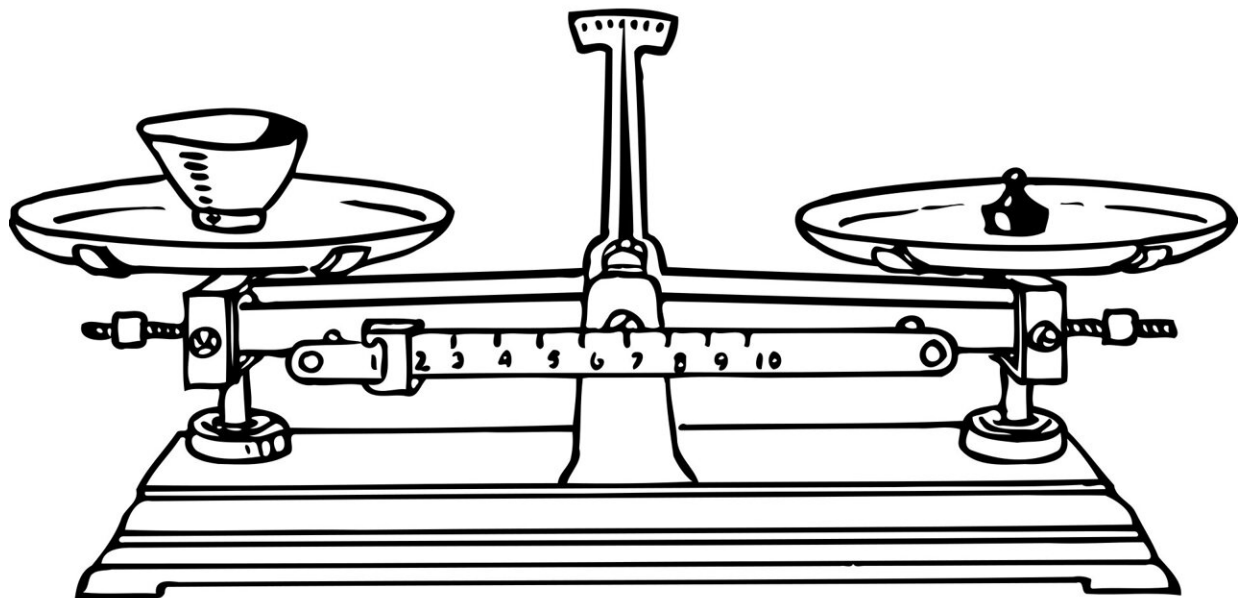# Building fairness into AI is crucial, and hard to get right

March 19 2024, by Ferdinando Fioretto

Credit: Pixabay/CC0 Public Domain

Artificial intelligence's capacity to process and analyze vast amounts of data has revolutionized decision-making processes, making operations in health care, finance, criminal justice and other sectors of society more efficient and, in many instances, more effective.

With this transformative power, however, comes a significant responsibility: the need to ensure that these technologies are developed and deployed in a manner that is [equitable and just](). In short, AI needs to be fair.

The pursuit of fairness in AI is not merely an ethical imperative but a requirement in order to foster trust, inclusivity and the [responsible advancement of technology](). However, ensuring that AI is fair is a major challenge. And on top of that, my research as a computer scientist [who studies AI]() shows that attempts to ensure fairness in AI can have unintended consequences.

Why fairness in AI matters

Fairness in AI has emerged as a [critical area of focus]() for researchers, developers and policymakers. It transcends technical achievement, touching on [ethical, social and legal dimensions of the technology]().

Ethically, fairness is a cornerstone of building trust and acceptance of AI systems. People need to trust that AI decisions that affect their lives—for example, hiring algorithms—are made equitably. Socially, AI systems that embody fairness can help address and mitigate historical biases—for example, those against women and minorities—fostering inclusivity. Legally, embedding fairness in AI systems helps bring those systems into alignment with anti-discrimination laws and regulations around the world.

Unfairness can stem from two primary sources: the input data and the algorithms. Research has shown that input data can [perpetuate bias]() in various sectors of society. For example, in hiring, algorithms processing data that reflects societal prejudices or lacks diversity can [perpetuate

"like me" biases. These biases favor candidates who are similar to the decision-makers or those already in an organization. When biased data is then used to train a machine learning algorithm to aid a decision-maker, the algorithm can propagate and even amplify these biases.

Why fairness in AI is hard

Fairness is inherently subjective, influenced by cultural, social and personal perspectives. In the context of AI, researchers, developers and policymakers often translate fairness to the idea that algorithms should not perpetuate or exacerbate existing biases or inequalities.

However, measuring fairness and building it into AI systems is fraught with subjective decisions and technical difficulties. Researchers and policymakers have proposed various definitions of fairness, such as demographic parity, equality of opportunity and individual fairness.

These definitions involve different mathematical formulations and underlying philosophies. They also often conflict, highlighting the difficulty of satisfying all fairness criteria simultaneously in practice.

In addition, fairness cannot be distilled into a single metric or guideline. It encompasses a spectrum of considerations including, but not limited to, equality of opportunity, treatment and impact.

Unintended effects on fairness

The multifaceted nature of fairness means that AI systems must be scrutinized at every level of their development cycle, from the initial design and data collection phases to their final deployment and ongoing evaluation. This scrutiny reveals another layer of complexity. AI systems are seldom deployed in isolation. They are used as part of often complex

and important decision-making processes, such as making recommendations about hiring or allocating funds and resources, and are subject to many constraints, including security and privacy.

Research my colleagues and I conducted shows that constraints such as computational resources, hardware types and privacy can significantly influence the fairness of AI systems. For instance, the need for computational efficiency can lead to simplifications that inadvertently overlook or misrepresent marginalized groups.

In our study on network pruning—a method to make complex machine learning models smaller and faster—we found that this process can unfairly affect certain groups. This happens because the pruning might not consider how different groups are represented in the data and by the model, leading to biased outcomes.

Similarly, privacy-preserving techniques, while crucial, can obscure the data necessary to identify and mitigate biases or disproportionally affect the outcomes for minorities. For example, when statistical agencies add noise to data to protect privacy, this can lead to unfair resource allocation because the added noise affects some groups more than others. This disproportionality can also skew decision-making processes that rely on this data, such as resource allocation for public services.

These constraints do not operate in isolation but intersect in ways that compound their impact on fairness. For instance, when privacy measures exacerbate biases in data, it can further amplify existing inequalities. This makes it important to have a comprehensive understanding and approach to both privacy and fairness for AI development.

The path forward

Making AI fair is not straightforward, and there are no one-size-fits-all

solutions. It requires a process of continuous learning, adaptation and collaboration. Given that [bias](#) is pervasive in society, I believe that people working in the AI field should recognize that it's not possible to achieve perfect fairness and instead strive for continuous improvement.

This challenge requires a commitment to rigorous [research](#), thoughtful policymaking and ethical practice. To make it work, researchers, developers and users of AI will need to ensure that considerations of [fairness](#) are woven into all aspects of the AI pipeline, from its conception through data collection and algorithm design to deployment and beyond.

This article is republished from [The Conversation](#) under a Creative Commons license. Read the [original article](#).

Provided by The Conversation