How to Solve AI Bias

By Robin Bradley
r.bradley2@herts.ac.uk
University of Hertfordshire

ABSTRACT

Bias in AI is a topic that impacts machine learning and artificial intelligence technology that learns from datasets and its training data. While gender discrimination and chatbots showing bias have recently caught people's attention and imagination, the overall area of how to correct and manage bias is in its infancy for business use. Further, little is known about how to solve bias in AI and how there could potent for malicious misuse at large scale. We explore this area and propose solutions to this problem.

1. INTRODUCTION

Bias is not a new phenomenon. In fact, bias and inequality have been part of human behaviour and society for thousands of years. It is only in the last century that women have given the right to vote. It some countries women are still treated differently to men. Racism is still an issue in society, with the public face of racism highlighted in football with monkey chants at black players still occurring at top level games around the world. American football has had to install the Rooney Law to try to combat racism in hiring of coaches. As for ageism, in society people are being affected by ageism, sometimes in subtle ways that don't seem that major. But these small acts of bias can have a major effect on people's mental health and they do actually discrimination against them. This micro bias was publicised well by a NY Times article [1] about the subject of ageism that stats, 'mere microaggressions … the forms ageism often takes: pervasive employment discrimination, biased health care, media caricatures or invisibility. When internalized by older adults themselves, ageist views can lead to poorer mental and physical health.'

When applying for jobs people have stopped including their date of birth. In fact, people have stopped including their gender and even their names on job applications to avoid bias in the recruitment process. Even this does not total avoid bias when reading resumes or during job applications. A paper by the University of California [2], discusses how we all have conscious biases that we use our views and bias to judge evidence and that ' how our perception may inadvertently have a negative impact on an applicant being considered.' Even someone's address or area where they live can create a bias in people's opinions. People can make false implications or assumptions about an individual's potential success or ethnicity based on their address. In the UK Police forces are using AI to identify future criminal [19]If someone is from an address in a lower socio-economic area, an unconscious bias could mean you assume that person did not do well at their last jobs and can also imply ethnicity, socio-economic status, and effect your judgement when assess whether they are the right person for the job or role.

Therefore, bias is in everyday life and society in general. But surely this bias does not apply to computers and AI?

2. AI METHODOLOGY

AI stands for Artificial Intelligence. A simple definition of AI is: an area of computer science that emphasizes the creation of intelligent machines that work and react like humans. In this basic definition, if we are trying to create intelligent machines that react and think like humans does this include bias? If they perfectly copy society and our behaviours and views, conscious and subconscious thoughts and logical thinking, obviously the answer is Yes! But AI is not perfect and also how it is built has an impact on whether it is biased.

At its foundation, AI has been built based on models. Models and patterns that are designed and programmed by humans. The AI algorithms learn from the humans building these models and using the associated training data to learn about what is the right and wrong answer. We train the AI by using sets of data and improving the AI's logic until it is acceptable. So, at its core people are still in control and the currently most AI is not autonomous. Although this is changing and with advancements in technology and by its nature, AI is learning by itself and developing a self-learning capability.

But at its core, AI and below that machine learning, are lots of complex decision trees and maths that make logical decisions based on the information that you provide the AI to process as part of the development and testing cycle.

2.1 Is AI, biased or otherwise, a good thing?

But AI is already being used for the good of society. Google and Amazon are using AI in their customer engagement to deliver better customer service. IBM have developed IBM Watson for Oncology that helps physicians quickly identify key information in a patient's medical record, surface relevant evidence and explore treatment options, quicker than humans can process the information.

Facial recognition is being used widely in China to help government surveillance and used in multiple countries as part of the criminal justice system.

