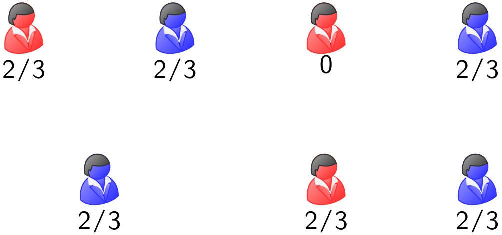
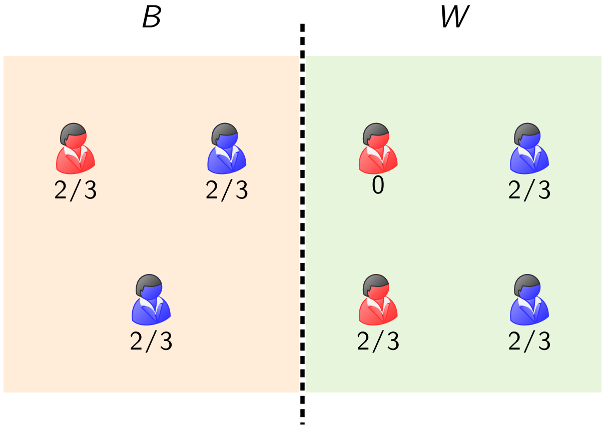
PHPE 308M/PHIL 209F
Fairness

Eric Pacuit, University of Maryland

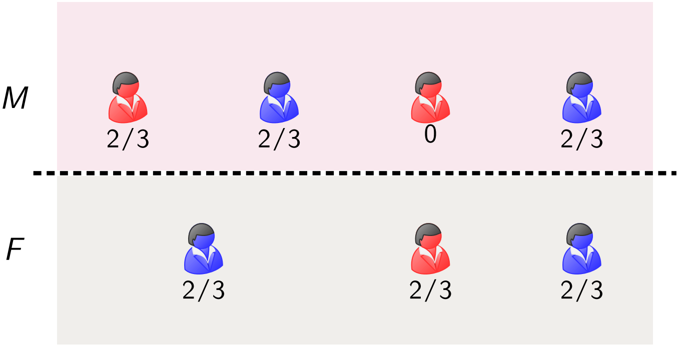
December 1, 2025

Rush T. Stewart (2022). *Identity and the limits of fair assessment*. Journal of Theoretical Politics, 34(3), pp. 415 - 442.

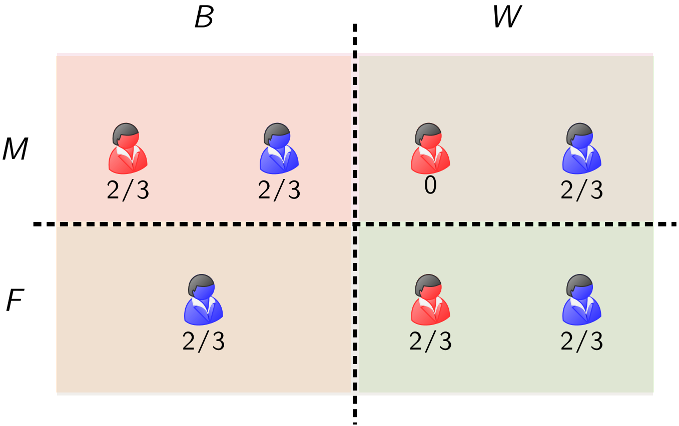

2/3
2/3
0
2/3
2/3
2/3
2/3



Does the algorithm satisfy (weak) calibration for the groups B and W?



Does the algorithm satisfy (weak) calibration for the groups M and F?



Does the algorithm satisfy (weak) calibration for the groups $B\&M$, $B\&F$, $W\&M$, and $W\&F$?

Setup

A single property y of interest.

Individuals in N either have property y or lack it: $Y : N \rightarrow \{0, 1\}$

Call a function $h : N \rightarrow [0, 1]$ an assessor.

For concreteness, interpret $h(i)$ as the assessor's probability that i has property y.

Setup

The quantity $P(Y = 1) = \mu$, for example, is the proportion of people in N that have property y, the prevalence of y in the population.

Call μ the base rate for y in N.

