PHPE 308M/PHIL 209F
Fairness

Eric Pacuit, University of Maryland

December 8, 2025

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

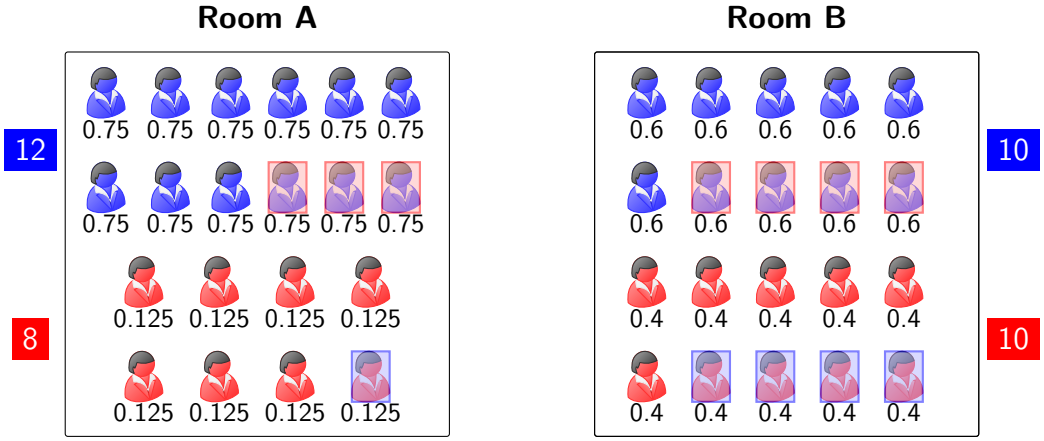
Recall the setup:

- 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

Infra-Marginality

Infra-marginality refers to a statistical phenomenon where observations are distributed away from ("infra" = below) a decision threshold or margin.

Infra-Marginality

Infra-marginality refers to a statistical phenomenon where observations are distributed away from ("infra" = below) a decision threshold or margin.

In the Hedden example:

- Room A: Coin probabilities are spread out far from the 0.5 threshold
- Room B: Coin probabilities cluster near 0.5

Infra-Marginality

Infra-marginality refers to a statistical phenomenon where observations are distributed away from ("infra" = below) a decision threshold or margin.

In the Hedden example:

- Room A: Coin probabilities are spread out far from the 0.5 threshold
- Room B: Coin probabilities cluster near 0.5

When you apply the same decision rule (predict Heads if $p > 0.5$), Room A predictions are more reliable because they're far from the cutoff (e.g., a coin with $p = 0.9$ almost always lands heads).

Infra-Marginality

Infra-marginality refers to a statistical phenomenon where observations are distributed away from ("infra" = below) a decision threshold or margin.

In the Hedden example:

- Room A: Coin probabilities are spread out far from the 0.5 threshold
- Room B: Coin probabilities cluster near 0.5

When you apply the same decision rule (predict Heads if $p > 0.5$), Room A predictions are more reliable because they're far from the cutoff (e.g., a coin with $p = 0.9$ almost always lands heads).

But Room B predictions are shakier (e.g., a coin with p = 0.51 is basically a toss-up, so "predicting Heads" is barely better than guessing.)

Infra-Marginality

The upshot for fairness: Even a perfectly calibrated, non-discriminatory algorithm can have different error rates across groups simply because their underlying distributions sit differently relative to the decision threshold.

This is an "innocent" statistical artifact, not evidence of bias in the algorithm itself.

Why Cosmic Cases Abstract Away

Hedden's case removes all morally relevant context:

- No socially salient groups (random assignment)
- No systematic misrepresentation practices
- No institutions or endorsement
- No stakes (nothing done with predictions)
- No repetition (one-shot game)

Result: Differential performance raises no moral concern.

Making It Real: Step 1

Add repetition and stakes:

- Game repeated multiple times
- Players forced to play, can't switch rooms
- Predictions used to allocate benefits/burdens

Making It Real: Step 1

Add repetition and stakes:

- Game repeated multiple times
- Players forced to play, can't switch rooms
- Predictions used to allocate benefits/burdens

$\Rightarrow$ Being stuck in the worse-performing room **becomes grounds for concern**.

Making It Real: Step 2

Add structural social position:

- Situate in a society with advantaged/disadvantaged groups
- Assignment to inframarginal room tracks social position
- We *know* the artifact exists but use the model anyway

Making It Real: Step 2

Add structural social position:

- Situate in a society with advantaged/disadvantaged groups
- Assignment to inframarginal room tracks social position
- We *know* the artifact exists but use the model anyway

$\Rightarrow$ **Presumptive grounds for complaint** according to predictive justice.

