AI and Fairness

Katja Maria Vogt

ValuesLab

Plan for Today

- Intro: Why fairness?
- An example: AI in job interviews
- What is fairness, as this notion is understood in the context of AI?
- How does fairness relate to other values?

Introduction: Why Fairness?

In discussions of AI and ethics, no value has received as much attention as fairness. Why?

Consider an Example

Example: An AI system is used to decide who among the applicants for a job get a first-round interview.

The AI assesses candidates w.r.t. a range of values: relevant experience, predicted productivity, creativity, collegiality, etc.

And now the question is: is the AI fair?

Socratic Premise

Socratic Premise: We need to know *what X is* in order to assess whether some Y is X.

Applied to our example: We need to know what fairness is in order to assess whether a given AI system is fair.

What is Fairness?

However, the question “what is fairness?” may be too large for current purposes…

Can we make the question more tractable by making it more specific?

Can we Make the Question More Specific?

Perhaps there are distinctive kinds of fairness: in games, in political institutions, in family life, and so on, and in AI.

We might ask: what is fairness in AI?

In other words: what is algorithmic fairness?

How Specific?

Can we make the question even more specific? Consider use-cases:

- What is fairness in AI w.r.t. job applications?
- What is fairness in AI in finance?
- What is fairness in AI in medicine?
- Etc.

Advantages, Disadvantages of Domain-Specific Notions

Advantage: Domain-specific notions of algorithmic fairness can capture characteristics of specific domains.

Disadvantage: If there's nothing in common, why talk about "algorithmic fairness," rather than "getting it right" in a domain?

Common Core? A Preliminary Gloss

Intuitively speaking, AI researchers who work toward “fair AI” seem to be concerned with the elimination of bias and discrimination.

Can we make this more precise?

Legal Ramifications

Perhaps the intuitions about eliminating bias/discrimination can be spelled out by appealing to relevant laws?

Legal Doctrines, Drago Plečko and Elias Bareinboim (2024)

Disparate treatment doctrine:

- "enforces the equality of treatment of different groups, prohibiting the use of the protected attribute (e.g., race) in the decision process." (3)

- "the disparate treatment doctrine ensures that there is no direct effect of the protected attribute on the outcome, which can be seen as the minimal fairness requirement." (4)

Continued, Drago Plečko and Elias Bareinboim (2024)

Disparate impact doctrine:

- “disparate impact doctrine focuses on outcome fairness, namely, the equality of outcomes among protected groups.” (4)
- “the disparate impact doctrine (in the extreme case), ensures that the protected attribute has no effect on the outcome.” (4)

Defining Algorithmic Fairness via the Law?

Should we define algorithmic fairness as compliance with law regarding anti-discrimination, disparate treatment, and disparate impact?

The Law is Broader

Here's a concern: the law is not specifically about AI.

That is, if we appeal to relevant legal doctrines, we need to *apply* them—precisify them, make them implementable—for AI.

Applying the Law to AI: Zemel (2013)

Group fairness, also called statistical parity ensures that “the overall proportion of members in a protected group receiving positive (negative) classification are identical to the proportion of the population as a whole.”

Zemel (2013), continued

Group fairness is important, but not enough.

It can lead to unfairness towards individuals:

“such as discriminating in employment while maintaining statistical parity among candidates interviewed by deliberately choosing unqualified members of the protected group to be interviewed in the expectation that they will fail.” (p.1)

Zemel (2013), continued

Individual fairness ensures that “any two individuals who are similar with respect to a particular task should be classified similarly.” (p.1)

Is Fairness Enough?

When philosophers discuss how people are treated, individually and as groups, fairness is just one of the values they invoke.

They also invoke justice, equality, autonomy, respect, and more.

Justice and Fairness

What is the difference between wanting an AI model to be just, and wanting it to be fair?

For reasons that aren’t entirely obvious, AI researchers tend to speak of fairness, not of justice.

Justice as More Comprehensive

Vredenburgh (2022) makes a proposal about the way the terms are used in research on AI: justice includes, but is more than fairness; AI fairness is, roughly, what the law demands.

Vredenburgh (2022)

Suppose that fairness is about complying with the law.

Laws typically undergo a process of revision; they can be imperfect, in need of updating due to historical change, unjust in some respects.

If the law is in some respects unjust, and if fairness is compliance with the law, fairness and justice come apart.

Vredenburgh (2022)

Fairness has little moral worth without just institutions: “without just institutions, fairness is of little moral worth. The value of fairness depends on the existence of just institutions in the background.” (138)

Take-Aways and Questions

- What is a compelling definition of fairness in AI?
- Is fairness enough? How does it relate to the law? To justice?
- What about other values? For example, we want an AI to be both fair and accurate. Can fairness and accuracy come apart?

Readings

- Drago Plečko and Elias Bareinboim (2024), “Causal Fairness Analysis”, *Foundations and Trends in Machine Learning*: Vol. 17, No. 3, 1–238.

- Kate Vredenburgh, “Fairness,” in Justin Bullock (ed.), *The Oxford Handbook of AI Governance*, OUP 2022.

- Richard Zemel et al, “Learning Fair Representations” (2013).