How fair AI can make us richer

Sandra Wachter[1]

We are all aware that artificial intelligence (AI) has now become an integral part of our lives. AI systems are behind mundane tasks such as displaying search results on Safari, preparing travel routs on Google and suggesting new music on Spotify. But algorithms also steer important parts of our lives: if we get admitted to university, if we are hired, fired or promoted, if we get insurance, social benefits or a loan, or even if we have to go to prison. Algorithms can touch almost every aspect of our lives.

The benefits are clear. They can help us make more efficient, cheaper, and consistent decisions. At the same time AI can introduce new risks and aggravate old ones, for example by replicating and exacerbating existing social and economic inequalities.[2] This should come as no surprise: algorithms can only be trained on existing and historical data and, if left alone, will inevitably pick up and learn from injustices and inequalities in past human decision-making.[3] One need only reflect on who in our society usually gets admitted to university, gets promoted, or receives loans and who does not to realise the magnitude of the risks if AI is allowed to lock in and preserve these existing biases.

AI and non-discrimination law

Of course, inequalities are not a new aspect of society. One could be tempted to think that the risks of AI can be mitigated simply by applying existing legal frameworks, such as European non-discrimination law, to algorithmic decision-making. Unfortunately, the law, at least in its current form, is quite powerless to help algorithmically disadvantaged groups for two reasons: the complaint-based system and the evidential requirements.[4]

Non-discrimination law is enforced through a complaint-based system. The system operates on a simple premise: the affected individual (or group) will be aware or feel that injustice is occurring, and raise a complaint in response. They will see, for example, that somebody is treated better than they are, or promoted over their head, or receives better insurance premiums, or acquires a spot at a university while they are denied. Or in cases where discrimination is less obvious, they will at least "feel" the subtle prejudices or norms that prevent them from succeeding. In other words, the affected party will notice that something is unfair and bring a complaint.

In the algorithmic world this comparative element of injustice is slowly being eroded. Individuals are very often unable to see when they are offered higher prices on products than others. Rather, they only see an algorithmically prepared version of the truth. People are also unaware that algorithms are able to infer intimate details such as ethnicity, gender, age or sexual

[1] Sandra Wachter, Associate Professor and Senior Research Fellow, E-mail: sandra.wachter@oii.ox.ac.uk Oxford Internet Institute, University of Oxford, 1 St. Giles, Oxford, OX1 3JS, UK. This work of the Governance of Emerging Technologies research programme at the Oxford Internet Institute has been supported by the British Academy Postdoctoral Fellowship grant nr PF2\180114, Luminate/Omidyar Group, and the Miami Foundation.

[2] VIRGINIA EUBANKS, AUTOMATING INEQUALITY: HOW HIGH-TECH TOOLS PROFILE, POLICE, AND PUNISH THE POOR (2018).

[3] CATHY O'NEIL, WEAPONS OF MATH DESTRUCTION: HOW BIG DATA INCREASES INEQUALITY AND THREATENS DEMOCRACY (2017).

[4] Sandra Wachter, Brent Mittelstadt & Chris Russell, *Why fairness cannot be automated: Bridging the gap between EU non-discrimination law and AI*, 41 COMPUTER LAW & SECURITY REVIEW 105567 (2021).

orientation about them and use this information to exclude them from seeing job advertisements or ads for housing. AI simply discriminates "behind their backs."

A complaint-based system loses its force if the complainant does not know that they have been harmed. Even if we are aware that algorithms have potentially disadvantaged us, problems remain. Proving illegal discrimination has occurred can be even harder when an algorithm is involved.

Algorithms often use very untraditional data sources such as social network data, shopping history, clicking behaviour, data on articles read and videos watched to evaluate a person and make decisions. It will be very hard for claimants to persuade a judge that these untraditional data sources such as eating or viewing habits are correlated with protected attributes such as gender, ethnicity or ability and thus constitute a seemingly "neutral provision, criterion or practice" that is, in fact, discriminatory.[5]

Compared to algorithmic decision-making, it will typically be easier to prove in court that human comprehensible decision-making provisions, criteria or practices have an adverse effect on protected groups. Convincing a judge that poor working conditions for part-time workers correlate with gender, that salary thresholds in loan decisions affect minorities, or that social benefits only granted to married couples affect the LGBT* community is an intuitive task. Intuitive links between these provisions and protected attributes are often based in lessons about historical inequalities in society.

Bias testing

Complex algorithmic models trained on diverse data and consisting of potentially millions of interdependencies are self-evidently not similarly intuitive. To close this gap, many researchers, companies, NGOs, and governments around the world are looking for solutions to mitigate and prevent bias in automated decision-making. One way to address this challenge are so-called bias or fairness tests.[6] If individuals are not aware that they are harmed or even if they are, they have difficulties proving a correlation between protected attributes and adverse effects. Technical tools such as bias tests may offer one potential solution to this conundrum.