Given a partition $\pi = \{G_1, \ldots, G_m\}$ of N, let $P_k = P(\cdot \mid G_k)$. So, $P_1(Y = 1) = \mu_1$ is the base rate for y in group 1 is μ_1 and $P_2(h = 0.5)$ is the proportion of people to which h assigns 0.5 in G_2, and so on.

Strong Calibration

An assessor is **(strongly) calibrated** if

$$P_k(Y = 1 \mid h = p) = p \text{ for all } p \in [0, 1] \text{ and } k = 1, 2, \ldots, m \text{ such that } P_k(h = p) > 0.$$

Strong Calibration

An assessor is **(strongly) calibrated** if

$$P_k(Y = 1 \mid h = p) = p \text{ for all } p \in [0, 1] \text{ and } k = 1, 2, \ldots, m \text{ such that } P_k(h = p) > 0.$$

E.g., consider weather forecasting. Suppose that each day, a forecaster announces a probability of rain for that day. The forecaster is calibrated if it rains on 10% of the days she announces that it will rain with probability 0.1, and it rains on 85% of the days she predicts rain with probability 0.85, etc.

Limitation Result

An assessor is **perfect** if $h(i) = Y(i)$ for all $i \in N$.

Limitation Result

An assessor is **perfect** if $h(i) = Y(i)$ for all $i \in N$.

Observation 1. Let h be an assessor for N. The following are equivalent:

1. h is calibrated for all binary partitions.
2. h is calibrated for all partitions.
3. h is perfect.

Limitation Result

An assessor is **perfect** if $h(i) = Y(i)$ for all $i \in N$.

Observation 1. Let h be an assessor for N. The following are equivalent:

1. h is calibrated for all binary partitions.
2. h is calibrated for all partitions.
3. h is perfect.

In other words, outside of the unrealistic case of perfect assessment, there will be bias in confidence against some group. Observation 1 complicates any automatic inference from failure of calibration for some group to *intentional* bias on behalf of the assessor.

Weak Calibration

An assessor h satisfies **weak calibration for groups** for a partition π if

$$P_k(Y = 1 \mid h = p) = P_j(Y = 1 \mid h = p) \text{ for all } G_k, G_j \in \pi$$

Put another way, among people assigned the same assessment score, the proportion of people who have property y is the same across all groups in the partition.

Limitation Result 2

An assessor h makes **perfect distinctions** if, for all $i, j \in N$, $Y(i) \neq Y(j)$ implies that $h(i) \neq h(j)$. So, for any score p, if $h(i) = p$ and $Y(i) = 1$, then for no individual j such that $Y(j) = 0$ is it the case that $h(j) = p$.

Limitation Result 2

An assessor h makes **perfect distinctions** if, for all $i, j \in N$, $Y(i) \neq Y(j)$ implies that $h(i) \neq h(j)$. So, for any score p, if $h(i) = p$ and $Y(i) = 1$, then for no individual j such that $Y(j) = 0$ is it the case that $h(j) = p$.

Observation 2. Let h be an assessor for N. The following are equivalent:

1. h satisfies predictive equity for all binary partitions.
2. h satisfies predictive equity for all partitions.
3. h makes perfect distinctions.

Limitation Result 2

An assessor h makes **perfect distinctions** if, for all $i, j \in N$, $Y(i) \neq Y(j)$ implies that $h(i) \neq h(j)$. So, for any score p, if $h(i) = p$ and $Y(i) = 1$, then for no individual j such that $Y(j) = 0$ is it the case that $h(j) = p$.

Observation 2. Let h be an assessor for N. The following are equivalent:

1. h satisfies predictive equity for all binary partitions.
2. h satisfies predictive equity for all partitions.
3. h makes perfect distinctions.

Aside from assessors that make perfect distinctions, scores will not “mean” the same thing for all groups; there will be bias against some group. In large populations, perfect distinctions is very difficult to achieve—not as difficult as perfect assessment, but difficult nonetheless.

