ACADEMIC REPORT:

Connor Faulkner
Department of Mathematics and Computing
Dundalk Institute of Technology

D00185637@student.dkit.ie

Introduction

Within this report, considerations for bias in technology will be discussed, firstly just an introduction into the discussion of bias in technologies, then examples that has been reported on of biases in different fields that I believe to be hugely influential in bring this to the light of the public. The cause of biases will be discussed next then this will lead into proposed strategies to mitigate bias and my own approach in my research topic.

INTRODUCTION

My research into bias in technology has been extensive, as we now live in a world where machine learning algorithms are used to automate simple and complex decision-making processes. As this is a huge area of topic, I will focus on computer models that make inferences from data about people, their identities and future behaviours. This will lead on to how organisations make decision on advertising, criminal sentencing and hiring to name a few, then talking about how lawless it is and the measures needed to reduce these biases.

Firstly, machine learning algorithms which are used in a range of areas can consume huge volumes of microdata that is collected from people and then used to influence them in a range of tasks from lending from a high-interest loan to recommendations of tv shows. I will try to show how bias is systemically creating less favourable situations to an individual within classified groups. Also, within this report on bias in technology I will look at specific cause of bias and best practices could detect and mitigate them. I want to recommend some policies that will promote the ethical deployment of these algorithms and fairness and how different approaches to bias detection can have an accuracy and fairness trade-off.

EXAMPLES OF ALGORITHMIC BIASES

Here I will discuss several examples that have been widely reported. Firstly, bias in online ads as founded by Sweeney that a 'sample of racially associated names and finds statistically significant discrimination in ad delivery based on searches of 2184 racially associated personal names' (Sweeney,2013). Essentially online search queries for names associated with African Americans were more likely to be return ads from a service relating to arrest records. That is not all for bias in online ads targeting Africans Americans, as she also found that high-interest credit cards were advertised to the user when it was inferred that the user was African American (Sweeney, 2014).

Bias in word associations is another area that has been observed. Researchers found that European names were associated as more pleasant than those of African-Americans, also in this research it found that the algorithm associated female names more than males with familial attributes such as 'wedding' and 'family whereas for career-related words the male names had a stronger association (Hadhazy, 2017). This algorithm has picked-up existing racial and gender biases shown by humans and this could have the reinforcing effect of bias if this learned association was used in a search-engine algorithm.

Bias in online recruitment, it was found that Amazon discontinued using recruitment algorithm after it discover the algorithm displaying gender bias. Within this report, it states that the algorithms used resumes

from a 10-year period, but which were predominantly male and as such the input training data was fed biased data. This algorithm looked for words relevant to skills sets and reference from the training data and as such it downgraded any resumes that contained the word 'women' (Dastin 2017).

Bias in facial recognition technology, this bias has a Netflix documentary discussing this and was hugely beneficial to understanding what bias in technology is. However, in this area of algorithm technology it was found that facial recognition systems from large companies failed to recognize darker-skinned complexions. In this study an error rate of 34.7% for dark-skinned women was found when trying to recognize the person. From this it was discovered the training sets were predominately white males and as such they were recognized 99% of the time (Hardesty, 2018).

Bias in criminal justice algorithms, probably the most impactful bias in the list as this algorithm help determine if some should go to jail. The COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) algorithm, is used in court by judges to infer if the defendant should be detained or released on bail, however a report found this to be bias towards African Americans (Angwin, 2016).

CAUSES OF BIAS

I believe there are two main causes that affect all bias in technologies which are incomplete or unrepresentative training data and human bias. Firstly, the incomplete training data is the data used to train these algorithms as in order to work they need data to infer from. However, if the data used to train algorithms have more representation than other groups of people then this will cause the algorithm to produce biased results. For examples from the Amazon hiring report, the training data has a huge representation of white males and lack of representation for women and from that the algorithm produced resumes only for males and excluded women simply because the data it was fed was biased. Another example of poor training data is when discussing facial recognition. From the reports above it was caused by having statistical insignificant amounts of underrepresented groups such as darker-skinned faces.

Historical human bias is another cause of bias in technology. The report of bias in the criminal justice algorithms it was found to prejudice towards the African Americans group. The algorithm called COMPAS uses training data from previous records of arrest, however, as it's been shown the African American group historically has been more likely to be arrested due to racism and or other inequalities with the system this will be reflected from the training data and as such cause a biased prediction on the person in question.

PROPOSED SOLUTIONS

I would like to discuss possible strategies to detect bias, firstly understanding the various causes of bias is the first step in reduce bias in any algorithm. However even if there are discovered flaws in the training sets and they are corrected this could still hamper the results as context is needed when trying to reduce bias. All algorithms need to be careful with sensitive information because even when the algorithm is blinded from this sensitive data it can still produce the same results if this data was used previous in a discriminatory way. This was discovered in a report that highlighted that stand in for sensitive data such as zip code can be infer by the algorithms like proxies for race or gender (Zarsky, 2015). Another example of algorithms using proxies for sensitive data is when Amazon excluded areas for same-day delivery system. Their decision was made by several factors that help with their profitability model but this excluded areas of poor and predominantly African Americans areas, which just transferred the proxies of sensitive data to racial classification (Ingold, 2016).