There are lots of practical uses of AI that are aiding society today. Like facial recognition, some are uses of AI are unpopular with groups of end consumers, but regardless of whether you think AI is good or bad thing, it here to stay. AI will only become more widely used in the years to come as the technology develops and becomes more usable, combined with the advancements in 5G, quantum computing and blockchain. McKinsey [3] sees AI delivering global economic activity of around $13 trillion by 2030. By the same year, PricewaterhouseCoopers reckons on $15.7 trillion - more than the current combined output of China and India. Tech investor Tej Kohli, however, believes the impact will be much faster and exponentially larger, however, potentially worth $150 trillion by 2025.

Whichever estimate of the size of the AI economy you believe, the question isn't if AI going to grow. It is how do we control and manage AI as it develops and ensure it remains a good force in our society.

3. REVIEW OF THE CURRENT STATE

Is AI unbiased or not? Scholars and research indicates there are issues in data quality and potential bias in AI processes. [15-17]. If we focus on the fact that AI is based on, developed and learns from the information that it processes, then the answer is, it depends. It depends on what information you provide the AI. Just like in human evolution, if you teach a child bad habits when it is learning, when it is an adult, if it isn't told or taught otherwise it will display these traits. AI is similar, if you train it with bad or bias data, unless you tell or teach it otherwise it will not work optimally and will contain some bias.

The question then becomes, is mirroring society in AI a good thing, which includes our conscious and unconscious bias? Or should all of AI be unbiased? The purist answer is that AI should be unbiased and as it is a computer/machine it shouldn't have any biased view. This inherently by the way AI is currently taught, should definitely include AI not have any unconscious bias views.

But actually, companies and societies are not black and white. Bias can be used in a good way to positively discriminate people, eg in healthcare. Therefore, AI can be 'biased' for the good of society if that is a conscious and informed decision for that specific customer journey or experience.

4. POTENTIAL ACTIONS TO CORRECT BIAS

In order to minimize bias, how do we define and measure fairness? How should we build the definitions of fairness in AI? To answer that we need to answer the question of 'What is fairness?' As per a Mckinsey study[7], in their article they claimed to 'identified at least 21 different definitions of fairness and said that even that was "non-exhaustive.' In the article it goes on to explain that fairness depends on the questions and subject matter and can be a complex questions. It continues in quoting 'Kate Crawford, co-director of the AI Now Institute at New York University, used an CEO image search highlight the complexities involved: how would we determine the "fair" percentage of women the algorithm should show? Is it the percentage of women CEOs we have today? Or might the "fair" number be 50 percent, even if the real world is not there yet? Much of the conversation about definitions has focused on individual fairness or treating similar individuals similarly, and on group fairness—making the model's predictions or outcomes equitable across groups, particularly for potentially vulnerable groups. '

Experts disagree on what is fairness, so how can we solve bias in AI. You also have to take into account the business and social objectives behind these AI models. What one person thinks is fair, might not be the next persons view. The same applies when creating AI models and testing data for the AI learning.

There is disagreement in society and the business world on the best way to resolve these issues and challenges. For example, some have suggested that creating different patterns and therefore decision trees for different groups may achieve fairness. This holds true as there is bias in some of the underlying models in data. But creating these models is time consuming and requires

unbiased expertise. Alternatively, unfortunately creating a single unified model and pattern for all people is impossible, so there does need to be adaptation of AI models and patterns dependant on the outcome and customer journey that the AI is trying to solve to try to maintain fairness and remove bias

4.1 **Using AI models to solve bias.**

Maybe AI can be the answer in solving bias in society?

If humans are nearly incapable of not being biased, even in an unconscious way, maybe using machines can actually help society to remove bias in processes and decision making? If we assume that AI is subjective and based on logical decision trees processed in Nano seconds, the logic is sound that AI can actually remove bias in decision making. AI will assess and review data, create and predict patterns that are based on logic and the data it is trained on.

But the problem is that humans create this training data. AI models are trained and therefore learn from the data it is provided. Data that is provided for and created by biased humans. Therefore, if AI is provided with training data that is bias it will learn from this bias data and use this as the basis of the patterns it creates to make its logic decisions going forward. So, training data and how you train AI is key to ensuring that you don't have biased AI.