Two Objections

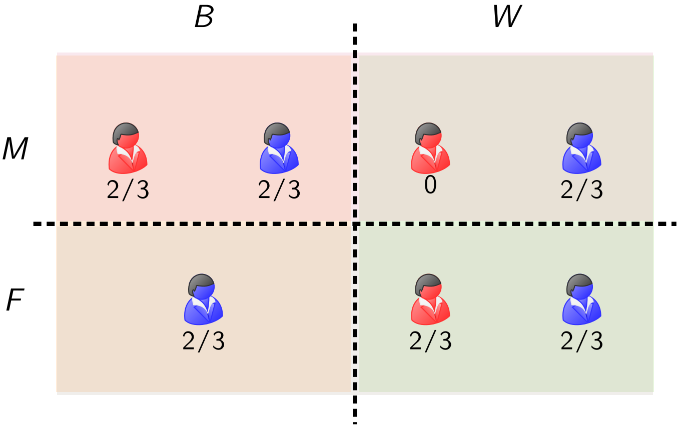
1. We might consider satisfying certain fairness constraints *approximately* rather than exactly. That is, we could confine the amount of bias to which any group is subject to a certain margin of tolerance.

Two Objections

1. We might consider satisfying certain fairness constraints *approximately* rather than exactly. That is, we could confine the amount of bias to which any group is subject to a certain margin of tolerance.

2. One might be inclined to think that, while (a particular type of) unbiased assessment for multiple partitions is often desirable, we have overshot the mark by requiring it for *all* partitions.

There are simple examples of populations that allow for a imperfect assessor that is simultaneously calibrated for, say, two different non-trivial ways of partitioning the population.



The algorithm is calibrated for both the $\{B, W\}$ and $\{M, F\}$ partitions.

- One one hand, requiring the satisfaction of a fairness constraint for some single partition is generally unsatisfactory since we may care about the fair treatment of groups from different partitions.

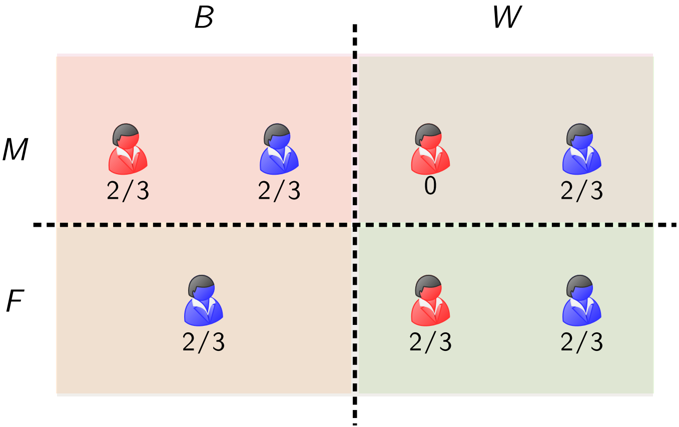
- One one hand, requiring the satisfaction of a fairness constraint for some single partition is generally unsatisfactory since we may care about the fair treatment of groups from different partitions.

- On the other hand, requiring any of the fairness constraints considered here be satisfied for *all* partitions of the population or all partitions of some cardinality places unrealistically high demands on assessment.

We cannot insist on any notion of statistical fairness for every subgroup of the population: for example, any imperfect classifier could be accused of being unfair to the subgroup of individuals defined ex-post as the set of individuals it misclassified. This simply corresponds to 'overfitting' a fairness constraint.

Michael Kearns, Seth Neel, Aaron Roth, Zhiwei, and Steven Wu (2018). *Preventing fairness gerrymandering: Auditing and learning for subgroup fairness*. In: Proceedings of the 35th International Conference on Machine Learning, Volume 80, Stockholm, Sweden, pp. 2564 - 2572. PMLR.

Intersectional Bias



Although the algorithm is calibrated for both the $\{B, W\}$ and $\{M, F\}$ partitions, it is **not** calibrated for the $\{B\&M, B\&F, W\&M, W\&F\}$ partition.

Intersectionality

Kimberlé Crenshaw, who introduced the term "intersectionality," makes use of a court case to explain how bias against black women, for example, is consistent with the lack of that form of bias against black people or against women.

Intersectionality

Kimberlé Crenshaw, who introduced the term “intersectionality,” makes use of a court case to explain how bias against black women, for example, is consistent with the lack of that form of bias against black people or against women.

In *DeGraffenreid v. General Motors*, five black women alleging discrimination by General Motor’s seniority-based system sued the company. Prior to 1964, General Motors did not hire black women. All of the black women hired after 1970 lost their jobs through a seniority-based layoff during a later recession.

K. Crenshaw (1989). *Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics*. University of Chicago Legal Forum 1989(Article 8): 139-167.