Making It Real: Step 3

Add realistic ML and institutions:

- Probabilities generated by ML model trained on unjust past
- Institution with power over subjects endorses the model
- Institution owes obligations of equal concern and respect

Making It Real: Step 3

Add realistic ML and institutions:

- Probabilities generated by ML model trained on unjust past
- Institution with power over subjects endorses the model
- Institution owes obligations of equal concern and respect

$\Rightarrow$ **Clear case of predictive injustice**.

Summary

Cosmic and situated approaches serve different purposes:

- Cosmic
  - What's fair in any possible world?
  - Useful for foundations
  - Hedden shows: differential performance $\neq$ automatic injustice

- Situated
  - What's fair *here*, given our history?
  - Useful for real-world guidance
  - PPP shows: context determines when it *is* injustice

Summary

Cosmic and situated approaches serve different purposes:

- Cosmic
  - What's fair in any possible world?
  - Useful for foundations
  - Hedden shows: differential performance $\neq$ automatic injustice
- Situated
  - What's fair *here*, given our history?
  - Useful for real-world guidance
  - PPP shows: context determines when it *is* injustice

Both matter, but situated norms capture what cosmic norms miss.

"In the presence of systematic background injustice, when our predictive models are tainted by that very injustice, and when institutions with power endorse them... we can acquire reasons to care about differential model performance which are dependent on those situated, contextual facts."

Human-in-the-Loop Decision Making and Explanations

The Promise of Human-in-the-Loop

The standard justification for algorithmic decision aids:

1. Algorithms make more accurate predictions than humans alone
2. But algorithms can make mistakes, especially in unusual cases
3. So: humans should use algorithms as **advice** while retaining final judgment
4. Human oversight catches algorithmic errors

The Promise of Human-in-the-Loop

The standard justification for algorithmic decision aids:

1. Algorithms make more accurate predictions than humans alone
2. But algorithms can make mistakes, especially in unusual cases
3. So: humans should use algorithms as **advice** while retaining final judgment
4. Human oversight catches algorithmic errors

Question: Does this actually work?

Algorithm-in-the-Loop

> *"[M]any important decisions are now made through an 'algorithm-in-the-loop' process where machine learning models inform people."*

Two questions:

1. What criteria characterize an ethical and responsible decision when a person is informed by an algorithm?
2. Do the ways that people make decisions when informed by an algorithm satisfy these criteria?

B. Green and Y. Chen (2019). *The Principles and Limits of Algorithm-in-the-Loop Decision Making*. Proceedings of ACM Human-Computer Interaction, Vol. 3, No. CSCW, Article 50.

Three Desiderata

Desideratum 1: Accuracy
People using the algorithm should make more accurate predictions than they could without it.

Desideratum 2: Reliability
People should accurately evaluate their own and the algorithm's performance, and calibrate their reliance accordingly.

Desideratum 3: Fairness
People should interact with the algorithm in ways that are unbiased with regard to race, gender, and other sensitive attributes.

Why These Three?

Accuracy: the stated goal of introducing algorithms

Reliability: necessary for:

- Correcting algorithmic errors
- Accountability (can't be accountable if you can't evaluate)
- Handling marginal/unusual cases

Fairness: algorithms may perform differently across groups; humans must not compound this

The Experiment

Participants: ∼1,900 Amazon Mechanical Turk workers

Two prediction tasks:

- Pretrial: Will defendant be rearrested or fail to appear?
- Loans: Will applicant default?

Setup: Participants see profiles and make predictions (0–100%)

Incentive: Paid based on accuracy (Brier score)

Six Conditions

Baseline	No algorithm prediction shown

Six Conditions

Baseline	No algorithm prediction shown
RA Prediction	See algorithm's prediction
Default	Algorithm's prediction pre-filled (can change)

Six Conditions

Baseline	No algorithm prediction shown
RA Prediction	See algorithm's prediction
Default	Algorithm's prediction pre-filled (can change)
Update	Make prediction first, then see algorithm, then revise
Explanation	See prediction plus which features drove it
Feedback	See prediction + told actual outcome after each case

Prediction status: Case 1 of 40

Defendant profile

Defendant #1 is a 29 year old black male. He was arrested for a drug crime. The defendant has previously been arrested 10 times. The defendant has previously been released before trial, and has never failed to appear. He has previously been convicted 10 times.

