Performativity and Prospective Fairness

Sebastian Zezulka, Konstantin Genin

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Look up our paper.

Research Group: Epistemology and Ethics of Machine Learning

Konstantin Genin
Group Leader

Inherent complexity of scientific problems. Interactions between morals and methodology.

Raysa Benatti
PhD student

AI and statistical evidence in legal settings.

Mykhailo Bogachov
Visiting PhD. Student

Normative implications of performativity in AI systems. Effects of LLMs on moral reasoning.

Sebastian Zezulka
PhD. Student

Dynamical perspectives on algorithmic fairness.

The Fundamental Question of Fair Machine Learning

Will machine learning algorithms will **reproduce** or **exacerbate** the structural inequalities reflected in their training data?

The Fundamental Question of Fair Machine Learning

Will machine learning algorithms will **reproduce** or **exacerbate** the structural inequalities reflected in their training data?

Widely cited as the ***motivation*** for AI fairness. However, the methodological solutions developed by researchers in algorithmic fairness are, surprisingly, **ill-suited** for answering this fundamental question.

The Generic Risk-Assessment Setting

Machine learners are concerned with learning a function that takes as **input**

- some features X,
- and a sensitive attribute A

and **outputs** a score R valuable for predicting an outcome Y.

The score R is meant to inform some important decision D that, typically, is causally relevant for the outcome Y.

Risk Assessment in Public Employment

The algorithm takes as **input**

- the education and employment history (X),
- and gender (A)

of a recently unemployed person, and **outputs** a risk score (*R*) of long-term unemployment (*Y*).

On the basis of the risk score (*R*), a case-worker allocates the person to some labor-market program (*D)* that is causally relevant for their employment prospects *(Y)* .

Risk Assessment in Public Employment

The risk score may support a number of different policies. For example,

- In Flanders: individuals at high risk of long-term unemployment are **prioritized —** they are contacted first by the public employment service (Desiere and Struyven, 2020).

- In Austria: risk scores classify the recent unemployed into (i) those with good prospects in the next six months; (ii) those with bad prospects in the next two years; and (iii) everyone else. Support measures **target** the third group. while offering **only limited support** to the first and second group (Allhutter et al., 2020).

Doris Allhutter, Florian Cech, Fabian Fischer, Gabriel Grill, and Astrid Mager. Algorithmic profiling of Job Seekers in Austria: How Austerity Politics Are Made Effective. Frontiers in Big Data, 3, 2020.

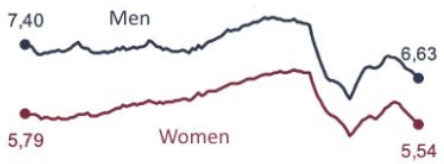
Sam Desiere and Ludo Struyven. Using Artificial Intelligence to classify Jobseekers: The Accuracy-Equity Trade-off. Journal of Social Policy, 50(2):367–385, 2020.

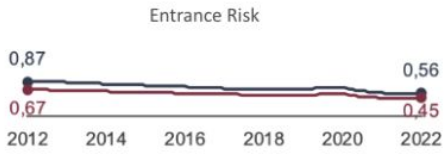
Risk Assessment in Public Employment

Advocates of the Austrian policy argue in terms of *efficiency*. Critics worry about exacerbating long-standing structural inequalities in the labor market.

The Gender Re-employment Gap

Risk of Entry into and Probability of Exit from Unemployment
Running annual average December 2012 to December 2022 in Percent
Germany



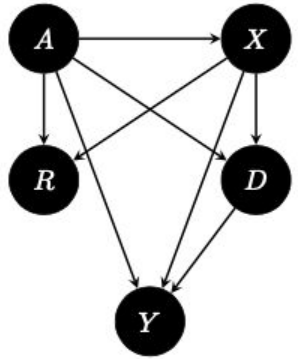


Source: Statistik der Bundesagentur für Arbeit

Algorithmic Fairness to The Rescue?

Almost all approaches in the fairness literature are **retrospective.**

- **Group-based** fairness proposals (Barocas et al., 2023) require that certain conditional independencies hold in the **training** distribution.
- **Causal** fairness proposals (Kilbertus et al., 2017) require that certain properties are satisfied by the **causal structure** giving rise to the **training distribution.**
- **Individual** fairness proposals (Dwork et al, 2012) require that similar people have similar **distributions** of outcomes (in the training distributions).



