PHPE 308M/PHIL 209F
Fairness

Eric Pacuit, University of Maryland

November 24, 2025

Redlining

Imagine a bank that wants to discriminate against black loan applicants, and suppose that black applicants tend to live in zip codes with higher than average default rates, although, within any given zip code, black applicants actually have the same average default rate as other applicants from the same area.

Redlining

Imagine a bank that wants to discriminate against black loan applicants, and suppose that black applicants tend to live in zip codes with higher than average default rates, although, within any given zip code, black applicants actually have the same average default rate as other applicants from the same area.

The bank can achieve its discriminatory agenda by assigning risk scores to applicants based purely on their zip code, and ignoring other relevant factors like income, credit history, and so on.

This is an idealized illustration of a real historical phenomena called "redlining," which lenders used to avoid giving mortgages to minority applicants in the 1930s.

Redlining 1

Race	Zip	Credit	Number	Default rate	Risk score
White	TR10	Good	90	$\frac{1}{10}$	$\frac{1}{4}$
White	TR10	Bad	30	$\frac{1}{5}$	$\frac{1}{4}$
White	TR11	Good	40	$\frac{1}{10}$	$\frac{3}{4}$
White	TR11	Bad	40	$\frac{1}{5}$	$\frac{3}{4}$
Black	TR10	Good	60	$\frac{1}{10}$	$\frac{1}{4}$
Black	TR10	Bad	20	$\frac{1}{5}$	$\frac{1}{4}$
Black	TR11	Good	60	$\frac{1}{10}$	$\frac{3}{4}$
Black	TR11	Bad	60	$\frac{1}{5}$	$\frac{3}{4}$

Redlining 1

Race	Zip	Credit	Number	Default rate	Risk score
White	TR10	Good	90	$\frac{1}{10}$	$\frac{1}{4}$
White	TR10	Bad	30	$\frac{1}{5}$	$\frac{1}{4}$
White	TR11	Good	40	$\frac{1}{10}$	$\frac{3}{4}$
White	TR11	Bad	40	$\frac{1}{5}$	$\frac{3}{4}$
Black	TR10	Good	60	$\frac{1}{10}$	$\frac{1}{4}$
Black	TR10	Bad	20	$\frac{1}{5}$	$\frac{1}{4}$
Black	TR11	Good	60	$\frac{1}{10}$	$\frac{3}{4}$
Black	TR11	Bad	60	$\frac{1}{5}$	$\frac{3}{4}$

All TR10 residents are assigned a risk score of 1/4

Redlining 1

Race	Zip	Credit	Number	Default rate	Risk score
White	TR10	Good	90	$\frac{1}{10}$	$\frac{1}{4}$
White	TR10	Bad	30	$\frac{1}{5}$	$\frac{1}{4}$
White	TR11	Good	40	$\frac{1}{10}$	$\frac{3}{4}$
White	TR11	Bad	40	$\frac{1}{5}$	$\frac{3}{4}$
Black	TR10	Good	60	$\frac{1}{10}$	$\frac{1}{4}$
Black	TR10	Bad	20	$\frac{1}{5}$	$\frac{1}{4}$
Black	TR11	Good	60	$\frac{1}{10}$	$\frac{3}{4}$
Black	TR11	Bad	60	$\frac{1}{5}$	$\frac{3}{4}$

All TR11 residents are assigned a risk score of 3/4

Redlining 1

Race	Zip	Credit	Number	Default rate	Risk score
White	TR10	Good	90	$\frac{1}{10}$	$\frac{1}{4}$
White	TR10	Bad	30	$\frac{1}{5}$	$\frac{1}{4}$
White	TR11	Good	40	$\frac{1}{10}$	$\frac{3}{4}$
White	TR11	Bad	40	$\frac{1}{5}$	$\frac{3}{4}$
Black	TR10	Good	60	$\frac{1}{10}$	$\frac{1}{4}$
Black	TR10	Bad	20	$\frac{1}{5}$	$\frac{1}{4}$
Black	TR11	Good	60	$\frac{1}{10}$	$\frac{3}{4}$
Black	TR11	Bad	60	$\frac{1}{5}$	$\frac{3}{4}$

