tsg insights

November 2021

Algorithmic bias – the cruel mirror artificial intelligence (AI) and machining learning (ML) reflects back at us

This whitepaper is the Sixth in a series which act as companion pieces:

- An overview of artificial intelligence (AI) and machine learning (ML).
- Introduction to the testing of AI and ML.
- Testing of AI (artificial intelligence) and ML (machine learning) – supervised learning.
- Testing of AI (artificial intelligence) and ML (machine learning) – unsupervised learning.
- Testing of AI (artificial intelligence) and ML (machine learning) – reinforcement learning.
- AI and machine learning – algorithmic bias – the cruel mirror AI and ML reflects back at us.

In this paper, we're focusing on algorithmic bias and how it can act as a cruel mirror, revealing our own hidden predispositions, often in a surprising and embarrassing way.

What is algorithmic bias?

Algorithmic bias is not an error, but a systemic fault that's rooted in the way algorithms work. In general, the more data an algorithm is exposed to, the better. But it can also have too much of the wrong data – or be lacking the right data in certain areas. It can also have dated data – just as humans might have "old-fashioned" views. As with people, this can result in prejudice in favour of or against an individual or group.

For AI and ML training, the underlying issue is that training data might include bias. The key risk is that the data bias will not necessarily be recognised as even existing until it's too late. Below are three real-life examples of this.

Amazon's recruitment system

In November 2018, Amazon decommissioned its AI based recruitment system. It was biased against women, rarely inviting them for interviews. This was ironic, given that Amazon's rationale for building the AI in 2014 was to eliminate bias.

For four years they had struggled to fix it, removing any references to gender, re-evaluating words and terms used in applications, and much else. Nothing worked.

So, what went wrong? Only Amazon know the precise answers and they're not divulging anything. Speculation suggests that the main problems were that the AI had been trained with 10 years' worth of historical data and that Amazon's engineering workforce is nearly all male.

This in itself could have led the AI to incorrectly judge that 9 out of 10 selected applicants should be male since this was the status quo. From there, the AI could have made several secondary conclusions. For example, female-only colleges would score lower and male-dominated sports higher, while more direct, "masculine" language was better. This would have effectively constructed underlying gender bias.

Untangling this appears to have been so complicated and cost-intensive that Amazon, one of the world' richest companies, decided to walk away from it altogether.

Kentucky bail decisions

In the US state of Kentucky in 2011a new law called HB 463 was introduced. It instructed judges to consult an algorithmic system to decide pre-trial about releasing suspects with or without bail payment. Decision factors included employment status, education and previous criminal records.

The aim was to grant release without bail consistently, more often and at less cost than the previous judge-only system. Today, the majority of release-without-bail decisions are made purely based on the algorithmic system.

It's partially successful, but discriminates against African Americans. Since 2011, their rate of release without bail has stayed at about 25%, while for white Americans it has increased from a similar 25% to 35%. This is despite changes being made to the algorithm to address the imbalance.

What's going wrong? It could be indirect bias. White Kentuckians affected mostly live in rural areas, where unemployment is lower, and income and education higher than in urban areas, where African Americans predominately live. The prime bias may therefore be location rather than race.

Microsoft's Tay and Zo chatbots

Tay was a chatbot developed and trained by Microsoft that was supposed to simulate a 19-year-old on Twitter. But as The Verge memorably put it, "Twitter taught Microsoft's AI chatbot to be a racist asshole in less than a day".*

What went wrong? Apparently, either by design or accident, Tay was inundated with alt-right requests. ML meant that it quickly picked up on the racist (and sexist) language used. Microsoft tried to retain Tay but that failed, so it was replaced by Zo. However, Zo got decommissioned as well. As Quartz noted: "Zo is politically correct to the worst possible extreme; mention any of her triggers, and she transforms into a judgmental little brat."**

So Microsoft went from "not politically correct" to "too politically correct". Given that Zo has not been replaced, finding the Goldilocks space of "just right" is clearly challenging.

*Source: 2016 article by James Vincent in The Verge **Source: 2018 article by Chloe Rose Stuart-Ulin in Quartz

What are the business risks of algorithmic bias?

At TSG we are expert testing practitioners, but we're not lawyers. The legal discussions regarding algorithmic decision systems (ADS) are ongoing and examples like the ones above are just part of the picture. One thing is for sure – it's complicated.

As usual with business risk, there are tangible and intangible implications. Tangible consequences would be legal challenges, possibly resulting in fines, regulatory interventions, and out-of-court settlements.

However, the non-tangible consequences could arguably be even more severe:

- Loss of reputation – Amazon received much negative publicity for their chatbot fiascos. If similar events occurred to a company with a lesser reputation, it could have a much more significant effect.

- Loss of talent – those directly affected by perceived or real bias are likely to avoid joining the company concerned, resulting in a loss of talent and, in a wider sense, diversity. Recruitment would become harder and more expensive.

- Unwanted regulatory attention – past transgressions will be noted and regulators are likely to intervene earlier and with more scrutiny. Uber's many issues with licensing in London in recent years are an example of this – their app had allowed too many uninsured, disqualified or previously disciplined drivers to re-enrol using false credentials.

What does the future hold?

As outlined in the Overview Whitepaper, AI has already experienced two “AI winters”. We are currently in the third “AI spring”. AI and ML are definitely advancing (as our Introduction Whitepaper highlights) but what about the ongoing issue of algorithmic bias?

There are in-depth discussions in progress concerning how the requirement to explain and justify any decision made by any algorithm might be enshrined in law. For AI and ML that could become a basic requirement, while being difficult to achieve in practice.

For example, if a deep neural network of the future declines your credit application and negatively affects your credit score, will the network operator be able to justify this decision? If not, what damages could you claim?

Given that supervised and unsupervised learning is based on existing data, algorithmic bias is a particular concern when using these approaches. These potential legal ramifications will make the quality assurance and testing of AI and ML systems a key consideration.

On the bright side, awareness of algorithmic bias is the first step towards avoiding it. Good quality assurance and testing are here to help. They can contribute towards building strong business cases for whatever AI or ML solutions that companies are looking to create.

TSG provides expert guidance on AI and ML, as well as assurance and testing services. We make change happen, safely and predictably. If you have any question about issues covered in this whitepaper or would like to know more about how we can help you, please contact us now. **Call: +44 (0) 207 469 1500** **Email: info@tsgconsulting.co.uk** **www.tsgconsulting.co.uk**