PHPE 308M/PHIL 209F
Fairness

Eric Pacuit, University of Maryland

December 3, 2025

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

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.

Prioritarian Performance Principle (PPP)

- PPP's targets are predictive models: mathematical objects which enable one to infer the probability of some target variable obtaining. It focuses on evaluating the performance of those models, and tells us to care more about model performance for systematically disadvantaged groups—but says nothing about which performance measures are relevant (there are many).

Prioritarian Performance Principle (PPP)

- PPP's targets are predictive models: mathematical objects which enable one to infer the probability of some target variable obtaining. It focuses on evaluating the performance of those models, and tells us to care more about model performance for systematically disadvantaged groups—but says nothing about which performance measures are relevant (there are many).

- PPP gives a necessary condition for predictive justice—not necessary and sufficient conditions. There are almost certainly other necessary conditions of predictive justice, but we will not try to defend them here.

Prioritarian Performance Principle (PPP)

- PPP's targets are predictive models: mathematical objects which enable one to infer the probability of some target variable obtaining. It focuses on evaluating the performance of those models, and tells us to care more about model performance for systematically disadvantaged groups—but says nothing about which performance measures are relevant (there are many).

- PPP gives a necessary condition for predictive justice—not necessary and sufficient conditions. There are almost certainly other necessary conditions of predictive justice, but we will not try to defend them here.

- PPP is prioritarian, because it says to prioritise the worst-off groups (though only up to a point)

The core argument for PPP rests on an empirical premise: **predictive and other representational models deployed within structurally unjust societies are both caused by that background injustice and help sustain it**.

The core argument for PPP rests on an empirical premise: **predictive and other representational models deployed within structurally unjust societies are both caused by that background injustice and help sustain it**.

The standard ways of representing the world that predominate in societies with systematic background injustice routinely centre the experiences of advantaged populations as 'normal'.

We accordingly optimise our epistemic practices to perform best for that 'normal' setting.

And, as a result, they perform worse for members of disadvantaged groups.

Digression: Statistical Discrimination

John W. Patty and Elizabeth Maggie Penn (2022). *Algorithmic Fairness and Statistical Discrimination*. Philosophy Compass.

Elizabeth Maggie Penn and John W. Patty (2025). *Classification algorithms and social outcomes*. American Journal of Political Science, pp. 1-18.

Algorithmic Fairness: Algorithmic fairness (AF) is a new term describing the study of how to evaluate rule-based procedures for making decisions about diverse individuals. At the heart of this study is the presumption that certain ways of discriminating between two or more individuals are undesirable (i.e., "unfair"), whereas others are less suspect, or even desirable (i.e., "permissible").

Algorithmic Fairness: Algorithmic fairness (AF) is a new term describing the study of how to evaluate rule-based procedures for making decisions about diverse individuals. At the heart of this study is the presumption that certain ways of discriminating between two or more individuals are undesirable (i.e., "unfair"), whereas others are less suspect, or even desirable (i.e., "permissible").

Statistical Discrimination: The literature on statistical discrimination (SD) is more established than that on AF. Rather than measuring and classifying disparities in algorithmic performance across groups, this literature squarely aims to identify the root causes of discrimination, and to disentangle disparate outcomes due to discrimination (i.e., disparate treatment) from those due to exogenous disparities across groups.

Example: Hiring

- Suppose that applicants for a job are from two different groups, “male” and “female.”

Example: Hiring

- Suppose that applicants for a job are from two different groups, "male" and "female."
- Every applicant is either qualified or not, but this is not directly observable. Rather, each applicant has taken a test, and the result of this test for applicant is positively correlated with whether he or she is qualified. To make things concrete, suppose that the test is scored on a $0 - 100$ point scale.