Good credit $\rightarrow$ default rate of $1/10$

Redlining 1

Race	Zip	Credit	Number	Default rate	Risk score
White	TR10	Good	90	$\frac{1}{10}$	$\frac{1}{4}$
White	TR10	Bad	30	$\frac{1}{5}$	$\frac{1}{4}$
White	TR11	Good	40	$\frac{1}{10}$	$\frac{3}{4}$
White	TR11	Bad	40	$\frac{1}{5}$	$\frac{3}{4}$
Black	TR10	Good	60	$\frac{1}{10}$	$\frac{1}{4}$
Black	TR10	Bad	20	$\frac{1}{5}$	$\frac{1}{4}$
Black	TR11	Good	60	$\frac{1}{10}$	$\frac{3}{4}$
Black	TR11	Bad	60	$\frac{1}{5}$	$\frac{3}{4}$

Bad credit $\rightarrow$ default rate of $1/5$

Does This Satisfy Calibration Within Groups?

Redlining 1

Race	Zip	Credit	Number	Default rate	Risk score
White	TR10	Good	90	$\frac{1}{10}$	$\frac{1}{4}$
White	TR10	Bad	30	$\frac{1}{5}$	$\frac{1}{4}$
White	TR11	Good	40	$\frac{1}{10}$	$\frac{3}{4}$
White	TR11	Bad	40	$\frac{1}{5}$	$\frac{3}{4}$
Black	TR10	Good	60	$\frac{1}{10}$	$\frac{1}{4}$
Black	TR10	Bad	20	$\frac{1}{5}$	$\frac{1}{4}$
Black	TR11	Good	60	$\frac{1}{10}$	$\frac{3}{4}$
Black	TR11	Bad	60	$\frac{1}{5}$	$\frac{3}{4}$

Check risk score 1/4 (all TR10 residents):

- White applicants: 90 good + 30 bad = 120 total
  - Defaults: $90 \times \frac{1}{10} + 30 \times \frac{1}{5} = 15$
  - Actual rate: $\frac{15}{120} = \frac{1}{8}$

Does This Satisfy Calibration Within Groups?

Redlining 1

Race	Zip	Credit	Number	Default rate	Risk score
White	TR10	Good	90	$\frac{1}{10}$	$\frac{1}{4}$
White	TR10	Bad	30	$\frac{1}{5}$	$\frac{1}{4}$
White	TR11	Good	40	$\frac{1}{10}$	$\frac{3}{4}$
White	TR11	Bad	40	$\frac{1}{5}$	$\frac{3}{4}$
Black	TR10	Good	60	$\frac{1}{10}$	$\frac{1}{4}$
Black	TR10	Bad	20	$\frac{1}{5}$	$\frac{1}{4}$
Black	TR11	Good	60	$\frac{1}{10}$	$\frac{3}{4}$
Black	TR11	Bad	60	$\frac{1}{5}$	$\frac{3}{4}$

Check risk score 1/4 (all TR10 residents):

- White applicants: 90 good + 30 bad = 120 total
  - Defaults: $90 \times \frac{1}{10} + 30 \times \frac{1}{5} = 15$
  - Actual rate: $\frac{15}{120} = \frac{1}{8}$
- Black applicants: 60 good + 20 bad = 80 total
  - Defaults: $60 \times \frac{1}{10} + 20 \times \frac{1}{5} = 10$
  - Actual rate: $\frac{10}{80} = \frac{1}{8}$

Does This Satisfy Calibration Within Groups?

Redlining 1

Race	Zip	Credit	Number	Default rate	Risk score
White	TR10	Good	90	$\frac{1}{10}$	$\frac{1}{4}$
White	TR10	Bad	30	$\frac{1}{5}$	$\frac{1}{4}$
White	TR11	Good	40	$\frac{1}{10}$	$\frac{3}{4}$
White	TR11	Bad	40	$\frac{1}{5}$	$\frac{3}{4}$
Black	TR10	Good	60	$\frac{1}{10}$	$\frac{1}{4}$
Black	TR10	Bad	20	$\frac{1}{5}$	$\frac{1}{4}$
Black	TR11	Good	60	$\frac{1}{10}$	$\frac{3}{4}$
Black	TR11	Bad	60	$\frac{1}{5}$	$\frac{3}{4}$

Check risk score 3/4 (all TR11 residents):

- White applicants: 40 good + 40 bad = 80 total
  - Defaults: $40 \times \frac{1}{10} + 40 \times \frac{1}{5} = 12$
  - Actual rate: $\frac{12}{80} = \frac{3}{20}$

Does This Satisfy Calibration Within Groups?

