He asks some of his white colleagues what kind of ads they see with the article and one of them tells Mark he sees an ad linking him to a website which compiles the top ten law schools for criminal law. Mark wants to make sure the ads he runs are not targeting minority groups and reinforcing cycles of historical disadvantage. Was he being shown a bail bond ad specifically because he is black or was it a coincidence? He could ask more of his colleagues what kinds of ads they see but that would hardly provide a representative sample. After all, they all work in the same office and live in the same city. Mark wants to know overall, out of all the visitors to his site, what advertisements are being shown to each demographic group. He opens up the dashboard from the advertising company he uses and navigates to a screen showing fairness metrics for each ad shown on the website. The tool confirms that the bail bond agency is targeting minorities. Luckily, there is an option in the tool to equalise the serving rates for all ads in this category. Mark enables that option and checks back in a week to find that all demographic groups have an equal chance of being served any kind of legal ad.

Vignette 2: Detecting gender biases in content recommendation systems

Jennifer is a data scientist at a streaming video company. She wants to test out the recommendation algorithm for suggesting videos to watch next. Jennifer is worried that her daughter is constantly being shown makeup tutorials and hardly ever watches sports highlights anymore, even though she used to be a devoted tennis fan. She wonders if the recommendation algorithm has internalised certain gender stereotypes from its training data and is projecting them onto impressionable users. She decides to run an experiment: she creates two fake accounts, one as a young woman her daughter’s age and one as a young man of the same age. She looks at the same content for one week on both and keeps track of what kinds of recommendations she gets. At the end, she tallies up the number of recommendations from each category and analyses her results. It appears that the algorithm did recommend slightly more sports videos but hardly ever recommended makeup videos without watching a few first to the male persona. Jennifer creates a report from her results and shows it to her manager. They agree to change the algorithm for young users to reduce representational assumptions which could cause self-fulfilling behaviour. Jennifer also talks to her daughter who says she looked up makeup tutorials on her own because she was interested and she enjoys them. She also tells Jennifer that she is still interested in tennis and would like to take lessons but she thought nothing exciting was happening since she wasn’t being recommended any highlights.

Future trends for the media sector

Fairness must be considered in many media applications. The list below highlights a few examples likely to be relevant in the next few years:

Investigative journalism has revealed many of the shortcomings in algorithms cited above and will continue to play a key role in keeping the public informed and discovering unfair algorithms even as government regulation of AI increases. It is therefore integral that journalists and media professionals understand how machine learning algorithms can affect fairness, how to spot unfair algorithms, and what is being done about it.

Preventing discriminatory advertising actively is an important goal instead of just restricting demographic information from being used explicitly. In 2019, the U.S. housing department sued Facebook for discriminatory targeted advertising practices related to housing[20]. Facebook responded by preventing advertisers from using demographic information to target ads. Assessing the actual outputs for bias and preventing unintentional or subversive discrimination would be better.

Considering ***representational fairness*** can help prevent recommendation algorithms that perpetuate ***stereotypes*** and help break so-called ***filter bubbles***. Echo chambers in media are created when users are consistently shown the same type of content which can lead to corralling users into one of just a few different viewpoints when in reality, people are more multifaceted.

Diverse hiring is important in every field. With algorithmic resume filtering and candidate selection becoming widely used, it is paramount that we make sure these systems do not incorporate biases and, moreover, promote diversity.

Chatbots, automated text platforms usually used to provide support in lieu of a human, are becoming increasingly common and powerful. Automated textual interaction overall is a big trend and using large AI models such as GPT-3 without fully understanding them poses many risks. We have discussed how large language models can incorporate representational biases from data on which they are trained but more generally, as models become more general-purpose it becomes harder to test every use case and ensure no harmful text will be generated. Care must be taken when using models like this and understanding when and how they might break is crucial.

Furthermore, as ***generative AI*** gets better at creative tasks, we will begin to see it used to generate new stories such as movie, TV, and video game plots. AI has already been used in the creative process of movie trailers[21] and advertisements[22]. Amidst growing calls for more diversity in film and TV, these models have the potential to introduce bias by rejecting ideas that do not cater to the majority and leaning into stereotypes if used naively.

[20] HUD archives, HUD charges Facebook with housing discrimination over company's targeted advertising practice (2019): https://archives.hud.gov/news/2019/pr19-035.cfm

[21] IBM THINK Blog, IBM Research Takes Watson to Hollywood with the First “Cognitive Movie Trailer” (2016): https://www.ibm.com/blogs/think/2016/08/cognitive-movie-trailer/

[22] IBM THINK Blog, Lexus Europe Creates World’s Most Intuitive Car Ad with IBM Watson (2018), https://www.ibm.com/blogs/think/2018/11/lexus-europe-creates-worlds-most-intuitive-car-ad-with-ibm-watson/

"***Deepfakes***" are AI-generated images or videos which combine the action of one scene with a face from another. More generally, AI-assisted computer-generated imagery will become more prevalent and it is important to make sure this technology is used fairly and does not deny opportunities to diverse performers and that there is no gap in quality for performers of different appearances. These tools also bring up issues of consent and misinformation, related ethical issues to fairness, when used maliciously.

Visual intelligence tools or tools for photo/visual analysis have been widely discussed for their shortcomings in relation to recognising people and especially dark-skinned people. Great strides will continue to be made in remedying this situation but the greater question of whether the benefits of ubiquitous visual tracking are worthwhile is an ongoing debate. A slippery-slope argument says it can be hard to separate legitimate uses from objectionable overreach by, say, oppressive governments discriminating against minority groups. The first step before creating an algorithm should be anticipating and preventing its misuse.

Voice recognition and real-time voice transcription/captioning has been revolutionised by AI but it can still struggle understanding accents, dialects and less common languages. This technology can and is used extensively in media applications for accessibility but the benefits can be limited to native speakers of the language of the developers.

Abuse and hate speech detection on online platforms is a laudable ambition. The anonymity and platform scale mean automation is usually the only solution but the reliability of these systems to track harmful speech from and against different groups fairly has been questioned[23]. Robust and fair abuse detection is essential for the wellbeing of all.

Goals for next 10 or 20 years

The disparate effects of algorithms are considered from inception and all relevant stakeholders are included in the model building process. Decision-making algorithms are monitored for bias in production in real time to account for drift and feedback loops. Data and models are transparent and easy to understand; decisions are explainable and logical based on causal features. Representational models make guarantees about preventing the perpetuation of harmful stereotypes.

[23] Sap, M., Card, D., Gabriel, S., Choi, Y., & Smith, N. A. (2019, July). The risk of racial bias in hate speech detection. In Proceedings of the 57th annual meeting of the association for computational linguistics (pp. 1668-1678).

AI4media

ARTIFICIAL INTELLIGENCE FOR
THE MEDIA AND SOCIETY

This project has received funding from the European Union's Horizon 2020
research and innovation programme under grant agreement No 951911

info@ai4media.eu www.ai4media.eu