But how do you create unbiased training data? Everyone has bias, even if it is unconscious bias, so how do we ensure the AI learns without bias?

4.2 Training your AI without bias

Let's follow the logic that AI only learns from what is input into the AI model, whether that be documents or data that it reads, objects or faces that it sees or sound or senses that it processes. If this training data is without bias, then the AI model will be without bias. Preparing your data for AI models becomes a key factor. [18]

Creating training data that is unbiased is key to creating AI without bias. AI only learns what it is trained and from the data, in whatever form, it processes. So how do humans that are inherently and unconsciously biased, create unbiased training data.

There are views that you should create teams that are with a mixed background, in genre, race, ethnicity, educational background and sexual orientation. But how big does your team have to be to create a full diverse team with totally unbiased views that cover all possible biases? The simple answer is that it can't. As per Tom Chatfield article [4] 'There's no such thing as a single set of ethical principles that can be rationally justified in a way that every rational being will agree to. Depending upon your priorities, your ethical views will inevitably be incompatible with those of some other people in a manner no amount of reasoning will resolve. Believers in a strong central state will find little common ground with libertarians; advocates of radical redistribution will never agree with defenders of private property; relativists won't suddenly persuade religious fundamentalists that they're being silly'. Everyone are not going to agree some of the time.

So, what is the answer? Maybe AI can help react unbiased training data?

4.3 How is AI helping remove bias from training data and therefore AI.

The good news is that we are becoming more aware of the problem and there is acceptance that this accepting there is bias in AI and that it is an issue is the first step to solving the problem. In a Datarobot report [5] '64% of all respondents say they are "very to extremely" confident in their ability to identify AI bias, nearly half of surveyed organizations (42%) also admit to being "very to extremely" concerned about AI bias occurring in their organizations'

So, most organisations are aware that bias is a problem in AI. But how to deal with it still a big issue. One way to deal with it is to us AI to detect bias in your AI models. Can AI be trained to spot things that do not fit the patterns that are expected and therefore have been created by bias? The logic stands that AI and computers can process and review data thousands of times quicker than humans, so this is a logical approach. During AI testing, these AI review tools can be used to flag areas of concern in the respective AI models and corresponding results so the AI developers, testers or conversational analysts can amend their test data accordingly. The test or training data can then be amended to improve the AI acceptance rate. Out of the 350 c-suite members interviewed by DataRobot [5], 56% were deploying AI algorithms to try to solve bias in their training data.

4.4 So how do you practical correct AI bias?

Training data needs to be reviewed and as unbiased as possible. Increasingly, companies and scholars have been trying to address this issue. In the emerging field of Human Centered Data Science (HCDS) [8, 9, 10, 11] have begun to investigate data science practices, showing the necessary, responsible, and increasingly accountable human activities that take place between data and models [9, 12, 13, 14].

In modern development teams, an agile test and learn process is mainly used and best suited to increasing the acceptance rates and removing bias from your AI. During your sprints, test results should be reviewed and compared to expectations. On a daily basis training data can be tweaked, based on the results of this continuous test and learn cycle. Using AI can help process this testing 1000+ times faster than is humanly possible, thus aiding removing bias quicker from the data. This agile process will require a good feedback process and loop within your testing and development team to quick amend the test data and continue with further test and learn cycles.

Also, as part of the end of sprint review process, a good feedback loop will allow analysis of these test result with your product owner. Having strong product ownership and accountability of the AI and related tools and applications impacting your customer journeys is key to ensuring the AI is actual meeting your customer and business objectives. Without this business input it is more likely that the technical and project teams could misinterpret testing results data and this can lead to creating even more bias in the AI algorithms.