(a) Causal structure G_{pre} *before* deploying an algorithmically-informed policy.

Algorithmic Fairness to The Rescue?

All these approaches in the fairness literature are **retrospective.** They all agree that fairness is a constraint

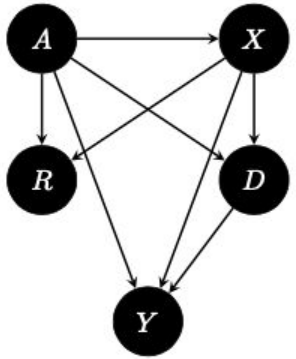
$$\Phi(P_{\text{pre}}, G_{\text{pre}}, M),$$

where M is a measure of similarity and

– P_{pre} is the joint distribution of (A,X,R,D,Y) and

— G_{pre} is the causal structure generating P_{pre}

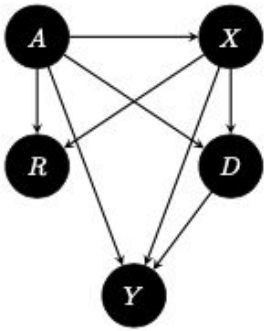
before the algorithm is deployed!



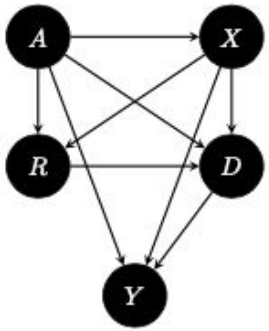
(a) Causal structure G_{pre} *before* deploying an algorithmically-informed policy.

Prospective Fairness

We claim that this approach is misguided: **prospective** fairness is a matter of **comparing** $\Phi(P_{\text{pre}}, G_{\text{pre}}, M)$ and $\Phi(P_{\text{post}}, G_{\text{post}}, M)$.



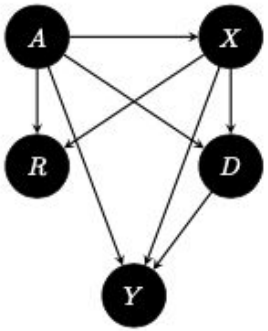
(a) Causal structure G_{pre} *before* deploying an algorithmically-informed policy.



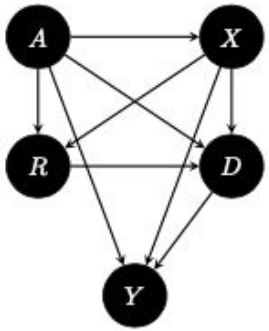
(b) Causal structure G_{post} *after* deploying an algorithmically-informed policy.

Prospective Fairness

E.g., comparing the gender re-employment gap **before** deployment with the gender re-employment gap **likely to be induced** by the algorithmic policy.



(a) Causal structure G_{pre} *before* deploying an algorithmically-informed policy.



(b) Causal structure G_{post} *after* deploying an algorithmically-informed policy.

Why isn't retrospective fairness enough?

In the worst case, retrospective fairness is **self-defeating.** Mishler and Dalmasso (2020) show that satisfying group-based fairness notions at the time of training **virtually ensures** that they will be violated after deployment.

Alan Mishler and Niccolò Dalmasso (2022) *Fair When Trained, Unfair When Deployed: Observable Fairness Measures are Unstable in Performative Prediction Settings.*

Why isn't retrospective fairness enough?

Group-based notions of fairness fall victim to **performativity**: the tendency of an algorithmic policy intervention to shift the distribution away from the one on which it was trained (Perdomo et al., 2020).

But they are undermined not by an **unintended** and **unforeseen** performative effect, but by the **intended**, and **foreseen** shift in distribution induced by algorithmic support, i.e. by the fact that :

$$P_{pre}(D \mid A, X, R) \neq P_{post}(D \mid A, X, R).$$

Juan Perdomo, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. *Performative prediction*. In International Conference on Machine Learning, pages 7599–7609. PMLR,

Why isn't retrospective fairness enough?

In the long-run, retrospective (group based) fairness constraints can **entrench** systemic inequality (D'Amour et al., 2020).