So even if creating training sets that can represent groups with sensitive information to prevent discriminatory results, it still can reproduce these results simply from proxies. So, the next topic tries to readdress this issue is creating an algorithm that isn't always accuracy but tries to capture some fairness. The goal is to stop reinforcing inequalities and as such developers of algorithms need to find context within the algorithms. This is a hard one to implement as how can a developer know the cost of societal fairness when decreasing or increasing accuracies and if one group gets affected to these changes. One implementation for the developers is that fixing bugs when maximising for accuracy can help and the addition of dataset which have underrepresented groups added for additional training of the algorithms can reduce unfair results.

However ethical frameworks are needed to completely encompass this discussion. In the United states it is the wild west as there is not all-encompassing legal framework around data, privacy and standards of the use of AI. In the EU, it has the General Data Protection Regulation (GDPR), which is a legal framework that sets guidelines for the collection and processing of personal information from individuals, which in most cases can help standardise the use of AI in Europe. They have also introduced guidelines called 'Ethics Guidelines for Trustworthy AI' which showcases seven principles: (1) human agency and oversight, (2) technical robustness and safety, (3) privacy and data governance, (4) transparency, (5) diversity, non-discrimination and fairness, (6) environmental and societal well-being, and (7) accountability (EU,2018). These guidelines set a clear path for the future of AI in Europe; however, it is not as simple as measuring fairness in the code but by determining like I said previously the trade-off between accuracy and fairness and context of the algorithm and data.

Other considerations for mitigating bias could possible be self-regulatory practices such as audits for bias with algorithms, for example Facebook completed a civil rights audit to find if its handling of issues and people from protected groups after it was revealed how it was handling these issues such as content moderation and privacy (Locklear, 2018).

MY APPROACH

From all my research into this area of bias in technology, it has help me grasp a better understand into my research in genetic data and more specifically, the Analysis of genetic substructure using genome-wide SNP and essentially from my work I would like to see how common variations of SNPs can separate races to geographically locations. I believe bias of race could be apparent within my data, it has been suggested in a report that similar data to mine has a white bias and this could be detrimental if this data were used in AI assisted medication prescription (Drozda, 2015). This could have effects on my research as it could be separating the white males and females into their geographical location more than the under-represented Chinese or Africans. My research is about separating the common variations of SNPs within different races so I believe my research is indirectly looking at potential bias in genomes that future algorithms could abuse if not regulated correctly. This data is protected by the European GDPR and looks of the information is anonymized to protect these individuals, however if for example this research was added on into a country like China and used to marginalize an ethnicity this could be hugely impactful if abused.

REFERENCES

Sweeney, Latanya. "Discrimination in online ad delivery." Rochester, NY: Social Science Research Network, January 28, 2013. Available at https://papers.ssrn.com/abstract=2208240 (last accessed 26/4/2021).

Sweeney, Latanya and Jinyan Zang. "How appropriate might big data analytics decisions be when placing ads?" PowerPoint presentation presented at the Big Data: A tool for inclusion or exclusion, Federal Trade Commission conference, Washington, DC. September 15, 2014.
Available at https://www.ftc.gov/system/files/documents/public_events/313371/bigdata-slides-sweeneyzang-9_15_14.pdf (last accessed 26/4/2021).

Hadhazy, Adam. "Biased Bots: Artificial-Intelligence Systems Echo Human Prejudices." Princeton University, April 18, 2017. Available at https://www.princeton.edu/news/2017/04/18/biased-bots-artificial-intelligence-systems-echo-human-prejudices (last accessed 26/4/2021).

Dastin, Jeffrey. "Amazon scraps secret AI recruiting tool that showed bias against women", Reuters, October 11, 2017.
Available at https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G (last accessed 26/4/2021).

Hardesty, Larry. "Study Finds Gender and Skin-Type Bias in Commercial Artificial-Intelligence Systems." MIT News, February 11, 2018. Available at http://news.mit.edu/2018/study-finds-gender-skin-type-bias-artificial-intelligence-systems-0212 (last accessed 26/4/2021).

Angwin, Julia, Jeff Larson, Surya Mattu, and Laura Kirchner. "Machine Bias." ProPublica, May 23, 2016. Available at https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing (last accessed 26/4/2021).

Zarsky, Tal. "Understanding Discrimination in the Scored Society." SSRN Scholarly Paper. Rochester, NY: Social Science Research Network, January 15, 2015. https://papers.ssrn.com/abstract=2550248
(last accessed 26/4/2021).

Ingold David," Amazon Doesn't Consider the Race of Its Customers. Should It?", Bloomberg.com. Available at http//www.bloomberg.com/graphics/2016-amazon-same-day (last accessed 26/4/2021).

EU Commission, "Ethics Guidelines for Trustworthy AI' ", 2018, https://ec.europa.eu/futurium/en/ai-alliance-consultation, (last accessed 26/4/2021).

Locklear, Mallory. "Facebook Releases an Update on Its Civil Rights Audit." Engadget (blog), December 18, 2018. Available at https://www.engadget.com/2018/12/18/facebook-update-civil-rights-audit/ (last accessed 26/4/2021).

Drozda K, Wong S, Patel SR, et al. Poor warfarin dose prediction with pharmacogenetic algorithms that exclude genotypes important for African Americans. Pharmacogenet Genomics. 2015;25(2):73-81. doi:10.1097/FPC.0000000000000108