Intersectionality

The district court rejected the plaintiffs' attempt to bring a suit on behalf of black women in particular rather than on behalf of black people or women. According to the court, the suit must present "a cause of action for race discrimination, sex discrimination, or alternatively either, but not a combination of both".

Intersectionality

The district court rejected the plaintiffs' attempt to bring a suit on behalf of black women in particular rather than on behalf of black people or women. According to the court, the suit must present "a cause of action for race discrimination, sex discrimination, or alternatively either, but not a combination of both".

The court noted that, while General Motors did not hire black women prior to 1964, they did hire female employees for a number of years prior to 1964. So there was no sex discrimination.

Intersectionality

And what if General Motors had hired black people—specifically black men—for a number of years prior to 1964? Crenshaw's point is that that would not really absolve General Motors of the charge of discrimination against black women. It certainly does not follow that there could be no discrimination against black women.

Concluding Remarks

There are multiple ways to carve a population, multiple social identities, for which it may be important to avoid biased assessments. Fixing a single partition of identities is overly restrictive, committing us to ignoring both relevant forms of bias against other groups and changing social context. Allowing even a set of partitions to ossify into *the* relevant partitions may fail to make us sufficiently attentive.

Concluding Remarks

There are multiple ways to carve a population, multiple social identities, for which it may be important to avoid biased assessments. Fixing a single partition of identities is overly restrictive, committing us to ignoring both relevant forms of bias against other groups and changing social context. Allowing even a set of partitions to ossify into *the* relevant partitions may fail to make us sufficiently attentive.

Where does this leave us? What the foregoing analysis helps us to make clear is that, not only is there a conflict between eliminating different forms of bias, but there are serious limits to the extent to which a given form of bias can be eliminated across different partitions.

Lily Hu. *Does calibration mean what they say it means; or, the reference class problem rises again*. forthcoming in Philosophical Studies.

Calibration and the Same Meaning Argument

Risk scores that are calibrated within groups ensure that a model is, on average, equally well-fit to those groups vis-a-vis the outcome that it predicts.
This property is often glossed in the literature as ensuring that scores “mean the same thing” for individuals of different groups.

Calibration and the Same Meaning Argument

Risk scores that are calibrated within groups ensure that a model is, on average, equally well-fit to those groups vis-a-vis the outcome that it predicts. This property is often glossed in the literature as ensuring that scores “mean the same thing” for individuals of different groups.

Hellman: “If a high-risk score means something different for blacks than for whites, then we do not know whether to believe (or how much confidence to have) in the claim that a particular scored individual is likely to commit a crime in the future.”

D. Hellman (2020). *Measuring algorithmic fairness*. Virginia Law Review 106, no. 4, 811866.

Calibration and the Same Meaning Argument

> Calibration is intuitively compelling and easily motivated. If it were violated by some algorithm, that would mean that the same risk score would have different evidential import for the two groups. Our probability that an individual is positive, given that they received a given risk score, would have to be different depending on the group to which the individual belongs. A given risk score, intended to be interpreted probabilistically, would in fact correspond to a different probability of being positive, depending on the individual's group membership. This seems to amount to treating individuals differently in virtue of their differing group membership. (Hedden)

The claim is that the fact that the algorithmically-output risk scores are (mis)calibrated within some group (e.g., race) speaks to the “meaning” or evidential value of particular individuals’ scores.

The claim is that the fact that the algorithmically-output risk scores are (mis)calibrated within some group (e.g., race) speaks to the “meaning” or evidential value of particular individuals’ scores.

The problem is that proponents of the Same Meaning picture provide no argument to back this inference from group probabilistic fact to individual probability.

They simply *assume* that the groups within which calibration is satisfied are indeed those groups that grant this inference.

Reference Class Problem

Alan Hájek (2007). *The reference class problem is your problem too*. Synthese, 156, pp. 563-585.

