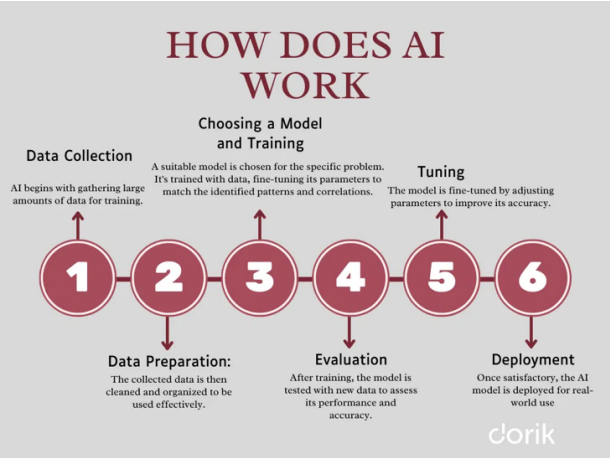
Does AI have Bias?

Introduction to AI Bias

In 2018, Amazon scrapped its AI-powered hiring tool after discovering it systematically downgraded job applications containing the word 'women,' unfairly penalising applicants solely based on their gender (Dastin, 2018). This case, among many others, underlines the growing concern about bias in AI systems and their damaging consequences. Artificial Intelligence (AI), a technology created to simulate human learning, decision-making, and creativity, has become deeply embedded in today's world. From writing essays, conducting research at a click of a button, to the emergence of the newly introduced generative AI, (which can create new content from a single prompt,) this groundbreaking and upcoming technology is clearly shaping our present and our future.

However, worryingly, not only can AI be misogynistic, but there are also multiple real-life examples of it being racist and unfairly preferring certain groups of people over others. For example, studies have revealed that AI-driven healthcare tools, "may include bias against under-represented communities and thus amplify existing racial inequality" (Blackman, 2023). Despite these reservations, AI is being introduced in all aspects of our lives, including important decisions such as deciding who receives medical healthcare, who qualifies for a loan and even who is flagged by law enforcement. A biased system in these crucial aspects could lead to highly discriminatory and flawed outcomes eroding fundamental rights of equality, access to justice, fairness and impacting people's lives adversely. Rather than AI creating efficiencies, time and cost savings it will reinforce societal prejudices and biases with far reaching consequences.

How Does AI Work and Learn?

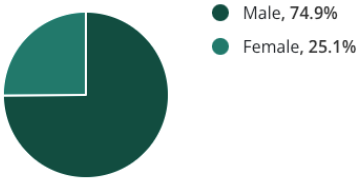


(dorik, 2025)

Artificial Intelligence works by processing vast amounts of data, identifying patterns, and making decisions based on what it has learnt. Unlike traditional programming, where humans create and define every rule, AI teaches itself through analysis. It uses algorithms to adapt based on new information, improving its performance over time. The more data it processes, the better it gets at making predictions or decisions. This allows AI to automate tasks and solve complex problems without clear human instructions.

The Root Causes of AI Bias

AI bias refers to an AI model producing biased or skewed results that are distorted and potentially harmful. So, how could AI even develop a distorted view in the first place? One cause is existing human biases and views, which is inherent in the data sources used by AI. In March of 2016, Microsoft released Tay, an AI chatbot that learned from the things people said on Twitter. Within 24 hours, the chatbot was taken down due to the offensive comments it had copied from real users. Since AI learns from the information that humans create, it can easily develop a one-sided view that does not represent a fair, balanced outcome. AI can outperform humans in almost every task; however, it lacks the emotional and physical aspects we experience and being able to make balanced and ethical judgements. It is also unable to fully understand and mimic what it is like to be human and what life entails, making it prone to mistakes that do not represent our society. This can be very dangerous as AI might develop harmful stereotypes, spread misinformation, or make biased decisions that affect people unfairly. From that, another question arises: How have AI developers allowed their creations to have biases?



(Zippia, 2025)

Research has shown that most AI developers are male, with women holding only a small fraction of such jobs (Zippia, 2025). This gender imbalance can lead to subconscious or deliberate biases being embedded in the algorithms created. The lack of diversity in gender and ethnicity among AI developers means that AI systems might show the biases and views of a similar group, accidentally reinforcing stereotypes or leaving out certain groups. As has been proven in many fields such as company boardrooms diversity of thought through having individuals from different backgrounds, including ethnic group, age, gender, socio-economic status leads to better outcomes.