But as we all know, projects can go wrong, projects have limited time and budgets. Also, companies are still mainly driven by increasing shareholder returns, i.e. profits. AI development teams and projects, whilst in the incubation phase might not be driven as much as other projects by these constraints, but as AI matures and acceptance rates increase, AI projects will become more beholden to increasing stakeholder value and quicker returns. With these increased pressures and constraints, there is a conflict between achieving a perfectly unbiased AI product and something that is good enough for release that it will drive value for the business.

5 CONCLUDING THOUGHTS

There needs to be an approach to stop AI bias at a macro level. Individual companies and development teams are doing their best to stop AI bias, but bias is still happening, potentially due to lack of knowledge, funding or business pressures.

As globalisation takes hold, this issue cannot be looked at in isolation. All countries and companies are trying to untangle the web of AI and with the challenges of bias. At a macro level there need to be more ownership by large corporates who are driving this AI market and creating the AI.

As per a Forbes [6] article, Bernard Marr discusses how some large companies are looking to solve this problem. ‘It’s possible AI may be the solution to, as well as the cause of this problem. Researchers at IBM are working on automated bias-detection algorithms, which are trained to mimic human anti-bias processes we use when making decisions, to mitigate against our own inbuilt biases. ‘ Amazon, Google and other large players in the AI market are all also looking at software or AI to defect and reduce bias in AI.

As we continue to buy and build AI off these large corporates their tools and methods become more important in controlling the landscape and fairness in the software. As AI becomes more mature, it also becomes more comples and sometimes AI is seen as a Blackbox. As it gets more complex, how do we understand how all the algorithms work? Also, will these tech companies who are developing the AI share all their algorithms to be tested and assessed for fairness and lack of bias? These patterns and models can be seen as these companies USP and therefore they do not share its inner working with the wider technology community for fear of losing market share and profits.

5.1 Can governments or regulators do more?

The simple answer is yes. Researchers at the Royal United Services Institute the defence and security think tank published a report and the RSU report commissioned by the Centre for Data Ethics and Innovation a government body warns of an absence

of national guidelines to govern algorithms [19]. The report highlighted “a lack of sufficient empirical evidence “ to help understand whether they are biased.”
The problem as with all areas of emerging technology is the sparsity of skills and the rapid rate of change. As the complex AI landscape is changing so quickly, do the governments of countries actually see this an immediate problem. The UK government in March 2019 launched an investigation in AI bias. Conducted by the Centre for Data Ethics and Innovation (CDEI), the investigation will focus on areas where AI has tremendous potential – such as policing, recruitment, and financial services – but would have a serious negative impact on lives if not implemented correctly. [20]
In many developed countries the governments are in power for 4 to 5 year terms and is managing effectively the developing AI technology going to materially affect their chances of getting re-elected? Some governments have task forces and are looking at AI and its impact, but are they looking at fairness and bias? Do they have the resources, expertise and reach and work at a pace that can keep up with change and materially effect the industry?

The other problem that this is a global issue. How do governments influence companies to do what is ethically the right thing? Global companies can easily move to the local with the regulation that suits their business needs.

There needs to be some global guidelines and principles that are put in place to manage AI and remove bias. This needs to be driven by the technology community alongside global regulatory and governing bodies. How these guidelines and principles are going to work needs to be transparently shared and communicated globally.

Is this happening fast enough? No. Can we catch up and put these principles in place before it is too late? Yes. But there needs to be material actions now before the AI models and algorithms become too mature that they have deeply ingrained bias in them.

Humans and machines together need to be used to setup principles, models and guidelines that can be used to reduce or minimise bias in AI and make it fair and ethical for everyone now and in the future.

