Australian Human Rights Commission

Fast Facts for Media

Using artificial intelligence to make decisions: Addressing the problem of algorithmic bias

This is an Australia-first technical paper – the result of a collaboration between the Australian Human Rights Commission and Gradient Institute, Consumer Policy Research Centre, CHOICE and CSIRO's Data61.

Algorithmic bias can result in decisions that are unfair, or even unlawful. This paper demonstrates how businesses can identify algorithmic bias in artificial intelligence (AI), and proposes steps they can take to address/mitigate the problem.

Addressing the problem of algorithmic bias also offers practical guidance for companies to ensure that when they use AI systems, their decisions are fair, accurate and comply with human rights.

About the hypothetical simulation we've used:

Addressing the problem of algorithmic bias uses a **hypothetical simulation** to test how algorithmic bias can arise. The hypothetical simulation we chose is: *an electricity retailer uses an AI-powered tool to decide how to offer its products to customers, and on what terms.*

This is not intended to be representative of a particular business or to imply that algorithmic bias in AI disproportionately affects the energy sector. Rather, we chose this example to demonstrate an everyday business decision-making scenario. The findings are applicable to a broad range of sectors and businesses.

We used **simulated data** in this paper instead of real aggregated data, made up of many individuals.

What is algorithmic bias?

- Algorithmic bias is a kind of propensity of AI-based decision making to lead to unfair outcomes.
- Algorithmic bias can arise in different ways. Sometimes the problem is with the design of the AI-powered decision-making tool itself. Sometimes the problem lies with the data set that was used to train the AI tool.

One good example of how the data used to train the AI tool can result in algorithmic bias is seen in Scenario 2 in the paper (see 4.5). In this scenario, out-of-date historical data is used in the AI system. Because the data is out of date, it does not reflect the fact that the gender pay gap has reduced. Making decisions based on the old data, the AI system rejects more women as 'unsuitable'. In other words, the AI system is used to conclude, wrongly, that women are less likely to be able to pay their bills. If current data had been used, more women would have been selected as 'suitable customers' by the AI system.

Consumer Policy Research Centre · GRADIENT INSTITUTE · CSIRO DATA 61 · CHOICE

Australian Human Rights Commission

Fast Facts for Media

How can algorithmic bias affect human rights?

Algorithmic bias can cause real harm. A decision influenced by algorithmic bias can lead to a person being unfairly treated, or even suffering unlawful discrimination, on the basis of characteristics such as race, age, sex or disability.

Key recommendations in the technical paper:

- Human rights should be considered whenever a company uses new technology, like AI, to make important decisions.
- Anyone who is considering the use of an AI system to make decisions should ensure that their decision-making process is fair and lawful. Fulfilling this responsibility starts **before the AI system is used in a live scenario**.
- The AI system should be rigorously designed and tested to ensure it does not produce outputs that are affected by algorithmic bias.
- Once the AI system is operating, it should be closely monitored throughout its lifecycle to check that algorithmic bias does not arise in practice.
- Using an AI system responsibly and ethically extends beyond simply complying with the narrow letter of the law.

This simulation highlights **five general approaches to mitigating algorithmic bias**. These approaches are potential tools in a 'toolkit' of mitigation strategies. Each set of circumstances needs to be assessed and the appropriate mitigation strategy applied.

1. **Acquire more appropriate data** – responsibly obtain additional data points or new types of information relating to individuals inaccurately represented, or under-represented in the data set.
2. **Pre-process the data** – this can involve operations such as reducing the influence of a protected attribute like gender before using an AI system to make decisions. This may prevent individuals from being treated differently based on protected attributes, lowering the risk of algorithmic bias.
3. **Increase the model complexity** – a simple model can be easier to test, monitor and interrogate but an over-simplified model will be less accurate and can lead to the model making generalisations that favour the majority group.
4. **Modify the AI system** – an AI system may be designed or modified to correct for existing societal inequalities, as well as other inaccuracies or issues in data sets causing algorithmic bias.
5. **Change the target** - finding a fairer measure to use as the target variable could help mitigate algorithmic bias.

The technical paper also outlines the considerations for business for each of the five approaches. **Ultimately though, if an AI system cannot produce fair, accurate results, it should not be used at all.**

You can download and read *Using artificial intelligence to make decisions: Addressing the problem of algorithmic bias* at tech.humanrights.gov.au

Consumer Policy Research Centre | GRADIENT INSTITUTE | CSIRO DATA 61 | CHOICE