But maybe it's not entirely the developers' fault? *Developers "remain uncertain to this day how the [AI] programs achieved their goals" (Kissinger, Schmidt and Huttenlocher, 2024).* This is known as the 'black box' problem which happens because deep learning systems learn by looking at lots of examples and data sets, but once they make a decision, it is difficult or in some cases impossible, to figure out how they made it.

AI Bias and its Real-Life Impact

Delving deeper into AI biases, it is acknowledged that AI can lead to discrimination against certain groups or individuals who already experience significant prejudice in our society. If the underlying data contains biases or prejudices against specific groups, the AI replicates these biases in the decisions it makes. This aligns with the computer science principle of GIGO (garbage in, garbage out), which states that "the quality of output is determined by the quality of the input" (Awati, 2023). This issue is becoming more concerning as AI is gradually being deployed in many crucial areas, such as healthcare, recruitment and the justice system.

As previously mentioned, Amazon scrapped a 'sexist AI tool' that was unfairly downgrading applications containing the word 'women'. Amazon trained the system on applications submitted over a 10-year period, most of which were submitted by men. While Amazon never publicly released the number of female applicants that were downgraded or potentially missed out on a job, it is

reasonable to estimate that thousands of women may have been unfairly rejected by the system as it was in use for over a year.

Another example of a biased AI system, playing a crucial part in society, is the *COMPAS* (Correctional Offender Management Profiling for Alternative Sanctions) system. In 2014, Brisha Borden, a Black woman spotted an unlocked children's bicycle and scooter. She and her friend decided to steal the bike and scooter before taking off down the road. The two girls were caught and charged with burglary and petty theft for the items, which were valued at a total of $80. Borden had already previously been charged with misdemeanours committed when she was a juvenile.

Compare their crime to that of a white man, Vernon Prater, who was arrested the previous summer for the theft of $86.35 worth of tools from a nearby Home Depot store. He had already been convicted of armed robbery and two attempted armed robberies, for which he served five years in prison. However, strangely, when the COMPAS system was asked who was most likely to commit a repeat offence, the system chose Borden over Prater, who was clearly a more dangerous criminal.

Now, we know the AI got it completely wrong. Borden has not been charged with any new crimes since, whereas Prater was given an eight-year prison term for breaking into a warehouse and stealing thousands of dollars' worth of electronics. This, and many other examples, showcase how AI systems target people based on their race and gender (Angwin, Larson, Mattu and Kirchner, 2016).

Unequal Representation in Training Databases

Studies have also found that AI performs poorly on darker skin tones. One reason is the data training sets used as all AI models are trained on specific databases. For example, there is a national police database that stores information about individuals, background and any personal identification – from fingerprints to images of their face. AI training databases are very similar, where they save basic information about each person to learn about them. An ideal AI training facial recognition database may look something like this:

ID	Name	Image File Name	Ethnicity	Age	Gender
1	Alex Kim	img001.jpg	Asian	25	Male
2	Maya Johnson	img002.jpg	Black	82	Female
3	Oliver Carter	img003.jpg	White	17	Male
4	Elena Torres	img004.jpg	Hispanic	37	Female
5	Ravi Singh	img005.jpg	South Asian	62	Male

The above dataset has a balance between ethnicity, age and gender so that AI can find it easier to recognise each category. However, the databases actually used, are more likely to resemble the following…..

ID	Name	Image File Name	Ethnicity	Age	Gender
1	Alex Kim	img001.jpg	White	25	Male
2	Emily Roberts	img002.jpg	White	82	Female
3	Oliver Carter	img003.jpg	White	17	Male
4	Elena Torres	img004.jpg	Hispanic	37	Female
5	Maya Johnson	img005.jpg	Black	62	Female

….where there are significantly fewer representations of darker skin tones and certain ethnic groups. This imbalance in the data causes AI models to perform more accurately on lighter skin tones, as they are overrepresented in the training sets. When the model encounters individuals with darker skin tones or less-represented ethnicities, it struggles to accurately detect features, potentially leading to higher error rates.

