The Theory of Artificial Immutability

Protecting Algorithmic Groups under Anti-Discrimination Law

Speaker: Professor Sandra Wachter (Oxford Internet Institute)[1] [2]

17 October 2022

In its *Frontier Talks* series, the Tech Cluster invites leading academics in their fields to discuss frontier research related to technology. This memo and the preceding interview summarise the first Frontier Talk of the academic year 2022-2023.

INTERVIEW WITH THE SPEAKER

by Réka Heszterényi

How did you come to work on the intersection between artificial intelligence (AI), algorithms and discrimination?

I have been working for a while on algorithms that exacerbate inequalities in society and have published on the topic. Bias is embedded in all aspects of our lives, so I was interested whether laws, especially anti-discrimination, are fit to deal with them. I concluded that the answer was no. In my new paper, The Theory of Artificial Immutability: Protecting Algorithmic Groups under Anti-Discrimination Law, I also discovered that algorithms not only discriminate against people who have always been harmed, but also those who have not been traditionally discriminated against. This was not anticipated by lawmakers, that is not how humans work. If algorithms are behaving in ways we did not anticipate and are causing harm, then we need to re-imagine the law.

Could you tell us a bit more about what the Oxford Internet Institute does?

Different communities deal with issues such as cybersecurity, competition aspects, online harm, and protecting children online. My research group, Governance of Emerging Technologies (GET), consists of 11 people working on fairness and bias, and privacy protection, all issues related to governing new technologies. They are computer scientists, political theorists, psychologists, lawyers – a very diverse set of people.

---

[1] The recording of this talk is available at: https://www.youtube.com/watch?v=Z4xA-uMnOKE.
[2] Sandra Wachter is Professor of Technology and Regulation at the Oxford Internet Institute, University of Oxford, where she researches the legal and ethical implications of AI, Big Data, and robotics as well as Internet and platform regulation. Her latest paper, upon which this talk is based, can be downloaded here: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4099100

Do you think that besides updating EU anti-discrimination law, there needs to be awareness raising among the public about the importance of the issue? How can academia contribute?

Awareness is extremely important. Both academics and the general public might not know that this type of decision-making is happening. Some practical examples: if we called attention to the fact that you are not getting into university because you have no dog or not getting a job because you are using Chrome, there would be an immediate reaction that "it's not fair". Nobody will feel unpassionate about this development. We need to act now because there is still time to have a discussion whether this is acceptable, while we are still grappling with traditional forms of discrimination.

Are there related trends that you see emerging, maybe a future worst-case scenario?

We should get ahead of the curve with algorithmic accountability in order to avoid opaque decision-making where nobody understands the rules behind it. For example, if you wanted to prepare your children for a good future, you would tell them to go to school and have good grades, write good applications, and reach out to reference letter writers. We know those are the criteria for applying to jobs or universities. But how am I supposed to prepare my children if the criteria are constantly changing? The unpredictability and uncertainty are the most worrying. Worst-case, we end up in a Kafkaesque situation where decisions are premade according to unknown rules.

Is there something else you would like to add?

Yes: I believe that the solution is to find a good middle ground. I don't want to tell companies or public entities how to make decisions. It is not about dictating how to make laws. Organisations should assess for themselves whether their selection criteria are actually good proxies. This is not only about equality and justice but also an incentive to make better decisions, just by putting effort into re-evaluating the criteria used.

FRONTIER TALK

Memo by Marina Sanchez Del Villar

On 17 October 2022 Professor Sandra Wachter spoke about her work on algorithmic discrimination, fairness, and artificial intelligence (AI). She argued that algorithmic groups should be protected by non-discrimination law and discussed how this could be achieved.

Algorithmic groups

Artificial intelligence is increasingly used to make life-changing decisions, including about who is successful with a job application and who gets into university.

To make these decisions, AI algorithms are banned from using protected attributes under antidiscrimination law, such as gender or race. Nevertheless, there have been cases where algorithms were found using these attributes to inform their decisions. For example, during 2020 the United Kingdom used an algorithm to predict the A-level scores of students

graduating from high school, which assigned disproportionately lower scores to students of colour. Or, the automated moderation tool from Google rates people of colour and the LQBTQ+ community more often as *toxic.*

AI, however, often creates groups that have not previously been used by humans and hence are not covered by non-discrimination law (e.g., 'dog owners' or 'sad teens'). Some of these groups may even be incomprehensible to humans (e.g., people classified by how fast they scroll through a page or the electronic signals they send). These are what Wachter calls *algorithmic groups*. Understanding how to regulate the use by AI of these new groups is important because decisions based on algorithmic groups can be harmful.

For example, if a loan applicant scrolls through the page quickly or uses only lower caps when filling out the form, their application is more likely to be rejected.[3] Job applicants who use browsers such as Microsoft Explorer or Safari, instead of Chrome or Firefox, are less likely to be successful.[4] Categorisation by an algorithm can be socially penalising: in China being a video gamer can lower your social score.

Non-discrimination law aims to protect against certain types of harms, such as equal access to employment, goods, and services, but it has never protected 'fast scrollers' or 'Safari users'. Granting these algorithmic groups protection will be challenging, as historically the European Court of Justice has remained reluctant to extend the law to cover new groups.

What constitutes a protected group?