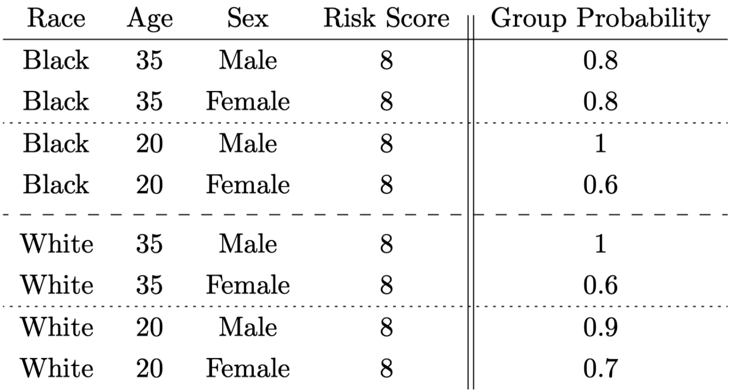
John Venn (1888). *The logic of chance: an essay on the foundations and province of the theory of probability, with especial reference to its logical bearings and its application to moral and social science, and to statistics*. Macmillan.

Hans Reichenbach (1971). *The theory of probability*. University of California Press.

The problem for the Same Meaning picture is that interpreting the evidential value of a given individual's score and comparing individuals' scores requires figuring *which* groupings and accordingly *which* calibration facts should apply for each.

The problem for the Same Meaning picture is that interpreting the evidential value of a given individual's score and comparing individuals' scores requires figuring *which* groupings and accordingly *which* calibration facts should apply for each.

When an individual belongs to many groups—when they are not only Black but also male and also 35 years-old and so also a 35 year-old Black male—it becomes clear that calibration within groups can only speak to what their score means when it picks out the right group for the individual: that is, the right reference class.



Race	Age	Sex	Risk Score	Group Probability
Black	35	Male	8	0.8
Black	35	Female	8	0.8
Black	20	Male	8	1
Black	20	Female	8	0.6
White	35	Male	8	1
White	35	Female	8	0.6
White	20	Male	8	0.9
White	20	Female	8	0.7

Does the score of 8 assigned to Jamal, a 35 year-old Black male, on average "mean the same thing" as a score of 8 assigned to Emily, a 20 year-old white female?

Race	Age	Sex	Risk Score	Group Probability
Black	35	Male	8	0.8
Black	35	Female	8	0.8
Black	20	Male	8	1
Black	20	Female	8	0.6
White	35	Male	8	1
White	35	Female	8	0.6
White	20	Male	8	0.9
White	20	Female	8	0.7

- Black, score of 8 $\Rightarrow$ 80%; White, score of 8 $\Rightarrow$ 80%

Race	Age	Sex	Risk Score	Group Probability
Black	35	Male	8	0.8
Black	35	Female	8	0.8
Black	20	Male	8	1
Black	20	Female	8	0.6
White	35	Male	8	1
White	35	Female	8	0.6
White	20	Male	8	0.9
White	20	Female	8	0.7

- Black, score of 8 ⇒ 80%; White, score of 8 ⇒ 80%
- 35 year-old, score of 8 ⇒ 80%; 20 year-old, score of 8 ⇒ 80%.

Race	Age	Sex	Risk Score	Group Probability
Black	35	Male	8	0.8
Black	35	Female	8	0.8
Black	20	Male	8	1
Black	20	Female	8	0.6
White	35	Male	8	1
White	35	Female	8	0.6
White	20	Male	8	0.9
White	20	Female	8	0.7

- Black, score of 8 ⇒ 80%; White, score of 8 ⇒ 80%
- 35 year-old, score of 8 ⇒ 80%; 20 year-old, score of 8 ⇒ 80%.
- **Male, score of 8 ⇒ 92.5%; Female, score of 8 ⇒ 67.5%.**

Calibration supposedly speaks to what scores mean, but what does Emily's score of 8 mean? In other words, given that her risk score is 8, what we should take her individual probability to be?

- If we take Emily as a white person with a score of 8, we may rationally infer that she has an 80% chance of the outcome.

- Taking her as a 20 year-old with a score of 8 also suggests an 80% chance.

- But if we take Emily as someone sexed female assigned an 8, we would rationally infer that she has a 67.5% chance.

- And if we take her as someone who is white, female, 20 years-old, and assigned a score of 8, then we would take her individual probability to be 70%.

- If we take Emily as a white person with a score of 8, we may rationally infer that she has an 80% chance of the outcome.
- Taking her as a 20 year-old with a score of 8 also suggests an 80% chance.
- But if we take Emily as someone sexed female assigned an 8, we would rationally infer that she has a 67.5% chance.
- And if we take her as someone who is white, female, 20 years-old, and assigned a score of 8, then we would take her individual probability to be 70%.