References

[1] NY Times: https://www.nytimes.com/2019/04/26/health/ageism-elderly-health.html Ageism: A 'Prevalent and Insidious' Health Threat by Paula Span, 2019
[2] University of California: https://ucnet.universityofcalifornia.edu/working-at-uc/your-career/talent-management/talent-acquisition-employment/how-unconscious-biases-may-impact-reviewing-a-resume.pdf - How Unconscious Biases May Impact Reviewing a Resume – prepared by Systemwide Talent Acquisition June 2016
[3] Forbes: Can The AI Economy Really Be Worth $150 Trillion By 2025? By Andrew Cave June 2019
[4] One Zero, Medium publication: https://onezero.medium.com/theres-no-such-thing-as-ethical-a-i-38891899261d There's No Such Thing As 'Ethical A.I.' by Tom Chatfield, , Jan 2020
[5] DataRobot Inc: https://www.datarobot.com/lp/the-state-of-ai-bias-in-2019/ The State of AI Bias in 2019 – 2019
[6] Forbes: https://www.forbes.com/sites/bernardmarr/2019/01/29/3-steps-to-tackle-the-problem-of-bias-in-artificial-intelligence/#e5426617a128 Steps to tackle the problem of bias in Artificial Intelligence, Bernard Marr, Jan 2019
[7] McKinsey and Company: https://www.mckinsey.com/featured-insights/artificial-intelligence/tackling-bias-in-artificial-intelligence-and-in-humans Tackling bias in artificial intelligence (and in humans) – by Jake Silberg and James Manyika June 2019
[8] Cecilia Aragon, Clayton Hutto, Andy Echenique, Brittany Fiore-Gartland, Yun Huang, Jinyoung Kim, Gina Neff, Wanli Xing, and Joseph Bayer. 2016. Developing a research agenda for human-centered data science. In Proceedings of the 19th ACM Conference on Computer Supported Cooperative Work and Social Computing Companion. ACM, 529–535
[9] Nadia Boukhelifa, Anastasia Bezerianos, Ioan Cristian Trelea, Nathalie Méjean Perrot, and Evelyne Lutton. 2019. An Exploratory Study on Visual Exploration of Model Simulations by Multiple Types of Experts. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. ACM, 644
[10] Michael Muller, Melanie Feinberg, Timothy George, Steven J Jackson, Bonnie E John, Mary Beth Kery, and Samir Passi. 2019. Human-Centered Study of Data Science Work Practices. In Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems. ACM, W15.
[11] Dakuo Wang, Justin D Weisz, Michael Muller, Parikshit Ram, Werner Geyer, Casey Dugan, Yla Tausczik, Horst Samulowitz, and Alexander Gray. 2019. Human-AI Collaboration in Data Science: Exploring Data Scientists' Perceptions of Automated AI. Proceedings of the ACM on Human-Computer Interaction 3, CSCW (2019), 211.
[12] Melanie Feinberg. 2017. A design perspective on data. In Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems. ACM, 2952–2963.
[13] Lisa Gitelman. 2013. Raw data is an oxymoron. MIT press.
[14] Sean Kandel, Andreas Paepcke, Joseph Hellerstein, and Jeffrey Heer. 2011. Wrangler: Interactive visual specification of data transformation scripts. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM, 3363–3372
[15] Samir Passi and Steven Jackson. 2017. Data vision: Learning to see through algorithmic abstraction. In Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing. ACM, 2436–2447.
[16] Samir Passi and Steven J Jackson. 2018. Trust in Data Science: Collaboration, Translation, and Accountability in Corporate Data Science Projects. Proceedings of the ACM on Human-Computer Interaction 2, CSCW (2018), 136.

[17] Kathleen H Pine and Max Liboiron. 2015. The politics of measurement and action. In Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems. ACM, 3147–3156.
[18] Tye Rattenbury, Joseph M Hellerstein, Jeffrey Heer, Sean Kandel, and Connor Carreras. 2017. Principles of data wrangling: Practical techniques for data preparation. " O'Reilly Media, Inc."
[19] Fears over police AI to identify future criminals "The Telegraph, Edward Malnick' February 2020
[20] UK government investigates AI bias in decision-making 'https://www.artificialintelligence-news.com by Ryan Daws' March 2019