Redlining 1

Race	Zip	Credit	Number	Default rate	Risk score
White	TR10	Good	90	$\frac{1}{10}$	$\frac{1}{4}$
White	TR10	Bad	30	$\frac{1}{5}$	$\frac{1}{4}$
White	TR11	Good	40	$\frac{1}{10}$	$\frac{3}{4}$
White	TR11	Bad	40	$\frac{1}{5}$	$\frac{3}{4}$
Black	TR10	Good	60	$\frac{1}{10}$	$\frac{1}{4}$
Black	TR10	Bad	20	$\frac{1}{5}$	$\frac{1}{4}$
Black	TR11	Good	60	$\frac{1}{10}$	$\frac{3}{4}$
Black	TR11	Bad	60	$\frac{1}{5}$	$\frac{3}{4}$

Check risk score 3/4 (all TR11 residents):

- White applicants: 40 good + 40 bad = 80 total
  - Defaults: $40 \times \frac{1}{10} + 40 \times \frac{1}{5} = 12$
  - Actual rate: $\frac{12}{80} = \frac{3}{20}$
- Black applicants: 60 good + 60 bad = 120 total
  - Defaults: $60 \times \frac{1}{10} + 60 \times \frac{1}{5} = 18$
  - Actual rate: $\frac{18}{120} = \frac{3}{20}$

Calibration is Satisfied!

Calibration Within Groups (Weak): For each risk score, the actual default rate is the same across racial groups.

Risk Score	White Actual Rate	Black Actual Rate
1/4	1/8	1/8
3/4	3/20	3/20

Calibration is Satisfied!

Calibration Within Groups (Weak): For each risk score, the actual default rate is the same across racial groups.

Risk Score	White Actual Rate	Black Actual Rate
1/4	1/8	1/8
3/4	3/20	3/20

The algorithm satisfies weak calibration within groups.

Calibration is Satisfied!

Calibration Within Groups (Weak): For each risk score, the actual default rate is the same across racial groups.

Risk Score	White Actual Rate	Black Actual Rate
1/4	1/8	1/8
3/4	3/20	3/20

The algorithm satisfies weak calibration within groups.

But is it fair?

But the Algorithm is Clearly Unfair

Redlining 1

Race	Zip	Credit	Number	Default rate	Risk score
White	TR10	Good	90	$\frac{1}{10}$	$\frac{1}{4}$
White	TR10	Bad	30	$\frac{1}{5}$	$\frac{1}{4}$
White	TR11	Good	40	$\frac{1}{10}$	$\frac{3}{4}$
White	TR11	Bad	40	$\frac{1}{5}$	$\frac{3}{4}$
Black	TR10	Good	60	$\frac{1}{10}$	$\frac{1}{4}$
Black	TR10	Bad	20	$\frac{1}{5}$	$\frac{1}{4}$
Black	TR11	Good	60	$\frac{1}{10}$	$\frac{3}{4}$
Black	TR11	Bad	60	$\frac{1}{5}$	$\frac{3}{4}$

The problem: The algorithm ignores credit score and uses zip code as a proxy for race.

But the Algorithm is Clearly Unfair

Redlining 1

Race	Zip	Credit	Number	Default rate	Risk score
White	TR10	Good	90	$\frac{1}{10}$	$\frac{1}{4}$
White	TR10	Bad	30	$\frac{1}{5}$	$\frac{1}{4}$
White	TR11	Good	40	$\frac{1}{10}$	$\frac{3}{4}$
White	TR11	Bad	40	$\frac{1}{5}$	$\frac{3}{4}$
Black	TR10	Good	60	$\frac{1}{10}$	$\frac{1}{4}$
Black	TR10	Bad	20	$\frac{1}{5}$	$\frac{1}{4}$
Black	TR11	Good	60	$\frac{1}{10}$	$\frac{3}{4}$
Black	TR11	Bad	60	$\frac{1}{5}$	$\frac{3}{4}$

The problem: The algorithm ignores credit score and uses zip code as a proxy for race.

