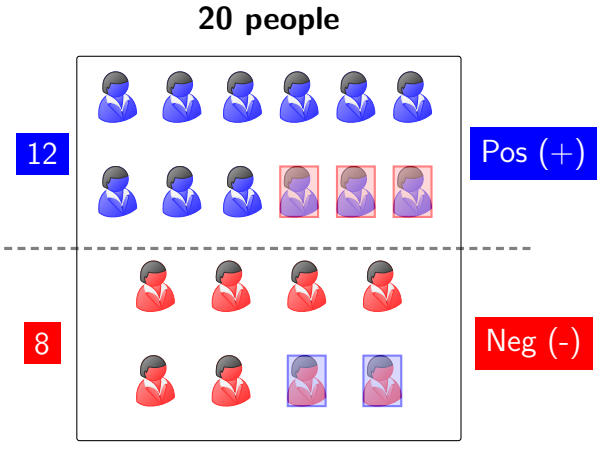
PHPE 308M/PHIL 209F
Fairness

Eric Pacuit, University of Maryland

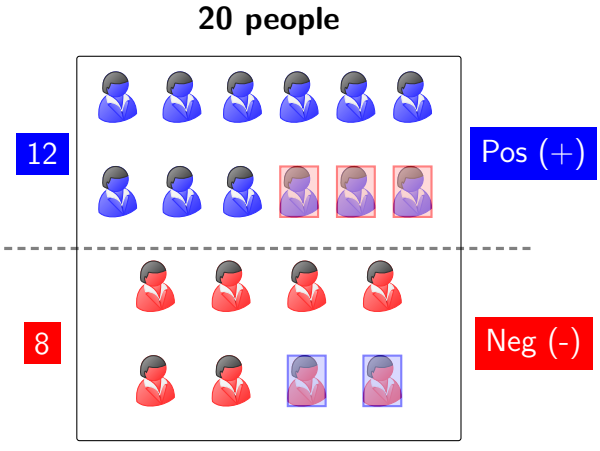
November 17, 2025

Example: Confusion Matrix



	Pred. +	Pred. −
Actual +	?? TP	?? FN
Actual −	?? FP	?? TN

Example: Confusion Matrix



	Pred. +	Pred. −
Actual +	?? TP	?? FN
Actual −	?? FP	?? TN

Accuracy: ??

Base Rate: ??

Example: Confusion Matrix

	Pred. +	Pred. −
Actual +	9 TP	2 FN
Actual −	3 FP	6 TN

Error Rates:

False Neg. Rate: ??

False Pos. Rate: ??

Given the truth, how often is the prediction wrong?

Predictive Value:

Pos. Predictive Value: ??

Neg. Predictive Value: ??

Given the prediction, how often is the prediction correct?

COMPAS Data

Overall population (18,293 defendants)

	Pred: High Risk	Pred: Not High Risk
Actual Recidivist	2921 TP	5489 FN
Actual Non-Recid.	1693 FP	8190 TN

Accuracy: $\frac{2921+8190}{2921+5489+1693+8190} \approx 0.607$

Base Rate: $\frac{2921+5489}{2921+5489+1693+8190} \approx 0.459$

Black Defendants (n = 9,779)

	Pred: High Risk	Pred: Not High Risk
Actual Recidivist	2174 TP	2902 FN
Actual Non-Recid.	1226 FP	3477 TN

Non-Black Defendants (n = 8,514)

	Pred: High Risk	Pred: Not High Risk
Actual Recidivist	747 TP	2587 FN
Actual Non-Recid.	467 FP	4713 TN

Predictive Parity

Black Defendants

	Pred +	Pred −
Actual +	2174	2902
Actual −	1226	3477

Non-Black Defendants

	Pred +	Pred −
Actual +	747	2587
Actual −	467	4713

Predictive Parity

Black Defendants

	Pred +	Pred –
Actual +	2174	2902
Actual –	1226	3477

$PPV = \frac{2174}{2174+1226} \approx 0.639$

Non-Black Defendants

	Pred +	Pred –
Actual +	747	2587
Actual –	467	4713

Predictive Parity

Black Defendants

	Pred +	Pred –
Actual +	2174	2902
Actual –	1226	3477

$PPV = \frac{2174}{2174+1226} \approx 0.639$

Non-Black Defendants

	Pred +	Pred –
Actual +	747	2587
Actual –	467	4713

$PPV = \frac{747}{747+467} \approx 0.615$

Predictive Parity

Black Defendants

	Pred +	Pred −
Actual +	2174	2902
Actual −	1226	3477

$PPV = \frac{2174}{2174+1226} \approx 0.639$

Non-Black Defendants

	Pred +	Pred −
Actual +	747	2587
Actual −	467	4713

$PPV = \frac{747}{747+467} \approx 0.615$

$$0.639 \approx 0.615$$

Predictive parity: conditional on the decision, individuals with different sensitive traits should be equally likely to have the same outcome.

Error Rate Balance

Black Defendants

	Pred +	Pred −
Actual +	2174	2902
Actual −	1226	3477

Non-Black Defendants

	Pred +	Pred −
Actual +	747	2587
Actual −	467	4713

Error Rate Balance

Black Defendants

	Pred +	Pred −
Actual +	2174	2902
Actual −	1226	3477

Non-Black Defendants

	Pred +	Pred −
Actual +	747	2587
Actual −	467	4713

$$FNR = \frac{2902}{2174+2902} \approx 0.572$$

$$FPR = \frac{1226}{1226+3477} \approx 0.261$$

Error Rate Balance

Black Defendants

	Pred +	Pred −
Actual +	2174	2902
Actual −	1226	3477

$$FNR = \frac{2902}{2174+2902} \approx 0.572$$
$$FPR = \frac{1226}{1226+3477} \approx 0.261$$

Non-Black Defendants

	Pred +	Pred −
Actual +	747	2587
Actual −	467	4713

$$FNR = \frac{2587}{747+2587} \approx 0.776$$
$$FPR = \frac{467}{467+4713} \approx 0.090$$

Error Rate Balance

Black Defendants

	Pred +	Pred −
Actual +	2174	2902
Actual −	1226	3477

$$FNR = \frac{2902}{2174+2902} \approx 0.572$$
$$FPR = \frac{1226}{1226+3477} \approx 0.261$$

Non-Black Defendants

	Pred +	Pred −
Actual +	747	2587
Actual −	467	4713

$$FNR = \frac{2587}{747+2587} \approx 0.776$$
$$FPR = \frac{467}{467+4713} \approx 0.090$$

$$0.572 \not\approx 0.776 \text{ and } 0.261 \not\approx 0.090$$

Error rate balance (equalized odds) requires that individuals differing only with respect to sensitive traits are equally likely to be misclassified by the algorithm.