The question then is, which grouping, each of which contains Emily as a member, carries the correct probability for her case? This is just the reference class problem.

In sum, the Same Meaning picture of calibration's distinctive normative edge hinges on a solution to the reference class problem, a solution that is presupposed and never explicitly defended.

The inference that calibration within groups ensures that individual scores "mean the same thing" as other scores is only safely drawn on an antecedent determination that that grouping within which scores are calibrated is the *right reference class* for the scores about which one is reasoning.

Multicalibration

Aaron Roth, Alexander Tolbert, and Scott Weinstein (2023). *Reconciling Individual Probability Forecasts*. in Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, pp. 101-110.

Benedikt Höltgen and Robert C Williamson (2023). *On the Richness of Calibration*. in Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, pp. 1124-1138..

As the number of individuals that an algorithm scores multiplies, so does the number of reference class problems and in turn, the number of groups within which scores must be calibrated for the Same Meaning picture to hold true.

As the number of individuals that an algorithm scores multiplies, so does the number of reference class problems and in turn, the number of groups within which scores must be calibrated for the Same Meaning picture to hold true.

That an algorithm might just satisfy calibration within the right set of groups is, to be sure, not impossible, though the *presumption* that it does is highly optimistic, bordering on wishful thinking.

Recall what motivates the Same Meaning picture: we are worried about our ability to properly interpret algorithmically-output risk scores. We worry that scores might systematically “mean different things” for individuals of different groups, and so we worry that we might treat individuals differently by misinterpreting what their individual risk scores mean for their true individual probability.

Recall what motivates the Same Meaning picture: we are worried about our ability to properly interpret algorithmically-output risk scores. We worry that scores might systematically "mean different things" for individuals of different groups, and so we worry that we might treat individuals differently by misinterpreting what their individual risk scores mean for their true individual probability.

But we would not have such worries in the first place if we could derive a solution to the reference class problem. For then we would simply know what individual scores "mean" because we would know what reference classes to use to determine their individual probabilities.

But then the Same Meaning story about calibration's normative significance would ring entirely hollow.

Summary

The Same Meaning picture, the predominant normative argument for calibration as a statistical criterion of algorithmic fairness, depends crucially on an inference **from group probabilities to individual probabilities**.

To make this inference properly is to successfully solve the reference class problem.

Summary

And yet it seems to me still true that the statistical fictions of averages and base rates and risk distributions bear on what it takes to treat real-life persons fairly.

If this is so, then our task is to figure how we might be able to reconcile these two viewpoints: the abstracted statistical view of a group and the concrete view of an actual person.

This is yet another matter of algorithmic fairness on which we are likely to find ourselves pulled in different directions.

S. Lazar and J. Stone (2024). *On the site of predictive justice*. Noûs, 58, pp. 730-754, https://doi.org/10.1111/nous.12477.

The Central Question

"What precisely goes wrong when Machine Learning (ML) goes wrong?"

The Central Question

"What precisely goes wrong when Machine Learning (ML) goes wrong?"

- Outcome Justice: Harms caused by **decisions** informed by ML systems.

The Central Question

“What precisely goes wrong when Machine Learning (ML) goes wrong?”

- Outcome Justice: Harms caused by **decisions** informed by ML systems.
- Predictive Justice: Moral grounds for criticizing the **predictions** themselves.

  Key claim: Predictions can be morally wrong independent of downstream effects.

Two Camps in the Algorithmic Fairness Debate

1. **Maximalists**: “Build just societies, not just algorithms”
   - Formal fairness criteria too abstract
   - Focus on regulation & abolition

2. **Minimalists**: “Optimize for epistemic standards only”
   - Let policymakers handle social impact
   - ML engineers shouldn’t make policy

Two Camps in the Algorithmic Fairness Debate

1. **Maximalists**: “Build just societies, not just algorithms”
   - Formal fairness criteria too abstract
   - Focus on regulation & abolition
2. **Minimalists**: “Optimize for epistemic standards only”
   - Let policymakers handle social impact
   - ML engineers shouldn’t make policy

The Paper’s Position: Predictive justice exists as a distinct site of moral concern, but it is **situated** (grounded in real social context), not abstract.

Prioritarian Performance Principle (PPP)

PPP: A model is **predictively just** only if its performance for systematically disadvantaged groups cannot be improved without a disproportionate decline in its performance for systematically advantaged groups.