Example: Hiring

- Suppose that applicants for a job are from two different groups, "male" and "female."
- Every applicant is either qualified or not, but this is not directly observable. Rather, each applicant has taken a test, and the result of this test for applicant is positively correlated with whether he or she is qualified. To make things concrete, suppose that the test is scored on a $0 - 100$ point scale.
- The employer can observe both the applicant's test score and his or her group membership, and suppose that the employer hires any applicant from group $g \in \{male, female\}$ if and only if his or her test score is greater than or equal to the employer's threshold for group g, denoted by $t(g) \in \{0, \ldots, 100, 101\}$

Example: Hiring

Both AF and SD are interested in the pair of thresholds used by the employer, $t(male)$ and $t(female)$.

This stylized setting allows us to clearly identify discrimination between the two groups: whenever $t(male) \neq t(female)$

Example: Hiring

Studies of Algorithmic Fairness tend to focus on questions like:

1. How do the thresholds affect the applicants' welfares?
2. What does it mean to treat applicants from both groups of applicants fairly?
3. Which pair(s) of thresholds (if any) treat both groups of applicants fairly?

Example: Hiring

Studies of Algorithmic Fairness tend to focus on questions like:

1. How do the thresholds affect the applicants' welfares?
2. What does it mean to treat applicants from both groups of applicants fairly?
3. Which pair(s) of thresholds (if any) treat both groups of applicants fairly?

Studies of Statistical Discrimination tend to focus on questions like:

1. How do the thresholds affect the employer's welfare?
2. Which pair(s) of thresholds maximize the employer's welfare?
3. What factors might justify the employer using different thresholds for the two groups?
4. How do these thresholds affect individual and group behavior?

Traffic Cameras, Fairness, and Discrimination

≡ ProPublica Donate

Chicago's "Race-Neutral" Traffic Cameras Ticket Black and Latino Drivers the Most

A ProPublica analysis found that traffic cameras in Chicago disproportionately ticket Black and Latino motorists. But city officials plan to stick with them — and other cities may adopt them too.

by Emily Hopkins and Melissa Sanchez

Jan. 11, 2022, 5 a.m. EST

Traffic Cameras, Fairness, and Discrimination

The study found that, in Chicago in 2020, "the ticketing rate for households in majority-Black ZIP codes jumped to more than three times that of households in majority-white areas. For households in majority-Hispanic ZIP codes, there was an increase, but it was much smaller."

Traffic Cameras, Fairness, and Discrimination

An Algorithmic Fairness perspective on this situation essentially asks why this disparity emerges and, more provocatively, how one might reduce or eliminate it.

Traffic Cameras, Fairness, and Discrimination

An Algorithmic Fairness perspective on this situation essentially asks why this disparity emerges and, more provocatively, how one might reduce or eliminate it.

This perspective is particularly helpful in this type of setting because, while this disparity has widened over the two decades since the cameras were introduced in Chicago, there is little reason to suspect that traffic cameras themselves are distinguishing between drivers based on their race or home neighborhood, per se.

Traffic Cameras, Fairness, and Discrimination

An Algorithmic Fairness perspective on this situation essentially asks why this disparity emerges and, more provocatively, how one might reduce or eliminate it.

This perspective is particularly helpful in this type of setting because, while this disparity has widened over the two decades since the cameras were introduced in Chicago, there is little reason to suspect that traffic cameras themselves are distinguishing between drivers based on their race or home neighborhood, per se.

In this specific case, this perspective allows one to see that the disparity is at least arguably due to speed limits and driving conditions being distributed in a "non-race blind" fashion across Chicago.

Traffic Cameras, Fairness, and Discrimination

Chicago Mayor Lori Lightfoot's administration described traffic cameras as "a tool in the toolkit to help alleviate" traffic fatalities and, from an empirical standpoint, Black Chicagoans were twice as likely to die in a traffic accident as white Chicagoans in 2017. Accordingly, Black Chicagoans are differentially treated by both traffic accidents and traffic tickets.

Traffic Cameras, Fairness, and Discrimination

From a Statistical Discrimination standpoint, on the other hand, one might ask why Chicago is using traffic cameras, in spite of the clear racial disparity in which citizens receive tickets.

Traffic Cameras, Fairness, and Discrimination

From a Statistical Discrimination standpoint, on the other hand, one might ask why Chicago is using traffic cameras, in spite of the clear racial disparity in which citizens receive tickets.