Why we should be fair

However, this leaves us with the question as to why companies and the public sector should change their business models and processes to reduce algorithmic bias and make fairer decisions. After all, the balance sheets of Big Tech firms do not currently appear to suffer too severely from the effects of sexism, racism, ableism or heterosexism. And here I want to offer three reasons as to why testing for bias is in the collective best interest of industry, the public sector, and society at large: namely ethical, legal, and financial reasons.

First, taking an active role in dismantling inequality is a laudable and ethical goal. Trying to optimise your processes to align with equality and inclusiveness demonstrates a commitment to fairness. Without going in ethical and political theory at length, fairness is simply the right

[5] Sandra Wachter, *Affinity Profiling and Discrimination by Association in Online Behavioral Advertising*, 35 BERKELEY TECH. LJ 367 (2020).
[6] Sahil Verma & Julia Rubin, *Fairness definitions explained*, *in* 2018 IEEE/ACM INTERNATIONAL WORKSHOP ON SOFTWARE FAIRNESS (FAIRWARE) 1–7 (2018).

thing to do because it makes decisions just, or based on desert or merit rather than luck. [7]It means ensuring the most talented or deserving people get through the door, rather than just those people who resemble successful candidates of the past.[8] With that said, ethical interests alone are often insufficient to motivate real social, political, and economic change. And we may not want to be reliant on the ethical conscience of other people to make the "right" decision.

So I offer a second reason to make AI fairer: it is legally mandated to thrive towards equality, at least in Europe. Non-discrimination law expects both the public and the private sector to take an active role in dismantling inequality.[9] Keeping things as they are and perpetuating the status quo is simply not good enough. The declared aim of European non-discrimination law is substantive or *de facto* equality. The goal is to erode inequalities, dismantle disparities and achieve parity and inclusion.[10] But here again we might run into the risk that some see this as a legal burden that motivates people to only do the bare minimum to demonstrate legal compliance.

Often when I give talks to academics, the public sector, and companies on algorithmic bias and AI fairness, the first question asked is about how to motivate buy-in among the private sector and powerful actors. This mirrors common questions asked of non-discrimination law about who in society should bear the costs of equality. These questions are sensible; for work on algorithmic bias, fairness, and equality to have impact, it must be adopted by the developers, deployers, and users of AI systems making critical decisions that impact peoples' lives.

In reality, important ethical and legal interests are unfortunately often insufficient to overcome concerns about how fairness or equality would impact on profits, business models, and other interests. In the context of development, for example, ethics is often the first thing abandoned when it runs counter to business interests.[11]

Setting aside the larger question of whether it is justifiable to prioritise profits over ethics and the law, why is this the case? Fairness and the rules of non-discrimination law are often seen as brakes on progress or as hurdles that hamper innovation and economic growth. Non-discrimination provisions are often portrayed as requiring "charity" or non-meritorious decision procedures. If true, greater fairness and equality would seemingly run counter to the power, financial and other interests of entrenched actors, and challenge their position in the market or society.

Thankfully, this framing of "equality as charity" is wrong. Substantive equality is an investment, not a price to pay. Therefore, I offer a third argument aimed specifically at powerful actors that are unwilling or unconvinced by ethics and the law alone: fairer AI can make you richer.

How fairness can make you richer

---

[7] Richard J. Arneson, *Equality and equal opportunity for welfare*, 56 PHILOSOPHICAL STUDIES 77–93 (1989).
[8] JOHN RAWLS, A THEORY OF JUSTICE (2009).
[9] SANDRA FREDMAN, DISCRIMINATION LAW (2002).
[10] TARUNABH KHAITAN, A THEORY OF DISCRIMINATION LAW (2015).
[11] Brent Mittelstadt, *Principles alone cannot guarantee ethical AI*, 1 NATURE MACHINE INTELLIGENCE 501–507 (2019).

People sometimes fear that non-discrimination law, if applied with the aim of substantive equality, leads to giving people a "leg up" who lack merit and relevant qualifications. According to this line of thinking, certain people only walk through the door because of the protected attributes that they possess. Giving people a chance who do not deserve it at the expense of more qualified people is what many people fear. But this is not what the law requires.

The very essence of non-discrimination law is not to push people through the door who do not deserve a chance on merit, but rather to realise that very talented people are currently "falling through the cracks" and never receive fair consideration. Non-discrimination law wants us to question the decision criteria we use and assess whether they are good proxies for merit. It wants us to determine whether our decision-making processes carry within them a legacy of oppression and inequality while failing to accurately measure competence.