Base Rates

Black Defendants

	Pred +	Pred −
Actual +	2174	2902
Actual −	1226	3477

Non-Black Defendants

	Pred +	Pred −
Actual +	747	2587
Actual −	467	4713

Base Rates

Black Defendants

	Pred +	Pred −
Actual +	2174	2902
Actual −	1226	3477

$\frac{2174+2902}{2174+2902+1226+3477} \approx 0.519$

Non-Black Defendants

	Pred +	Pred −
Actual +	747	2587
Actual −	467	4713

Base Rates

Black Defendants

	Pred +	Pred −
Actual +	2174	2902
Actual −	1226	3477

$\frac{2174+2902}{2174+2902+1226+3477} \approx 0.519$

Non-Black Defendants

	Pred +	Pred −
Actual +	747	2587
Actual −	467	4713

$\frac{747+2587}{747+2587+467+4713} \approx 0.392$

Base Rates

Black Defendants

	Pred +	Pred −
Actual +	2174	2902
Actual −	1226	3477

$\frac{2174+2902}{2174+2902+1226+3477} \approx 0.519$

Non-Black Defendants

	Pred +	Pred −
Actual +	747	2587
Actual −	467	4713

$\frac{747+2587}{747+2587+467+4713} \approx 0.392$

$$0.519 \not\approx 0.392$$

The base rates of recidivism are not equal.

So, there is a conflict between different notions of fairness when analyzing the COMPAS algorithm.

Brian Hedden (2021). *On statistical criteria of algorithmic fairness*. Philosophy & Public Affairs, 49(2), pp. 209 - 231.

Fairness

"I want to focus not on whether an algorithm is unfair to individuals, or whether it is unfair to groups. Rather, I want to focus on whether it is unfair to individuals *in virtue of their membership in a certain group*."

Fairness

How does this notion of fairness differ from the others?

Fairness

How does this notion of fairness differ from the others?

- One can be unfair to an individual without being unfair to them in virtue of their group membership.

Fairness

How does this notion of fairness differ from the others?

- One can be unfair to an individual without being unfair to them in virtue of their group membership.
- It is not obvious that fairness is owed to groups, as opposed to individuals.

Fairness

How does this notion of fairness differ from the others?

- One can be unfair to an individual without being unfair to them in virtue of their group membership.
- It is not obvious that fairness is owed to groups, as opposed to individuals.
- Granting the notion of unfairness to groups, one can perhaps be unfair to an individual in virtue of their membership in a certain group without being unfair to that group itself, for instance if one treats a single individual worse because of their race or gender but at the same time takes other actions that are to the net benefit of that group.

Consider 10 different fairness criteria.
Are any of them **necessary** for an algorithm to be fair?

Fairness (1)

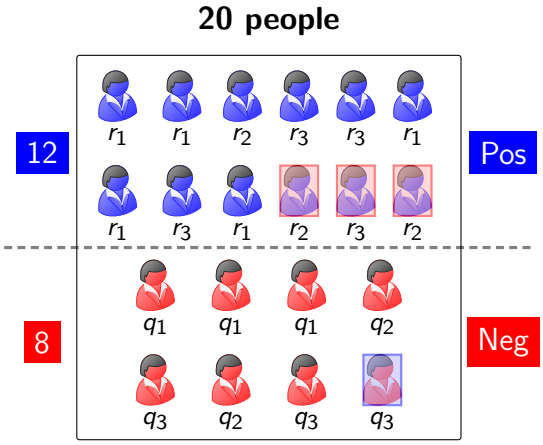
Calibration Within Groups: For each possible risk score, the (expected) percentage of individuals assigned that risk score who are actually positive is the same for each relevant group and is equal to that risk score.

Fairness (1)

Calibration Within Groups: For each possible risk score, the (expected) percentage of individuals assigned that risk score who are actually positive is the same for each relevant group and is equal to that risk score.

The idea is that fairness requires a given risk score to "mean the same thing" for each relevant group. We want the assignment of a given risk score to have the same evidential value, regardless of the group to which the individual belongs.

Calibration



risk score	proportion Pos
r_1	1.0
r_2	1/3
r_3	3/4
q_1	0
q_2	0
q_3	1/3

Fairness (2)

Equal Positive Predicative Value: The (expected) percentage of individuals Predicted to be positive who are actually positive is the same for each relevant group.

Equal Negative Predicative Value: The (expected) percentage of individuals Predicted to be negative who are actually negative is the same for each relevant group.

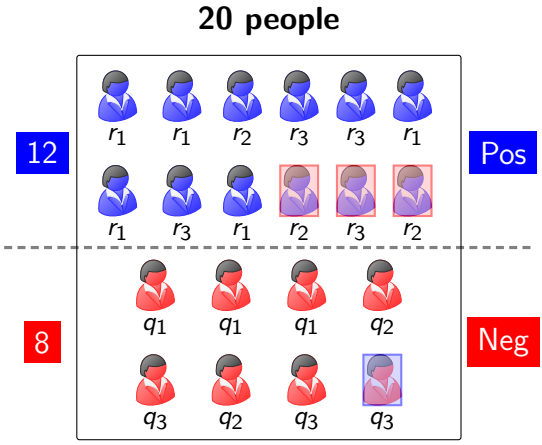
Fairness (2)

Equal Positive Predicative Value: The (expected) percentage of individuals Predicted to be positive who are actually positive is the same for each relevant group.

Equal Negative Predicative Value: The (expected) percentage of individuals Predicted to be negative who are actually negative is the same for each relevant group.

The idea is that fairness requires a prediction of positive to mean the same thing, or to have the same evidential value, regardless of the group to which the individual belongs (similarly for a prediction of negative).

Pos/Neg Predictive Value



Pos Predicative Value: 9/12

Neg Predicative Value: 7/8

Fairness (3)

Equal False-Positive Rates: The (expected) percentage of actually negative individuals who are falsely predicted to be positive is the same for each relevant group.