The natural question, therefore, is what makes a group worthy of protection. Currently, there is no coherent framework to study this question. Wachter offers a taxonomy to best summarise the state of the literature.

1. Immutability and choice. Does the subject have the choice to change the attributes? Immutability of an attribute signals that it should be protected under antidiscrimination law. Some sensitive choices, like religion, are also protected.

2. Relevance, arbitrariness, and merit. Sex or race should not be relevant for a loan repayment, but AI/machine learning finds correlations and patterns and can make everything relevant. Hence, allowing algorithmic groups to be included in decision-making because of their 'relevance' would de facto mean disabling antidiscrimination law.

3. Historical oppression, stigma, and structural disadvantage. Traditional protected groups experience ongoing oppression, while algorithmic groups (e.g., IE users, dog owners) do not experience comparable stigmas.

4. Social saliency and communal and cultural identity. Generally, being part of a group is important either for the individual or the society. Many of the attributes that define algorithmic groups are not socially salient (how fast you browse, how fast you move

[3] https://money.cnn.com/2013/08/26/technology/social/facebook-credit-score/index.html
[4] https://www.economist.com/the-economist-explains/2013/04/10/how-might-your-choice-of-browser-affect-your-job-prospects

the retina). It is therefore hard to make the case about social saliency to protect these groups under the law.

According to these criteria, algorithmic groups do not meet the threshold to be protected under antidiscrimination law.

Why is discrimination wrong?

Although algorithmic groups are not protected under current discrimination law, we know that AI uses these groups to discriminate. Is there a case to be made about discrimination in the context of AI decision making being wrong? If so, then it would be equally immoral to discriminate based on algorithmic group than it is to discriminate based on traditional groups.

Wachter's taxonomy of what makes discrimination wrong includes:

1. Assumption of moral superiority and stigmatisation. This may be described as demeaning an individual, considering the discriminated as of lower moral value or treating them with disrespect, stigmatising or subjecting them to subordinating acts.

2. The existence of a power difference.

AI developers are simply optimising a process, without any moral stance about who is included in their groups. Additionally, as the groups can be constantly changing, the same individual is not constantly under power imbalance.

Therefore, according to the above tests for undesirable discrimination, algorithmic groups do not meet the criteria to be considered protected groups.

What is the aim of antidiscrimination law?

A third attempt to fit algorithmic groups inside the antidiscrimination law considers the aim of such law. If algorithmic groups generate harm to consumers in the same way as traditionally protected groups, then there should be room to include them under antidiscrimination law.

According to Wachter, the aims of antidiscrimination law include:

1. Progression towards substantive equality among groups.

2. The attainment of equal opportunity between individuals.

3. The eradication of lingering effects of past discrimination.

4. The end of oppression and domination.

5. The combat of systemic subordination and stigmatisation.

When it comes to algorithmic groups, there is no identifiable group consistently disadvantaged more than others, as the groups and their composition can be constantly changing. Because it is unlikely that there is one group that is consistently benefiting with respect to the others, algorithmic groups would not qualify for protection under antidiscrimination law, as defined by its aims.

A new theory of harm

Wachter suggests that we think about a new theory of harm from algorithmic groups.

We can break the law down into the basic goals it wishes to attain: liberties, freedoms, or access to basic goods and services (such as education, pursuit of a profession, or healthcare).

The law anticipates that the person in power can have prejudices. With algorithmic groups, the prejudice might no longer be present, but the harm still is. The harm is the same as that originally imagined in traditional antidiscrimination law, only the mode, perpetrator, and process of bringing about that harm are different.

Wachter proposes five types of artificial immutability. The source is different from that of traditionally protected groups, as algorithmic groups are created by the AI use itself.

1. Opacity: for example, in the context of a loan application, the applicant has no control over which elements are used as eligibility criteria.

2. Vagueness: vague decision criteria allow only very little agency over the process.

3. Instability: the composition of the groups is constantly changing, so an applicant cannot prepare a good portfolio without being fully aware of the criteria. This additionally means that not all applicants are evaluated following the same criteria.

4. Empirical incoherence: AI is getting rid of the causal link between two data points (think about the theoretical connection between liking the colour green and being more likely to repay a loan). This is connected to the relevance question: relevance should not be about being correlated but about having a causal, ethically acceptable, link.

5. Lack of social concept, involuntariness and invisibility: sometimes there is not even a human word to describe the criterium (as in groups defined by the speed of retinal movement).

Wachter's requirements for transparency, stability, and empirical coherence challenge the normative and ethical acceptability of algorithmic groups. They can hence be considered under antidiscrimination law.

What can we do?

Some advocate for making the inclusion criteria in algorithmic groups public. This results into what Wachter calls the transparency fallacy: more transparency *alone* does not solve the problem generated by the algorithmic groups. For example, telling people an individual that face recognition software is being used does not make that person move her eyes slower – in contrast to the more traditional criteria of better grades will get you a seat at law school. The main issue that needs to be addressed is autonomy and power over the process, not just transparency.

Not all types of artificial immutability are problematic. Wachter argues that we need to consider the application to understand if these should be classified as discriminatory under the law. For example, age is generally considered a protected attribute, but in a labour context anti child labour laws that use age as an exclusionary criterion are acceptable because it is important to protect children. We need to have a similar set-up for algorithmic groups.