- Credit score perfectly predicts default risk
- But TR11 has more Black residents than TR10
- So using zip code systematically disadvantages Black applicants

But the Algorithm is Clearly Unfair

Redlining 1

Race	Zip	Credit	Number	Default rate	Risk score
White	TR10	Good	90	$\frac{1}{10}$	$\frac{1}{4}$
White	TR10	Bad	30	$\frac{1}{5}$	$\frac{1}{4}$
White	TR11	Good	40	$\frac{1}{10}$	$\frac{3}{4}$
White	TR11	Bad	40	$\frac{1}{5}$	$\frac{3}{4}$
Black	TR10	Good	60	$\frac{1}{10}$	$\frac{1}{4}$
Black	TR10	Bad	20	$\frac{1}{5}$	$\frac{1}{4}$
Black	TR11	Good	60	$\frac{1}{10}$	$\frac{3}{4}$
Black	TR11	Bad	60	$\frac{1}{5}$	$\frac{3}{4}$

The problem: The algorithm ignores credit score and uses zip code as a proxy for race.

- Credit score perfectly predicts default risk
- But TR11 has more Black residents than TR10
- So using zip code systematically disadvantages Black applicants

Calibration within groups is not sufficient for fairness.

What aspects of the Redlining example generate the obvious unfairness.

What aspects of the Redlining example generate the obvious unfairness.

- If, as in the actual historical case, the creators of the algorithm crafted it with the intention of disadvantaging black applicants, then it is obvious that the designer's actions in designing and constructing the algorithm themselves constitute a source of injustice and unfairness.

What aspects of the Redlining example generate the obvious unfairness.

- If, as in the actual historical case, the creators of the algorithm crafted it with the intention of disadvantaging black applicants, then it is obvious that the designer's actions in designing and constructing the algorithm themselves constitute a source of injustice and unfairness.

- Even if the designers of the algorithm did not explicitly intend to disadvantage black applicants, one could argue that the correlations between race, zip code and default rates are themselves the product of unjust social economic historical trends, and hence that it is unjust to apply an algorithm that exploits those correlations without recognizing, and in some way compensating for, their unjust historical origin.

Is there anything intrinsically unfair about the redlining algorithm or its predictions in and of themselves?

Is there anything intrinsically unfair about the redlining algorithm or its predictions in and of themselves?

Perhaps the most obvious thing to say here is that the algorithm is intrinsically unfair simply in virtue of its using zip codes as a **proxy** for race.

Problem: There is good reason to think that fairness sometimes requires predictive algorithms to explicitly base their predictions on group membership traits like gender and race.

Problem: There is good reason to think that fairness sometimes requires predictive algorithms to explicitly base their predictions on group membership traits like gender and race.

> *[I]t is often necessary for equitable risk assessment algorithms to explicitly consider protected characteristics. In the criminal justice system, for example, women are typically less likely to commit a future violent crime than men with similar criminal histories. As a result, gender-neutral risk scores can systematically overestimate a woman's recidivism risk, and can in turn encourage unnecessarily harsh judicial decisions. Recognizing this problem, some jurisdictions, like Wisconsin, have turned to gender-specific risk assessment tools to ensure that estimates are not biased against women.* *(Corbett-Davies and Goel)*

Sam Corbett-Davies and Sharad Goel (2018). *The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning*. https://arxiv.org/abs/1808.00023.

It is difficult to define exactly what it means for a predictive feature to be used as a proxy for a group membership trait.

It is difficult to define exactly what it means for a predictive feature to be used as a proxy for a group membership trait.

On what grounds can one say that zip code counts as a proxy for race in the above case, while other variables that are also correlated with race do not count as proxies?

This problem is further compounded when we recall that the predictive algorithms whose fairness we hope to assess are often proprietary, meaning that we do not actually know exactly which predictive features are being employed by the algorithm.

It is clear that merely citing the use of a proxy variable does not helpfully identify what is intrinsically wrong with the algorithm in the Redlining example.

If the algorithm used some other features rather than zip code to obtain the same predictions, it would still be just as unfair.

It is clear that merely citing the use of a proxy variable does not helpfully identify what is intrinsically wrong with the algorithm in the Redlining example.

If the algorithm used some other features rather than zip code to obtain the same predictions, it would still be just as unfair.

Claim: There is something *intrinsically unfair* in the predictions themselves, and that we should not need to refer to the predictive features used by the algorithm in order to diagnose that unfairness.

But as we have just seen, the most popular statistical criterion of algorithmic fairness from the literature, calibration within groups, is unable to identify any unfairness.