Equal False-Negative Rates: The (expected) percentage of actually positive individuals who are falsely predicted to be negative is the same for each relevant group.

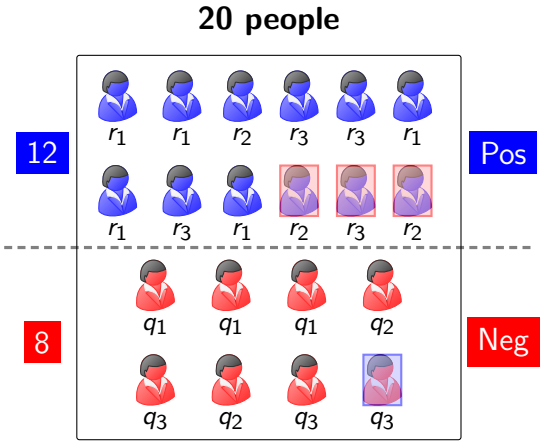
Fairness (3)

Equal False-Positive Rates: The (expected) percentage of actually negative individuals who are falsely predicted to be positive is the same for each relevant group.

Equal False-Negative Rates: The (expected) percentage of actually positive individuals who are falsely predicted to be negative is the same for each relevant group.

The idea is that fairness requires individuals from different groups who exhibit the same behavior to, on balance, be treated the same by the algorithm in terms of whether they are Predicted to be positive or negative. It would be unfair, for instance, if individuals from one group who are actually negative tended to be Predicted to be positive at higher rates than actually negative members of the other group.

False Pos/Neg Rate



False Pos Rate: 3/10

False Neg Rate: 1/10

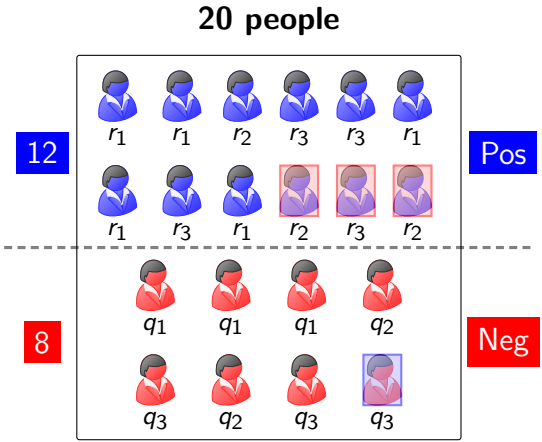
Fairness (4)

Balance for the Positive Class: The (expected) average risk score assigned to those individuals who are actually positive is the same for each relevant group.

Balance for the Negative Class: The (expected) average risk score assigned to those individuals who are actually negative is the same for each relevant group.

These are generalizations of the previous two conditions from the case of binary predictions to the case of risk scores, and are motivated in the same way.

Average Risk Scores



Average Pos Risk Score:

$(5 * r_1 + r_2 + 3 * r_3 + q_3)/10$

False Neg Rate:

$(3 * q_1 + 2 * q_2 + 2 * q_3 + 2 * r_2 + r_3)/10$

Fairness (5)

Equal Ratios of False-Positive Rate to False-Negative Rate: The (expected) ratio of the false-positive rate to the false-negative rate is the same for each relevant group.

Equal Overall Error Rates: The (expectation of) the number of false positives and false negatives, divided by the number of individuals, is the same for each relevant group.

Fairness (5)

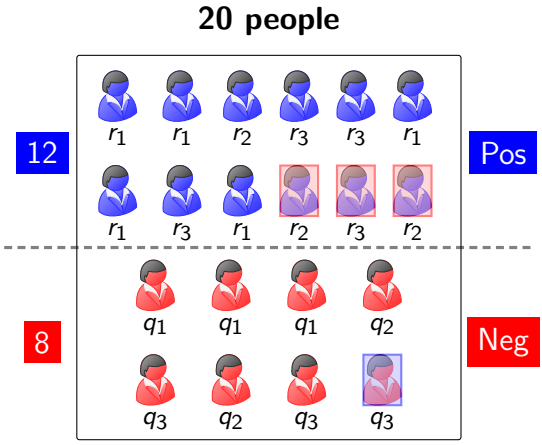
Equal Ratios of False-Positive Rate to False-Negative Rate: The (expected) ratio of the false-positive rate to the false-negative rate is the same for each relevant group.

Equal Overall Error Rates: The (expectation of) the number of false positives and false negatives, divided by the number of individuals, is the same for each relevant group.

The idea is that fairness requires assigning equal relative weights to the two main error types, false positives and false negatives, for the various groups. It would be unfair, for instance, if the algorithm tended to err on the side of caution for one group while tending to do the reverse for the other group.

Equal Overall Error Rates incorporates the thought that it would be unfair if an algorithm were simply less accurate for one group than for another.

Ratio/Error Rate



False Pos to False Neg: 3 : 1

Error Rate: 4/20

Fairness (6)

Statistical Parity: The (expected) percentage of individuals Predicted to be positive is the same for each relevant group.

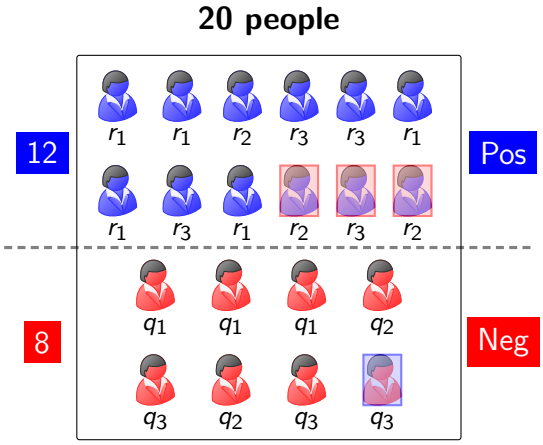
Fairness (6)

Statistical Parity: The (expected) percentage of individuals Predicted to be positive is the same for each relevant group.

The idea is that the percentage of individuals predicted to be positive be the same for each relevant group.

However, this criteria is in fact widely rejected, because it is insensitive to differences in base rates (ratios of actual positives to actual negatives) across groups. Indeed, when base rates differ across groups, this criterion will be violated by an omniscient algorithm which perfectly Predicts people's behavior. But a perfect algorithm would, presumably, not be unfair simply in virtue of differing base rates.