Risk assessment

The risk score algorithm predicts that this person is 40% likely to fail to appear in court for trial or get arrested before trial. **The prediction has been set to this value, but you are free to predict another value.**

Make a Prediction

How likely is this defendant to fail to appear in court for trial or get arrested before trial?

○ 0% ○ 10% ○ 20% ○ 30% ◉ 40% ○ 50% ○ 60% ○ 70% ○ 80% ○ 90% ○ 100%

Continue

Prediction status: Case 1 of 40

Applicant profile
Loan applicant #1 has applied for a loan of $30,375, with an interest rate of 19.52%. The loan will be paid in 36 monthly installments of $1,121.43. The applicant has an annual income of $80,000 and a "Good" credit score. The applicant has a mortgage out on their home.

Risk assessment
The risk score algorithm predicts that this person is 40% likely to default on their loan. Compared to the average applicant, the following attributes make this applicant notably

- Higher risk: Interest rate.
- Lower risk: Home ownership.

Make a Prediction
How likely is this applicant to default on their loan?

○ 0% ○ 10% ○ 20% ○ 30% ○ 40% ○ 50% ○ 60% ○ 70% ○ 80% ○ 90% ○ 100%

Continue

Brier Score

Evaluated the quality of each prediction using the **Brier score**. When presented with a loan applicant who does not default on their loan, for example, a prediction of 0% risk would yield a score of 1, a prediction of 100% would yield a reward of 0, and a prediction of 50% would yield a score of 0.75.

Participant Prediction Score

The **participant prediction score** is the average Brier score attained among the 40 predictions that each participant made.

Performance gain:

$$Gain_t = \frac{S_t - S_B}{S_R - S_B}$$

where S_t, S_B, and S_R represent the average prediction scores of participants in the treatment t, of participants in Baseline, and of the risk assessment, respectively.

Desideratum 1: Accuracy

Treatment	Pretrial	Loans
Baseline	0%	0%
RA Prediction	46%	68%
Default	53%	57%
Explanation	58%	70%
Update	**60%**	**82%**
Feedback	1%	33%
Algorithm alone	100%	100%

Performance gain relative to baseline, as fraction of algorithm's improvement.

Accuracy: Takeaways

Good news:

- Most treatments improved accuracy over baseline
- Update performed best in both domains

Accuracy: Takeaways

Good news:

- Most treatments improved accuracy over baseline
- Update performed best in both domains

Bad news:

- No treatment matched the algorithm alone
- Feedback made things *worse*
- Participants shifted to extreme predictions (0% or 100%)

Why Did Update Work?

Two mechanisms:

1. **Training effect:** Seeing algorithm predictions helped participants make better *initial* predictions over time
2. **Anchoring on own judgment:** Starting with your own prediction prevents over-reliance on algorithm

The Update participants' *initial* predictions (before seeing algorithm) were better than Baseline participants' predictions.

Desideratum 2: Reliability

Can participants evaluate performance?

Self-Evaluation

Participants were asked: “How confident were you in your decisions?”

Self-Evaluation

Participants were asked: "How confident were you in your decisions?"

Finding: No significant positive relationship between confidence and actual performance in any condition.

In some conditions, the relationship was *negative*: more confident participants performed worse.

Algorithm Evaluation

Participants were asked: "How accurate do you think the risk score algorithm is?"

Algorithm Evaluation

Participants were asked: "How accurate do you think the risk score algorithm is?"

Finding: Participant assessments were not positively correlated with actual algorithm performance.

In some conditions, participants rated the algorithm as *less* accurate when it was actually *more* accurate.

Calibration

Did participants rely more on the algorithm when it was performing well?

Calibration

Did participants rely more on the algorithm when it was performing well?

Finding: In most conditions, no relationship between algorithm quality and reliance.

In some conditions, participants relied *less* on the algorithm when it was more accurate.

Reliability: Summary

Conclusion: No treatment satisfied the reliability desideratum.

Desideratum 3: Fairness

Defendant profiles included race (Black or White).

Race was *not* used by the algorithm.

Question: Did participants respond to the algorithm differently depending on defendant race?

Desideratum 3: Fairness

Defendant profiles included race (Black or White).

Race was *not* used by the algorithm.

Question: Did participants respond to the algorithm differently depending on defendant race? Yes!

All treatments exhibited disparate interactions. But, making your own prediction first may reduce bias in how you incorporate algorithmic advice.