We need a new criterion to help us clearly diagnose the sense in which the predictions of the algorithm in the Redlining example are intrinsically unfair.

Base Rate Tracking

Redlining 1

Race	Zip	Credit	Number	Default rate	Risk score
White	TR10	Good	90	$\frac{1}{10}$	$\frac{1}{4}$
White	TR10	Bad	30	$\frac{1}{5}$	$\frac{1}{4}$
White	TR11	Good	40	$\frac{1}{10}$	$\frac{3}{4}$
White	TR11	Bad	40	$\frac{1}{5}$	$\frac{3}{4}$
Black	TR10	Good	60	$\frac{1}{10}$	$\frac{1}{4}$
Black	TR10	Bad	20	$\frac{1}{5}$	$\frac{1}{4}$
Black	TR11	Good	60	$\frac{1}{10}$	$\frac{3}{4}$
Black	TR11	Bad	60	$\frac{1}{5}$	$\frac{3}{4}$

- Average risk score for white applicants is

$$(90 * \frac{1}{4} + 30 * \frac{1}{4} + 40 * \frac{3}{4} + 40 * \frac{3}{4}) / 200 = 9/20.$$

Base Rate Tracking

Redlining 1

Race	Zip	Credit	Number	Default rate	Risk score
White	TR10	Good	90	$\frac{1}{10}$	$\frac{1}{4}$
White	TR10	Bad	30	$\frac{1}{5}$	$\frac{1}{4}$
White	TR11	Good	40	$\frac{1}{10}$	$\frac{3}{4}$
White	TR11	Bad	40	$\frac{1}{5}$	$\frac{3}{4}$
Black	TR10	Good	60	$\frac{1}{10}$	$\frac{1}{4}$
Black	TR10	Bad	20	$\frac{1}{5}$	$\frac{1}{4}$
Black	TR11	Good	60	$\frac{1}{10}$	$\frac{3}{4}$
Black	TR11	Bad	60	$\frac{1}{5}$	$\frac{3}{4}$

- Average risk score for black applicants is

$$(60 * \frac{1}{4} + 20 * \frac{1}{4} + 60 * \frac{3}{4} + 60 * \frac{3}{4})/200 = 11/20.$$

Base Rate Tracking

Redlining 1

Race	Zip	Credit	Number	Default rate	Risk score
White	TR10	Good	90	$\frac{1}{10}$	$\frac{1}{4}$
White	TR10	Bad	30	$\frac{1}{5}$	$\frac{1}{4}$
White	TR11	Good	40	$\frac{1}{10}$	$\frac{3}{4}$
White	TR11	Bad	40	$\frac{1}{5}$	$\frac{3}{4}$
Black	TR10	Good	60	$\frac{1}{10}$	$\frac{1}{4}$
Black	TR10	Bad	20	$\frac{1}{5}$	$\frac{1}{4}$
Black	TR11	Good	60	$\frac{1}{10}$	$\frac{3}{4}$
Black	TR11	Bad	60	$\frac{1}{5}$	$\frac{3}{4}$

- Average default rate for white applicants is

$$(90 * \frac{1}{10} + 30 * \frac{1}{5} + 40 * \frac{1}{10} + 40 * \frac{1}{5}) / 200 = 27/200.$$

Base Rate Tracking

Redlining 1

Race	Zip	Credit	Number	Default rate	Risk score
White	TR10	Good	90	$\frac{1}{10}$	$\frac{1}{4}$
White	TR10	Bad	30	$\frac{1}{5}$	$\frac{1}{4}$
White	TR11	Good	40	$\frac{1}{10}$	$\frac{3}{4}$
White	TR11	Bad	40	$\frac{1}{5}$	$\frac{3}{4}$
Black	TR10	Good	60	$\frac{1}{10}$	$\frac{1}{4}$
Black	TR10	Bad	20	$\frac{1}{5}$	$\frac{1}{4}$
Black	TR11	Good	60	$\frac{1}{10}$	$\frac{3}{4}$
Black	TR11	Bad	60	$\frac{1}{5}$	$\frac{3}{4}$

- Average default rate for black applicants is

$$(60 * \frac{1}{10} + 20 * \frac{1}{5} + 60 * \frac{1}{10} + 60 * \frac{1}{5}) / 200 = 28/200.$$

Base Rate Tracking

	Black	White	Difference
Average Risk Score	11/20	9/20	2/20
Average Default Rate	28/200	27/200	1/200

The difference between the average risk scores of the two groups is 20 times as great as the difference between their actual default rates.