Percentage Predicted Positive



% Predicted Pos: 12/20

Fairness (7)

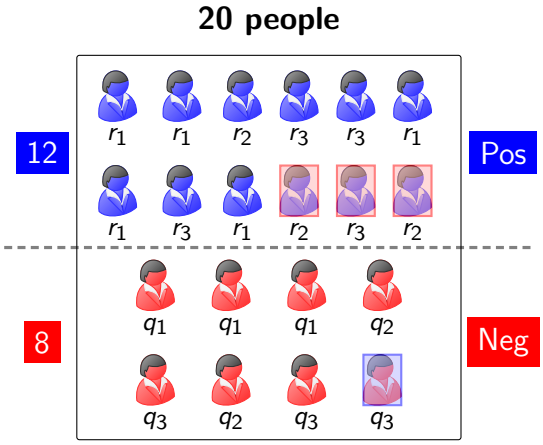
Equal Ratios of Predicted Positives to Actual Positives: The (expectation of) the number of individuals predicted to be positive, divided by the number of individuals who are actually positive, is the same for each relevant group.

Fairness (7)

Equal Ratios of Predicted Positives to Actual Positives: The (expectation of) the number of individuals predicted to be positive, divided by the number of individuals who are actually positive, is the same for each relevant group.

This improves on the previous previous criterion. When base rates differ, this requires that differences in base rates yield corresponding differences in the rates at which individuals are Predicted to be positive.

Ratio Predicated to Actual



Predicted Pos: Actual Pos: 12/10

Impossibility

Theorem (Kleinberg, Mullainathan, and Raghavan 2016) No algorithm (for Predicting risk scores) can satisfy Calibration Within Groups, Balance for the Positive Class and Balance for the Negative Class, unless either

1. the base rates are equal across the relevant groups, or
2. the algorithm makes perfect predictions (assigning risk score 1 to all actual positives and risk score 0 to all actual negatives).

J. Kleinberg, S. Mullainathan, and M. Raghavan (2016). *Inherent trade-offs in the fair determination of risk scores*. https://arxiv.org/abs/1609.05807.

Impossibility

Theorem (Chouldechova 2017) No algorithm (for binary predictions) can satisfy Equal False-Positive Rates, Equal False-Negative Rates, and Equal Positive Predicative Value unless

1. the base rates are equal across the relevant groups, or
2. the algorithm makes perfect predictions (assigning 1 to all actual positives and 0 to all actual negatives).

A. Chouldechova (2017). *Fair prediction with disparate impact: A study of bias in recidivism prediction instruments*. https://arxiv.org/abs/1610.07524.

Impossibility

Theorem (Miconi) No algorithm can satisfy more than one of (i) Equal False-Positive Rates and Equal False-Negative Rates, (ii) Equal Positive Predicative Value and Equal Negative Predicative Value, and (iii) Equal Ratios of Predicted Positives to Actual Positives unless

1. the base rates are equal across the relevant groups, or
2. the algorithm makes perfect predictions (assigning 1 to all actual positives and 0 to all actual negatives).

T. Miconi (2017). *The impossibility of "fairness": a generalized impossibility result for decisions*. https://arxiv.org/abs/1707.01195.

"These results suggest some of the ways in which key notions of fairness are incompatible with each other." (Kleinberg et al. 2016)

Impossibility

We might interpret these results as showing that fairness dilemmas are inevitable: whatever we do, we cannot help being unfair or biased.

Impossibility

We might interpret these results as showing that fairness dilemmas are inevitable: whatever we do, we cannot help being unfair or biased.

Alternatively, we might interpret them as showing that not all of these statistical criteria are *necessary* conditions for an algorithm to be fair or unbiased. Which criteria, then, are genuine conditions of fairness?

A Perfectly Fair Algorithm

Suppose that there are a bunch of coins of varying biases.

Each individual in the population is

1. randomly assigned a coin; and
2. randomly assigned to one of two rooms, A and B.

Goal: For each person, Predict whether that person's coin will land heads or tails. That is, our aim is to Predict, for each person, whether they are a heads person or a tails person.

Luckily, each coin comes labeled with its bias, with a real number in the interval $[0, 1]$ indicating its bias, or its objective chance of landing heads.

A Perfectly Fair Algorithm

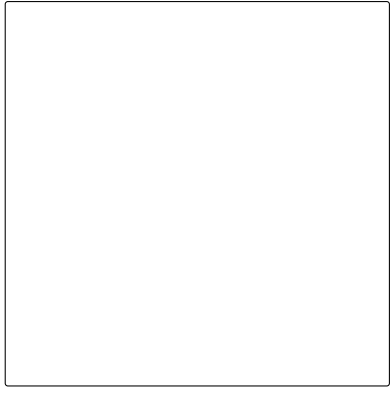
For each person, take their coin and read its label.

- If the coin label says x, assign that person a risk score of x.
- if $x > 0.5$, then Predict that they are a heads person (positive)
- if $x < 0.5$, then Predict that they are a tails person (negative).
- if $x = 0.5$, then randomize prediction (but "sidestep this issue by assuming that none of the coins are labeled "0.5").