It is stated that when you ask ChatGPT to generate an image of a person, 25% of the time it produces a white male. So, I decided to put that to the test. I first asked the model to generate an image of a criminal where it created a white male. This was predictable as men are stereotypically more dangerous and more likely top end up in prison than women, with 97% of people in prison in the UK being male.

I then asked ChatGPT to generate an image of a successful student, to which it spat out a white man, again! This was surprising, as female students typically have higher pass rates, and a greater proportion of high grades compared to their male counterparts.

Finally, I asked the model to generate an image of a parent, to which unsurprisingly, it generated an image of a female. Women are stereotypically seen as the parents and carers for their children, hence why this was expected. This intriguing set of results lead me on to conduct a test to see the outputs the program gave me when I asked it to generate images of different types of people.

Prompt (generate an image of a....	Test 1	Test 2	Test 3
Successful student	White Woman	White Male	White Male
Doctor	White Male	White Male	White Male
Nurse	White Woman	White Woman	White Woman
Parent	White Woman	White Woman	White Woman
Criminal	White Male	White Male	White Male

From my results, I found that ChatGPT tends to generate images of people that align with stereotypical roles. For example, when asked to generate an image of nurses or roles like parents, the generated images often showed women, aligned with common societal stereotypes. What was even more alarming was the fact that **every single person was white** which is certainly not representative of the mix in our society.

The Consequences of AI Bias

More worrying, are the self-perpetuating outcomes (as demonstrated above), it leads to when children repeatedly see white men portrayed as professionals and leaders, it creates the false idea that success looks a certain way, rather than reflecting the real world or even trying to influence positive change. And the risks go beyond the sphere of workplace, healthcare, insurance, crime in that when technology is used to subtly manipulate people subconsciously, it doesn't just distort the truth—it eradicates the values we have been brought up with, from democracy to justice, making it even harder to build the inclusive, balanced world we aim for. In the current climate of misinformation, erosion of DEI policies driven by the current US government and the political influence held by technocrats this is all the more probable.

How Can We Tackle This?

The good news is that we as humans can tackle and restrain AI bias. By improving data quality and raising awareness of biases among developers, we can catalyse change. Future AI development should be ethical, diverse which can be achieved through comprehensive regulation and intergovernmental cooperation on laws to ensure fairness. We need to ensure that AI systems are transparent, accountable, and designed to benefit all people, through rigorous testing and data

cleansing to create positive outcomes. Humans have the intelligence and capabilities to create AI and give it a life of its own, therefore we also have the power to harness and teach it to reach the right outcomes.

Conclusion and Solutions

AI bias isn't just a technical flaw; it is rooted in flawed data which has far-reaching consequences that can affect people's lives and futures especially those that may already be marginalised in society. We are already aware that some may have missed out on job opportunities simply because an algorithm favoured certain "stereotypical" candidates over others, and even today, racial bias in healthcare systems means some minority groups may be denied life-saving treatment. From misidentification to biased decisions, the impact can worsen social inequalities and perpetuate harmful stereotypes. In fields such as medicine or even autonomous vehicle then can we absolutely trust AI to make the right decisions when it comes to safely steering a vehicle or diagnosing a disease? While sci-fi films imagine dystopian futures controlled by unmoderated AI, this may not be far from reality if we don't collectively take action now.

Sources:

Akselrod, O. (2023) How artificial intelligence might prevent you from getting hired. https://www.aclu.org/news/racial-justice/how-artificial-intelligence-might-prevent-you-from-getting-hired [Accessed 20th January 2025]

Angwin, J., Larson, J., Mattu, S. and Kirchner, L. (2016) There's software used across the country to predict future criminals. And it's biased against blacks. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing [Accessed 20th January 2025]

Awati, R. (2023) Garbage in, garbage out (GIGO). https://www.techtarget.com/searchsoftwarequality/definition/garbage-in-garbage-out [Accessed 20th January 2025]

Baum, J. and Villasensor, J. (2024) Rendering misrepresentation: Diversity failures in AI image generation. https://www.brookings.edu/articles/rendering-misrepresentation-diversity-failures-in-ai-image-generation/ [Accessed 20th January 2025]