As the ProPublica article describes, Chicago Mayor Lori Lightfoot proposed lowering the minimum speed at which a speeding ticket would be issued.

Traffic Cameras, Fairness, and Discrimination

From a Statistical Discrimination standpoint, on the other hand, one might ask why Chicago is using traffic cameras, in spite of the clear racial disparity in which citizens receive tickets.

As the ProPublica article describes, Chicago Mayor Lori Lightfoot proposed lowering the minimum speed at which a speeding ticket would be issued.

This proposal, which was adopted by the Chicago City Council in 2021, prompted some to question how much Mayor Lightfoot cared about racial disparities, as opposed to the City of Chicago's serious structural deficit.

Traffic Cameras, Fairness, and Discrimination

From a Statistical Discrimination standpoint, on the other hand, one might ask why Chicago is using traffic cameras, in spite of the clear racial disparity in which citizens receive tickets.

As the ProPublica article describes, Chicago Mayor Lori Lightfoot proposed lowering the minimum speed at which a speeding ticket would be issued.

This proposal, which was adopted by the Chicago City Council in 2021, prompted some to question how much Mayor Lightfoot cared about racial disparities, as opposed to the City of Chicago's serious structural deficit.

The question of "is Mayor Lightfoot more interested in racial equality or city revenue?" is directly analogous to the seminal question in statistical discrimination, "is that employer simply maximizing profits or are they racist?"

Algorithmic Fairness vs. Statistical Discrimination

In terms of the traffic camera example, we can also distinguish the AF and SD viewpoints as

- Algorithmic Fairness: Can we make Chicago's traffic enforcement more fair? If so, how?
- Statistical Discrimination: Why did Chicago use an unfair traffic enforcement algorithm?

Rationality vs. Fairness

A key contrast between the AF & SD approaches revolves around the question of rationality or, in slightly different terms, *efficiency*.

Many SD theories are focused on how the pursuit of efficiency (e.g., by an employer, job applicant, government, or other individuals) can generate behavior that is discriminatory.

On the other hand, AF is less concerned with efficiency (partly because the basic framework does not presume anything about individuals' motives/goals).

Game-Theoretic Models of Discrimination

- When the workers' sensitive trait is observed by the employer at the time of making the hiring decision, individuals with different sensitive traits may be treated by the employer differently in the sense that the hiring rule for one group is different from the hiring rule applied to a different group.

Game-Theoretic Models of Discrimination

- When the workers' sensitive trait is observed by the employer at the time of making the hiring decision, individuals with different sensitive traits may be treated by the employer differently in the sense that the hiring rule for one group is different from the hiring rule applied to a different group.

- This, in turn, leads to each worker's incentive to invest in obtaining qualification endogenously depending on the worker's sensitive trait.

Game-Theoretic Models of Discrimination

- When the workers' sensitive trait is observed by the employer at the time of making the hiring decision, individuals with different sensitive traits may be treated by the employer differently in the sense that the hiring rule for one group is different from the hiring rule applied to a different group.

- This, in turn, leads to each worker's incentive to invest in obtaining qualification endogenously depending on the worker's sensitive trait.

- Accordingly, discriminatory behavior by the employer may emerge as a result of the equilibrium played by the employer and worker depending on the worker's sensitive trait (in game theoretic terms, this is referred to as equilibrium selection).

For example, it can be the case that the employer believes that women invest in qualification with some positive probability, but that men do not. In this case, the employer may (correctly) be willing to hire women whose test scores are high enough but (correctly) never hire a male applicant regardless of his or her test score.

This type of discriminatory equilibrium can emerge even if men and women are otherwise identical.