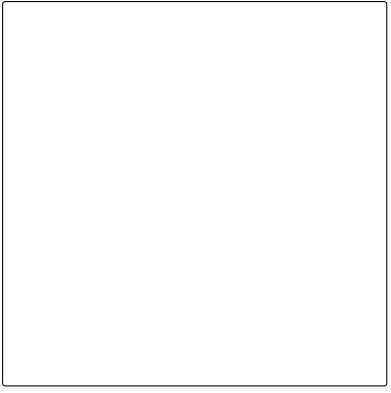
A Perfectly Fair Algorithm

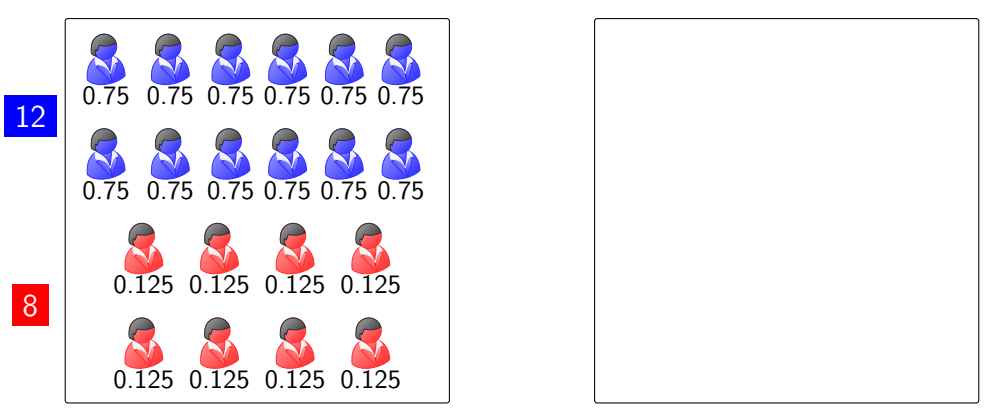
- This algorithm is perfectly fair and unbiased, and in particular, it is not unfair to any people in virtue of their room membership.
- The algorithm predictions are not sensitive to individuals' room membership. And the sole feature on which its predictions are based (the labeled bias of the coin) is clearly the relevant one to focus on and is neither a proxy for, nor caused or explained by, room membership.
- Indeed, it is not just that the algorithm is in no way unfair to individuals in virtue of their membership in a certain room; there is seemingly no unfairness of any kind anywhere in this situation.
- This algorithm is uniquely optimal; no alternative can be expected to do as well or better at Predicting whether individuals are heads people or tails people.

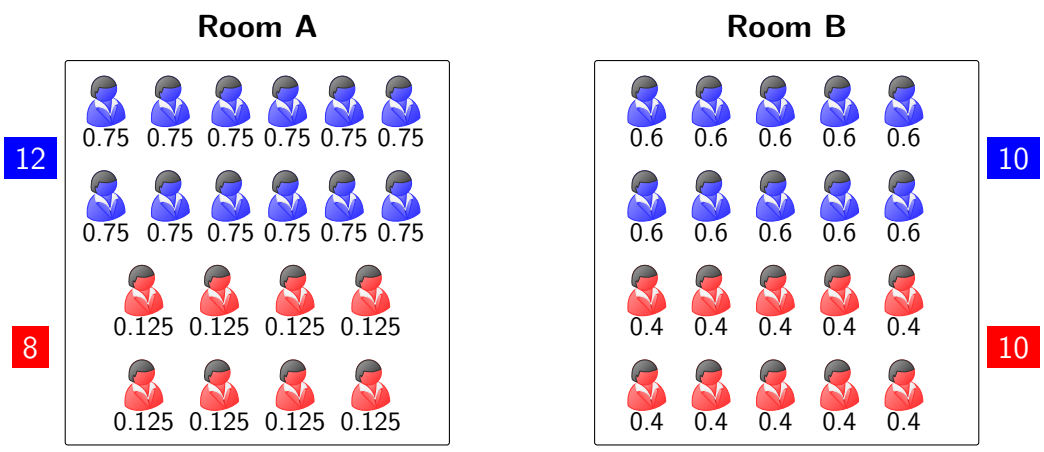
Room A

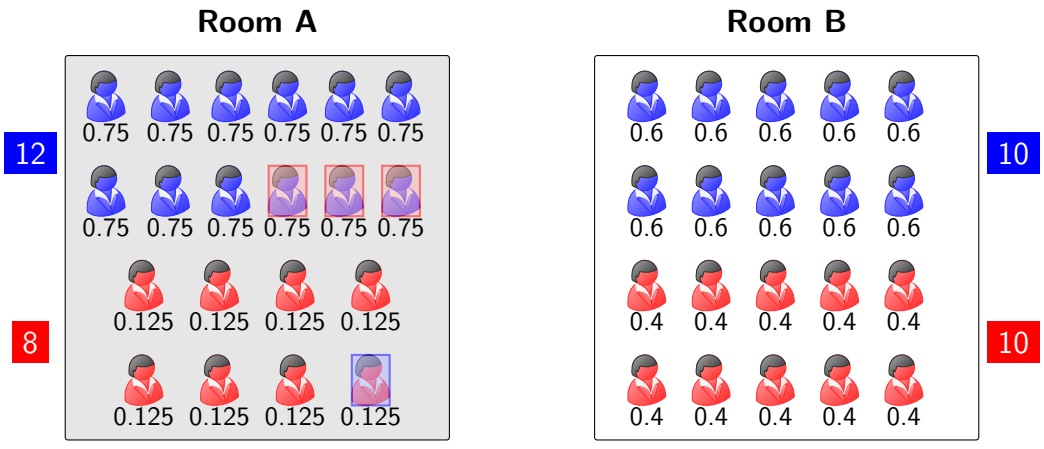


Room B



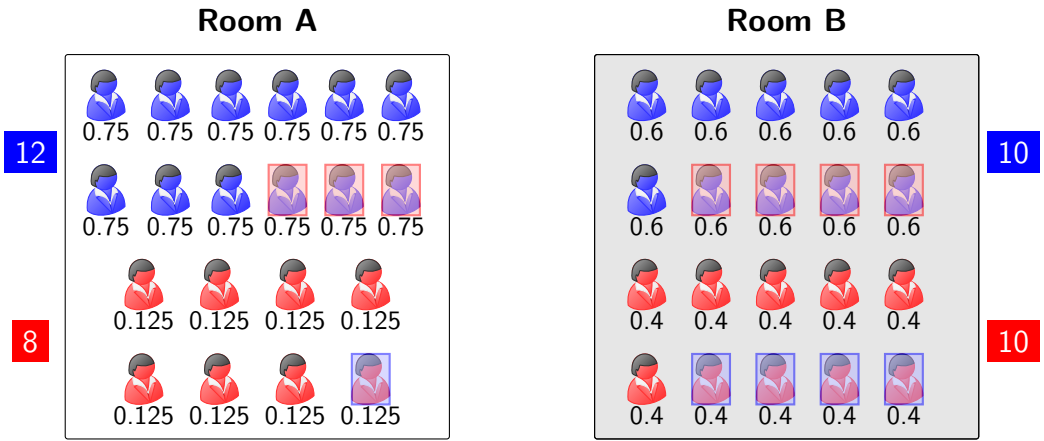

Room A
Room B
12
8
0.75 0.75 0.75 0.75 0.75 0.75
0.75 0.75 0.75 0.75 0.75 0.75
0.125 0.125 0.125 0.125
0.125 0.125 0.125 0.125


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



Room A: $0.75 * 12 + 0.125 * 8 =$ 10 people are actually heads people.