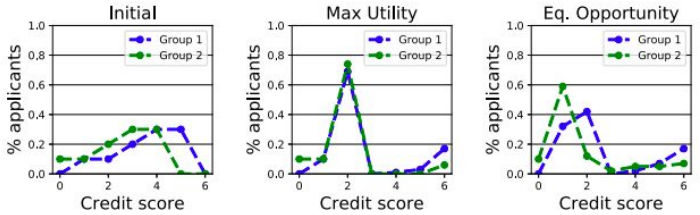


Figure 2: Initial credit score distributions of the two groups (far left) and final states after 20K steps of the environment using a max-util agent (center) and EO agent (right). The credit distributions start with group 2 slightly disadvantaged, but the groups converge to the similar distributions under the max-util agent, while the EO agent maintain unequal credit distributions between groups.

Alexander D'Amour, Hansa Srinivasan, James Atwood, Pallavi Baljekar, D. Sculley, and Yoni Halpern. *Fairness Is Not Static: Deeper Understanding of Long Term Fairness via Simulation Studies.* In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, FAT* '20, page 525–534

Prospective Fairness

But how do we estimate the relevant measures of ***post-deployment*** inequality from ***pre-deployment data?*** Especially in the face of performative effects?

Prospective Fairness

But how do we estimate the relevant measures of ***post-deployment*** inequality from ***pre-deployment data?*** Especially in the face of performative effects?

Introduce potential outcome random variables Y^d, which take the value of the outcome under assignment to program d.

Prospective Fairness: Identifiability

Theorem (Identification of $P_{\text{post}}(Y = y \mid A = a)$)

Suppose that CONSISTENCY, UNCONFOUNDEDNESS, NO UNPRECEDENTED DECISIONS, STABLE CATE and NO FEEDBACK hold. Suppose also that $P_{\text{post}}(A = a) > 0$. Then, $P_{\text{post}}(Y = y \mid A = a)$ is given by

$$\sum_{(x,d)\in\Pi_{\text{post}}} P_{\text{pre}}(Y = y \mid A = a, X = x, D = d) P_{\text{pre}}(X = x \mid A = a) \underbrace{P_{\text{post}}(D = d \mid A = a, X = x)}_{\text{ALGORITHMIC EFFECT}},$$

where $\Pi_t = \{(x, d) \in \mathcal{X} \times \mathcal{D} : P_t(X = x, D = d \mid A = a) > 0\}$.

Sebastian Zezulka, Konstantin Genin (2023). *Performativity and Prospective Fairness.* Accepted: Algorithmic Fairness Through the Lens of Time, NeurIPS Workshop.

Prospective Fairness: The Assumptions

$$Y = \sum_{d \in \mathcal{D}} Y^d \mathbb{1}[D = d]. \qquad \text{CONSISTENCY}$$

$$Y^d \perp\!\!\!\perp_t D \mid A, X. \qquad \text{UNCONFOUNDEDNESS}$$

$$P_{\text{pre}}(D = d \mid A = a, X = x) > 0 \text{ if } P_{\text{post}}(D = d \mid A = a, X = x) > 0 \qquad \text{NO UNPRECEDENTED DECISIONS}$$

$$P_{\text{pre}}\left(Y^d \mid A = a, X = x\right) = P_{\text{post}}\left(Y^d \mid A = a, X = x\right) \qquad \text{STABLE CATE}$$

$$P_{\text{pre}}\left(A = a, X = x\right) = P_{\text{post}}\left(A = a, X = x\right) \qquad \text{NO FEEDBACK}$$

- General *ceteris paribus* assumption: We want to isolate the effect of the policy.
- Assumptions do **not** rule out the intended direct effect of the algorithm on the decisions.

$$P_{\text{pre}}\left(D = d \mid A = a, X = x\right) \neq P_{\text{post}}\left(D = d \mid A = a, X = x\right) \qquad \text{ALGORITHMIC EFFECT}$$

Prospective Fairness: The Paper

Thank You!

Sebastian Zezulka, Konstantin Genin (2023). *Performativity and Prospective Fairness.* Accepted: Algorithmic Fairness Through the Lens of Time, NeurIPS Workshop.