Cosmic vs. Situated Principles

- Cosmic Principles
  - Hold true at all possible worlds
  - Abstract, universal

Cosmic vs. Situated Principles

- Cosmic Principles
  - Hold true at all possible worlds
  - Abstract, universal

- Situated Principles
  - Make essential reference to actual circumstances
  - Context-dependent

PPP as a Situated Principle

The Predictive Parity Principle (PPP) is explicitly **situated**:

PPP as a Situated Principle

The Predictive Parity Principle (PPP) is explicitly **situated**:

- References background structural injustice (extrinsic to the model)

PPP as a Situated Principle

The Predictive Parity Principle (PPP) is explicitly **situated**:

- References background structural injustice (extrinsic to the model)
- Degree of injustice depends on:
  - Stakes of the prediction
  - Causal history of the model
  - Grounds for differential performance

PPP as a Situated Principle

The Predictive Parity Principle (PPP) is explicitly **situated**:

- References background structural injustice (extrinsic to the model)
- Degree of injustice depends on:
  - Stakes of the prediction
  - Causal history of the model
  - Grounds for differential performance
- A model can be unjust in one country but just in another:

  Example: US recidivism instruments might be permissible in Norway

The Cosmic Approach: Hedden's Case

Recall the setup:

- People randomly assigned to Room A or Room B
- Each person given a biased coin with objective probability noted
- Algorithm: Predict "Heads" if probability > 0.5, else "Tails"

The Cosmic Approach: Hedden's Case

Recall the setup:

- People randomly assigned to Room A or Room B
- Each person given a biased coin with objective probability noted
- Algorithm: Predict "Heads" if probability > 0.5, else "Tails"

The algorithm performs worse for Room B by standard fairness metrics.

The Cosmic Approach: Hedden's Case

Recall the setup:

- People randomly assigned to Room A or Room B
- Each person given a biased coin with objective probability noted
- Algorithm: Predict "Heads" if probability > 0.5, else "Tails"

The algorithm performs worse for Room B by standard fairness metrics.

Yet it seems perfectly fair: it simply reports objective probabilities.

The Cosmic Approach: Hedden's Case

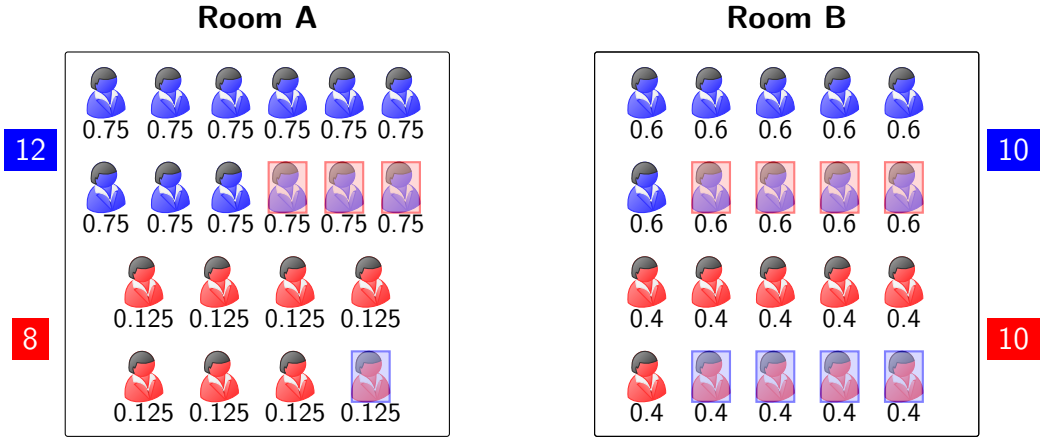
Recall the setup:

- People randomly assigned to Room A or Room B
- Each person given a biased coin with objective probability noted
- Algorithm: Predict "Heads" if probability > 0.5, else "Tails"

The algorithm performs worse for Room B by standard fairness metrics.

Yet it seems perfectly fair: it simply reports objective probabilities.

Key insight: Differential performance can arise from innocent statistical artifacts, not injustice.


Room A
Room B
12
8
10
10
0.75 0.75 0.75 0.75 0.75 0.75
0.75 0.75 0.75 0.75 0.75 0.75
0.125 0.125 0.125 0.125
0.125 0.125 0.125 0.125
0.6 0.6 0.6 0.6 0.6
0.6 0.6 0.6 0.6 0.6
0.4 0.4 0.4 0.4 0.4
0.4 0.4 0.4 0.4 0.4