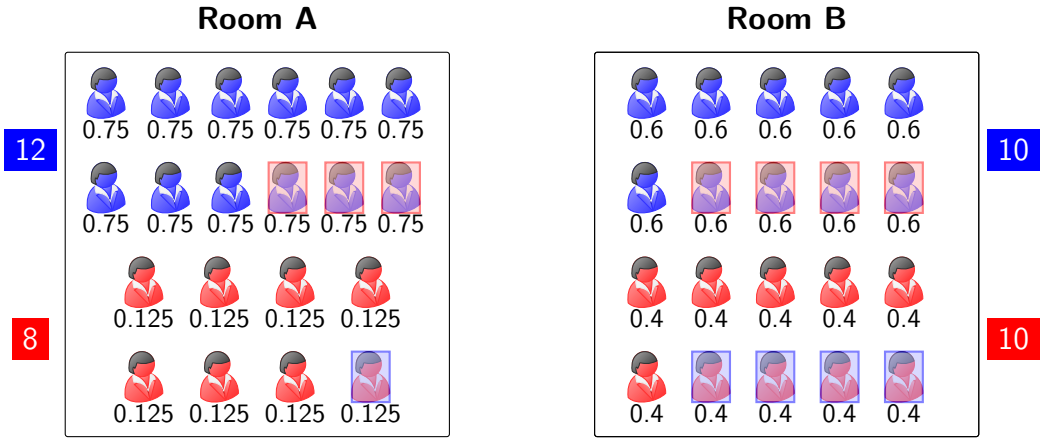
Room A: $0.25 * 12 + 0.875 * 8 =$ 10 people are actually tails people.



Room B: $0.6 * 10 + 0.4 * 10 = 10$ people are actually heads people.

Room B: $0.4 * 10 + 0.6 * 10 = 10$ people are actually tails people.

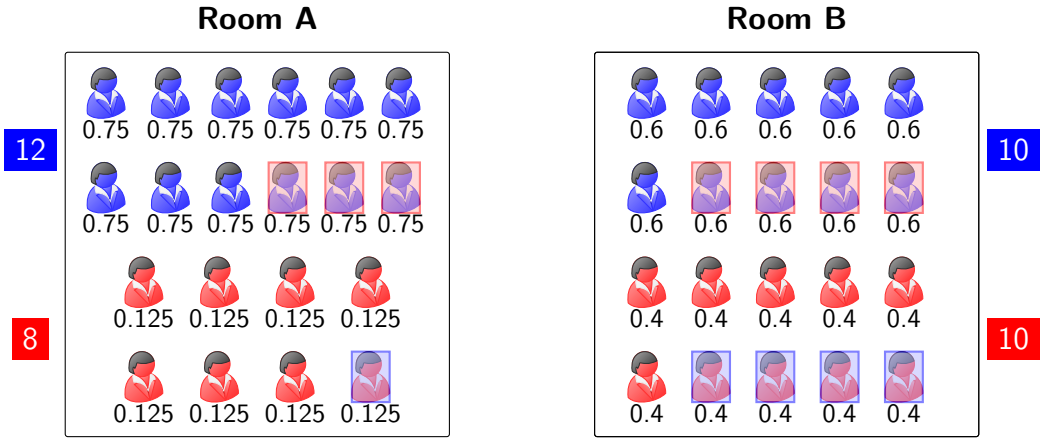
Balance for the Positive Class is Violated



Room A Room B

$$(9 * 0.75 + 1 * 0.125)/10 = 0.6875 \quad \neq \quad 0.52 = (6 * 0.6 + 4 * 0.4)/10$$

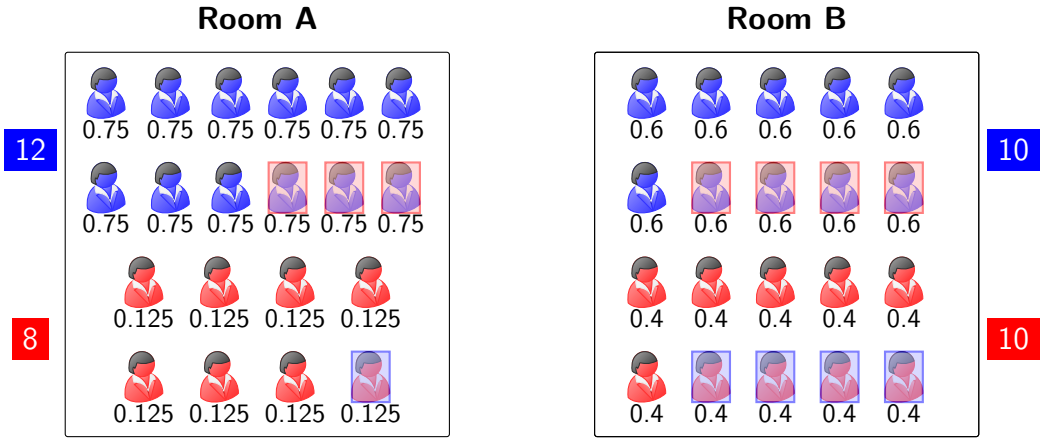
Balance for the Negative Class is Violated



Room A	Room B

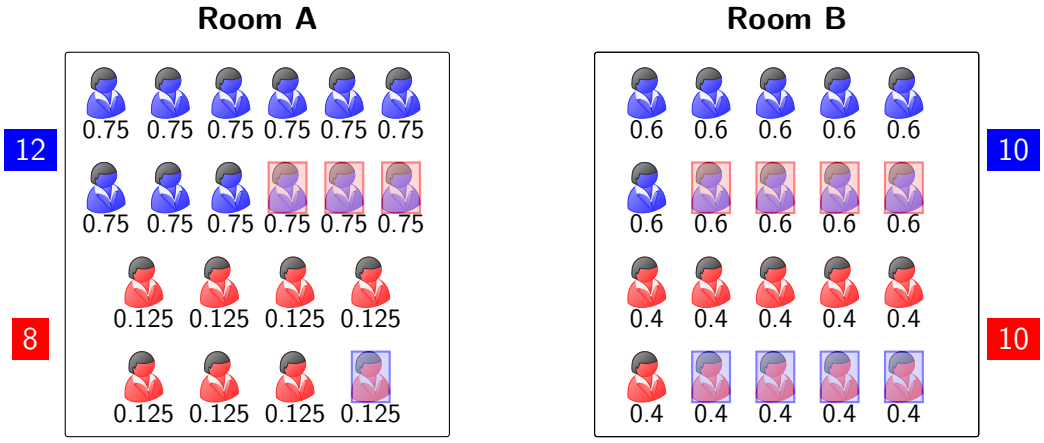
$$(3 * 0.75 + 7 * 0.125)/10 = 0.3125 \quad \neq \quad 0.48 = (4 * 0.6 + 6 * 0.4)/10$$