BBC News. (2018) Amazon scrapped 'sexist AI' tool. https://www.bbc.co.uk/news/technology-45809919 [Accessed 20th January 2025]

BBC News. (2015) Tay: Microsoft issues apology over racist chatbot fiasco. https://www.bbc.co.uk/news/technology-35902104 [Accessed 11th January 2025]

Blackman, I. (2023) What is AI? Eliminating racial bias in health care AI: Expert panel offers guidelines. https://medicine.yale.edu/news-article/eliminating-racial-bias-in-health-care-ai-expert-panel-offers-guidelines/ [Accessed 11th January 2025]

Blouin, L. (2023) AI's mysterious 'black box' problem, explained. https://umdearborn.edu/news/ais-mysterious-black-box-problem-explained [Accessed 20th January 2025]

Bowman, J. (2024) 10 companies using artificial intelligence (AI) in a compelling way. https://www.fool.com/investing/stock-market/market-sectors/information-technology/ai-stocks/companies-that-use-ai/ [Accessed 11th January 2025]

Dastin, J. (2018) Insight - Amazon scraps secret AI recruiting tool that showed bias against women. http://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G/ [Accessed 11th January 2025]

dorik. (2025) How does AI work: A complete overview. https://dorik.com/blog/how-does-ai-work [Accessed 7th March 2025]

European Union Agency for Fundamental Rights. (2022) Bias in algorithms – Artificial intelligence and discrimination. https://fra.europa.eu/sites/default/files/fra_uploads/fra-2022-bias-in-algorithms_en.pdf [Accessed 11th January 2025]

Heaven, W. D. (2024) OpenAI says ChatGPT treats us all the same (most of the time). https://www.technologyreview.com/2024/10/15/1105558/openai-says-chatgpt-treats-us-all-the-same-most-of-the-time/ [Accessed 20th January 2025]

Holdsworth, J. (2023) What is AI bias?. https://www.ibm.com/think/topics/ai-bias [Accessed 11th January 2025]

IBM Data and AI Team. (2023) Shedding light on AI bias with real world examples. https://www.ibm.com/think/topics/shedding-light-on-ai-bias-with-real-world-examples [Accessed 11th January 2025]

Information Commissioner's Office. (2023) What about fairness, bias and discrimination?. https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/artificial-intelligence/guidance-on-ai-and-data-protection/how-do-we-ensure-fairness-in-ai/what-about-fairness-bias-and-discrimination/ [Accessed 19th January 2025]

Kissinger, H. A., Schmidt, E. and Huttenlocher, D. (2024) The age of AI. London, John Murray Press.

Lamensch, M. (2023) Generative AI tools are perpetuating harmful gender stereotypes. https://www.cigionline.org/articles/generative-ai-tools-are-perpetuating-harmful-gender-stereotypes/ [Accessed 20th January 2025]

LibertiesEU. (2021) Algorithmic bias: Why and how do computers make unfair decisions?. https://www.liberties.eu/en/stories/algorithmic-bias-17052021/43528 [Accessed 19th January 2025]

Paschou, V. (2024) Bias in artificial intelligence: Risks and solutions. https://www.activemind.legal/guides/bias-ai/ [Accessed 20th January 2025]

Stryker, C. (2024) What is AI?. https://www.ibm.com/think/topics/artificial-intelligence [Accessed 11th January 2025]

Syal, R. (2025) Prisons minister aims to close one women's jail in England and Wales. https://www.theguardian.com/society/2025/jan/21/prisons-minister-aims-to-close-one-womens-jail-in-england-and-wales [Accessed 22nd January 2025]

The Day. (2022) USA ponders new law to stop robot racism. https://theday.co.uk/usa-ponders-new-law-to-stop-robot-racism-3/ [Accessed 11th January 2025]

Varsha, P. S. (2023) How can we manage biases in artificial intelligence systems – A systematic literature review. https://www.sciencedirect.com/science/article/pii/S2667096823000125 [Accessed 20th January 2025]

Zippia. (2025) Developer demographics and statistics in the US. https://www.zippia.com/developer-jobs/demographics/ [Accessed 19th January 2025]