We know for example that reference letters reflect strong racial and gender bias. Very often women and people of colour are described as "hard-working" whereas their white male counterparts are described as "trailblazers" and "geniuses" even when they perform equally well.[12]

Grades and exam scores likewise often carry strong gender and racial bias which means that people of colour and especially girls of colour receive worse grades even where their capabilities are objectively comparable to their peers.[13]

Salaries and promotion track records are also marked by race[14], ability[15], gender and sexuality[16] lines. People from disenfranchised communities are often overseen for promotions or offered lower salaries for reasons of prejudice and bigotry, not competence.

But we make decisions and teach our algorithms on these biased data and decision-making criteria on a daily basis. Reference letters can make or break if we get a job, grades decide if we are admitted into a good university, our salaries define if we get loans or housing, promotions, or healthcare. We use them as proxies for ground truth and merit, but these data types are not representative of the whole truth. Unfortunately, data always inherits and reflects the unequal status quo and thus not all data is a good proxy for merit or competence, and this comes at a high price.

Every bank that fails to give a loan to someone that could repay is losing money. Every talented employee who missed out because they are a minority is a lost opportunity for economic growth. Every university that does not admit a gifted student because of their gender identity, sexuality or ability will lose the chance to improve their international reputation and attract funding.

Injustice has a high price, for individuals and a democratic society, but also for the economy.

Which bias test to use?

---

[12] CAROLINE CRIADO PEREZ, INVISIBLE WOMEN: EXPOSING DATA BIAS IN A WORLD DESIGNED FOR MEN 102 (2019).

[13] RENI EDDO-LODGE, WHY I'M NO LONGER TALKING TO WHITE PEOPLE ABOUT RACE 66–67 (2020).

[14] JEAN HALLEY, AMY ESHLEMAN & RAMYA MAHADEVAN VIJAYA, SEEING WHITE: AN INTRODUCTION TO WHITE PRIVILEGE AND RACE (2011).

[15] PAUL DAVID HARPUR, ABLEISM AT WORK: DISABLEMENT AND HIERARCHIES OF IMPAIRMENT (2019).

[16] JEAN HALLEY & AMY ESHLEMAN, SEEING STRAIGHT: AN INTRODUCTION TO GENDER AND SEXUAL PRIVILEGE (2016).

This brings us back to the question that motivated this special issue: how can we fix algorithmic bias and make AI fairer? There are many different bias tests currently available with sometimes contradicting and mutually exclusive normative assumptions. In previous work[17] I have shown how these tests can be classified as either bias preserving (i.e. measuring fairness based on equal error rates across groups) and bias transforming (i.e. measuring fairness based on equal decision rates across groups). The trouble with bias preserving tests is that they trust the status quo to be accurate, neutral and fair, which is unfortunately too often not the case. Bias transforming tests, on the other hand, focus on parity of outcomes between groups. They flag up when disparity occurs between groups and do not take equality in the status quo for granted.

I have argued elsewhere[18] that when AI is used to make important decisions about people in Europe in protected sectors, bias transforming tests such as Conditional Demographic Disparity (CDD) are legally preferable if not mandated. They align closely with the central aim of European non-discrimination law, to achieve substantive or *de facto* equality, and give deployers the opportunity to assess and question the validity of the conditioning variables and possibly change them according to contextual interests and requirements.

Nonetheless, choosing the right bias test will depend on purpose and application, sector, legal jurisdiction, and the accuracy of the dataset. Put simply, "contextual equality"[19] is the aim of the law and thus a feature not a bug, even if that means that a "one size fits all" solution for fairness is impossible and perhaps not desirable. Luckily, however, there are a wide range of tools already available that let deployers choose the fairness test and metric most appropriate for their purpose.

Tackling bias and fairness in AI is an essential task. When we feed data about our biased world and societies into an algorithm without questioning how talent, merit, and inequality have been conceptualised and measured in our past decisions, we are losing out on good people.

However, if we see AI as a mirror of society that shows us where inequalities exist, we can use this knowledge as a starting point to rethink our selection strategies and criteria for housing, insurance, parole, education, hiring, and other critical areas of life. We would not only make algorithms more accurate and fairer but also get "richer"—from a financial perspective—but, more importantly, as a society.

---

[17] Sandra Wachter, Brent Mittelstadt & Chris Russell, *Bias preservation in machine learning: the legality of fairness metrics under EU non-discrimination law*, 123 W. VA. L. REV. 735 (2020).
[18] Wachter, Mittelstadt, and Russell, *supra* note 4.
[19] *Id.* at 1–2.