Equal False-Positive Rates is Violated



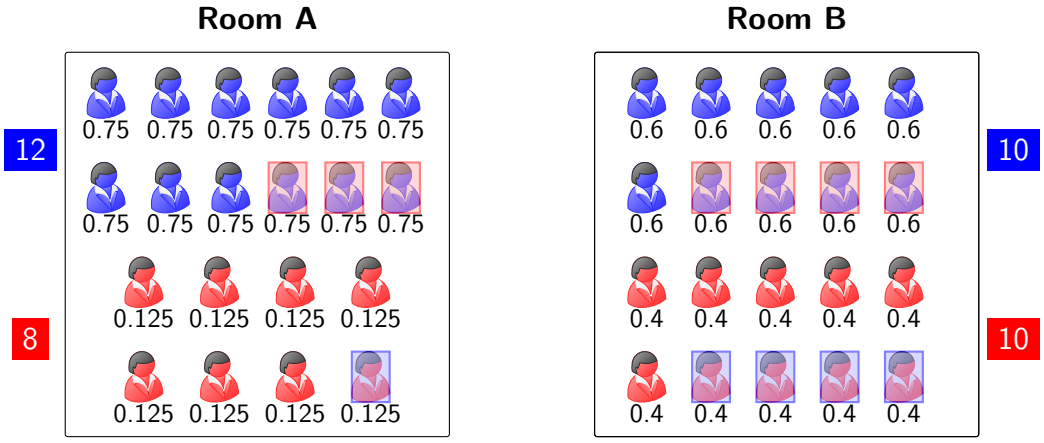
Room A		Room B
(False Pos Rate) 3/10	$\neq$	4/10 (False Pos Rate)

Equal False-Negative Rates is Violated



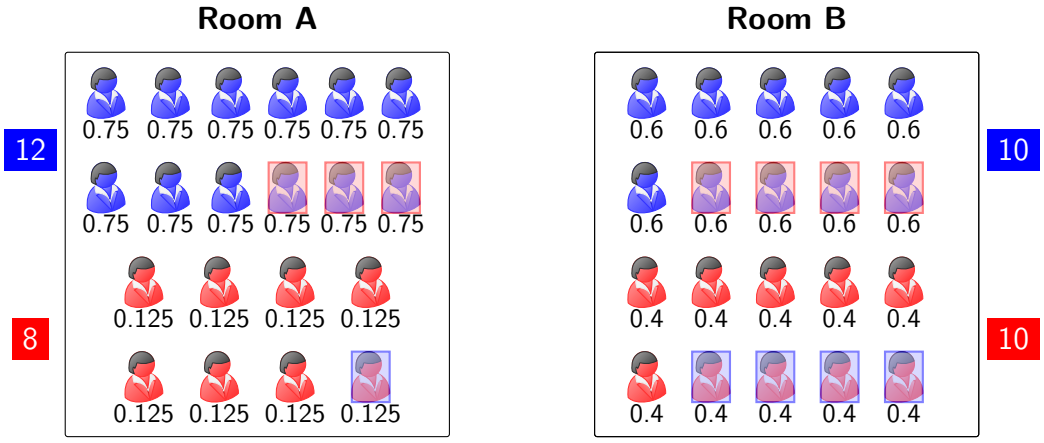
Room A		Room B
(False Neg Rate) 1/10	$\neq$	4/10 (False Neg Rate)

Equal Positive Predicative Value is Violated



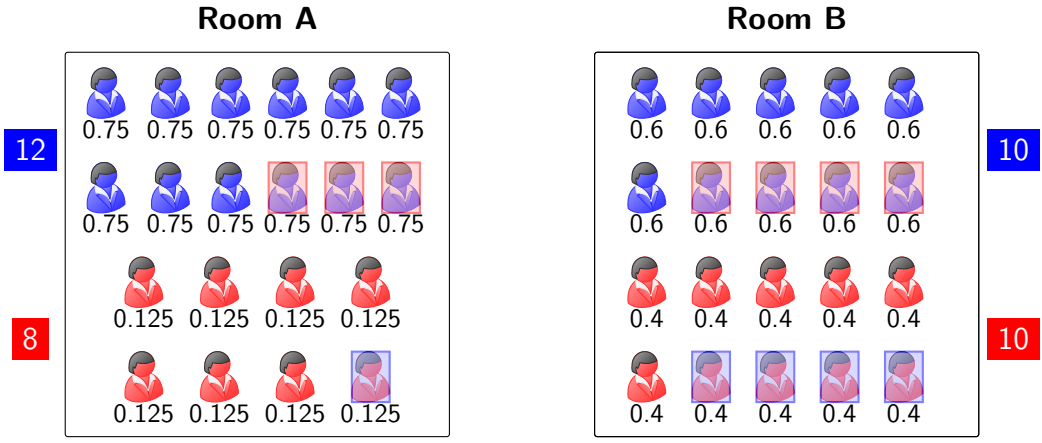
Room A	Room B
(Pos Predicative Value) $9/12$	$\neq$ $6/10$ (Pos Predicative Value)

Equal Negative Predicative Value is Violated



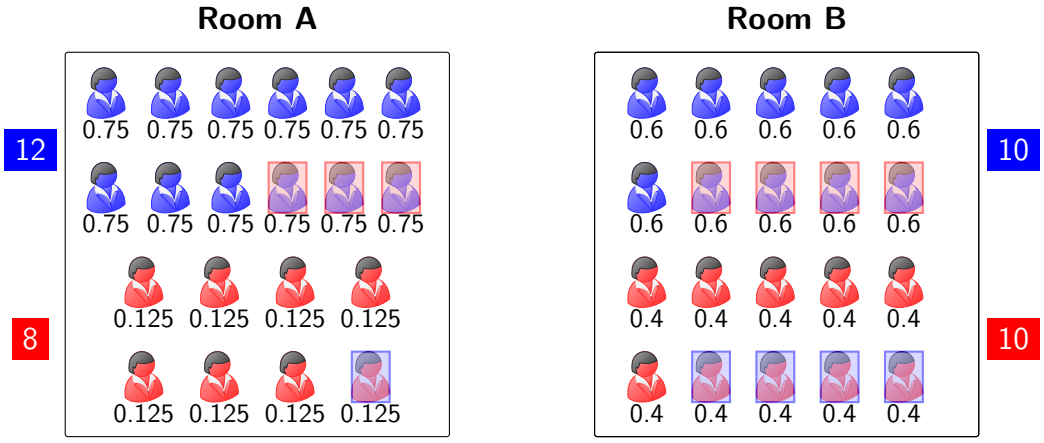
Room A	Room B
(Neg Predicative Value) $7/8$	$\neq$ $6/10$ (Neg Predicative Value)

Equal Ratios of False-Positive Rate to False-Negative is Violated



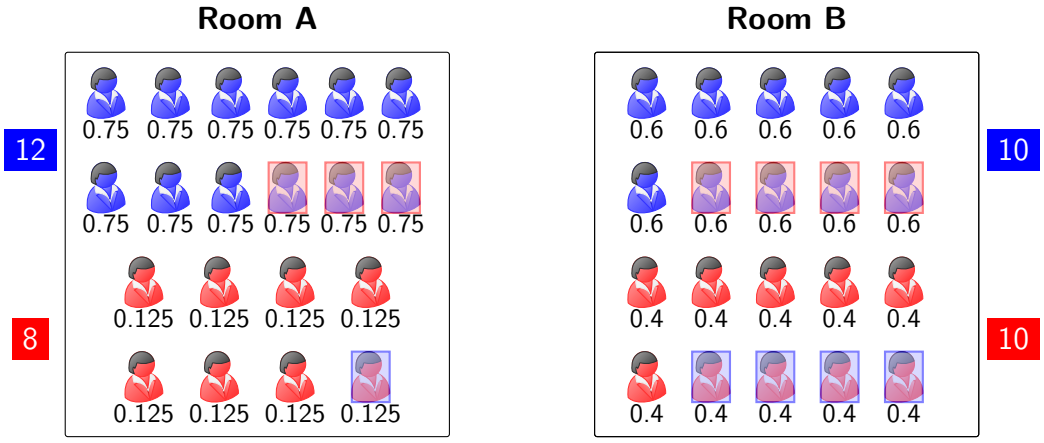
(Ratio False Pos: False Neg) $3:1 \neq 1:1$ (Ratio False Pos: False Neg)

Equal Overall Error Rates is Violated



Room A		Room B
(Overall Error Rate) $4/20$	$\neq$	$8/20$ (Overall Error Rate)

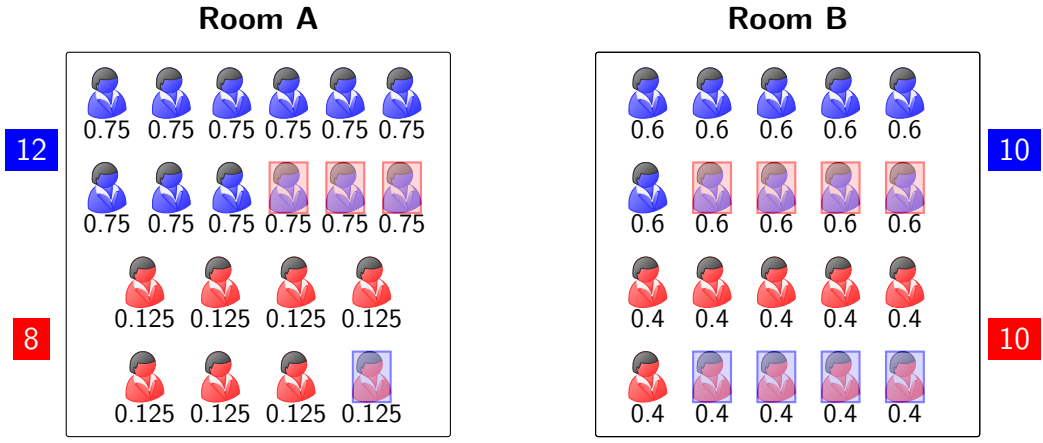
Statistical Parity is Violated



Room A Room B

(% Predicted to be Pos) 12/20 $\neq$ 10/20 (% Predicted to be Pos)

Equal Ratios of Predicted to Actual Positives is Violated



(Ratio of Pred Pos:Actual Pos) 12 : 10 $\neq$ 10 : 10 (Ratio of Pred Pos:Actual Pos)

	Room A	Room B
Avg Score of Positives	0.6875	0.52
Avg Score of Negatives	0.3125	0.48
False Pos Rate	3/10	4/10
False Neg Rate	1/10	4/10
Pos Predicative Value	3/4	3/5
Neg Predicative Value	7/8	3/5
Ratio False Pos: False Neg	3	1
Overall Error Rate	4/20	8/20
% Predicted to be Pos	12/20	10/10
Ratio of Pred Pos:Actual Pos	12/10	10/10

Let me emphasize the limited nature of my argument.

I am not claiming that the case of people, coins, and rooms is realistic or completely analogous to cases like COMPAS. Of course it is not. In my example, room membership is not socially constructed, is not the basis of historical oppression, and does not influence what features people have or how they "behave" (whether their coins land heads).

But my argument does not depend on my example being realistic.

But my argument does not depend on my example being realistic.

1. simplifications and idealizations can help clarify issues by abstracting away from messy complicating factors. In real-life cases, group membership influences what features individuals have, thereby raising the thorny issue of basing predictions on "proxies" for group membership.

But my argument does not depend on my example being realistic.

1. simplifications and idealizations can help clarify issues by abstracting away from messy complicating factors. In real-life cases, group membership influences what features individuals have, thereby raising the thorny issue of basing predictions on "proxies" for group membership.

2. only arguing that none of the above criteria (except Calibration Within Groups) are necessary for fairness. And to conclude that some criterion is not necessary for fairness, all you need is a single case where fairness is satisfied but the criterion violated. That is what I have sought to provide.