Base Rate Tracking

If an algorithm assigns one group a higher average risk score than another, that discrepancy has to be justified by a corresponding discrepancy between the base rates of those two groups, and the magnitudes of those discrepancies should be equivalent.

Base Rate Tracking

In slogan form: an algorithm should only treat one groups as much more risky than another if it really is much more risky.

Base Rate Tracking: The difference between the average risk scores assigned to the relevant groups should be equal to the difference between the (expected) base rates of those groups.

Applying Base Rate Tracking to the Redlining algorithm:

Since the difference between the average risk scores assigned to white and black applicants is 20 times greater than the corresponding difference between their base rates, we can say that the algorithm treats black applicants unfairly in comparison to white applicants.

Applying Base Rate Tracking to the Redlining algorithm:

Since the difference between the average risk scores assigned to white and black applicants is 20 times greater than the corresponding difference between their base rates, we can say that the algorithm treats black applicants unfairly in comparison to white applicants.

If we were to rely only on calibration within groups, then we would need to refer to the designers' intentions, or the unjust historical origins of the relevant correlations, or the internal workings of the algorithm, in order to diagnose the unfairness in this case.

Applying Base Rate Tracking to the Redlining algorithm:

Since the difference between the average risk scores assigned to white and black applicants is 20 times greater than the corresponding difference between their base rates, we can say that the algorithm treats black applicants unfairly in comparison to white applicants.

If we were to rely only on calibration within groups, then we would need to refer to the designers' intentions, or the unjust historical origins of the relevant correlations, or the internal workings of the algorithm, in order to diagnose the unfairness in this case.

But base rate tracking allows us to directly identify the algorithm as intrinsically unfair on the basis of its predictions alone.

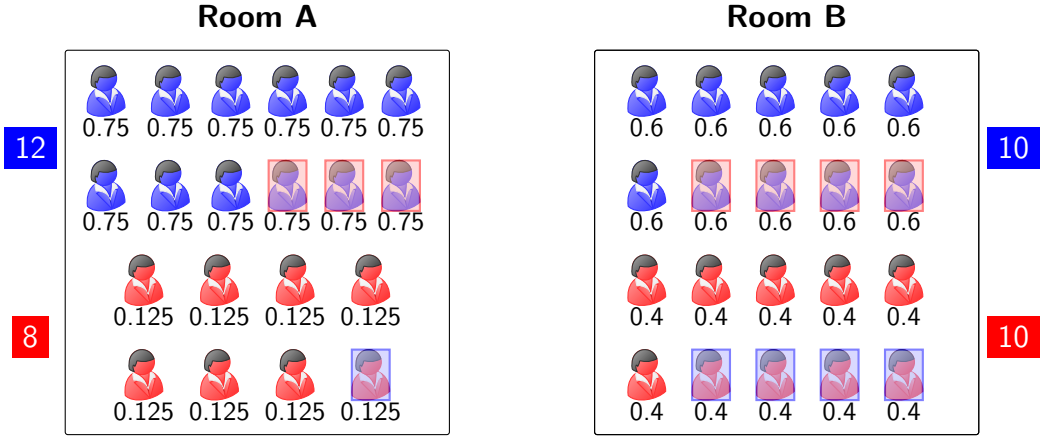
Unlike calibration within groups, base rate tracking really is a statistical criterion of algorithmic fairness, i.e., a necessary condition that any fair algorithm must satisfy.

Base Rate Tracking allows us to directly identify the algorithm as intrinsically unfair on the basis of its predictions alone. Given the lack of information that is generally available regarding the design process and internal architecture of predictive algorithms, this is important, since it shows that base rate tracking allows us to identify algorithmic unfairness in many cases where we would otherwise be unable to do so.

Base Rate Tracking is motivated by a natural philosophical intuition regarding the nature of fairness: that any difference in the way that an algorithm treats two groups needs to be justified by a corresponding difference in the relevant behaviors/properties of the two groups.

It is unfair to treat white loan applicants as if they have a much lower average risk of defaulting compared to black applicants if they do not actually have a much lower default rate.

Base Rate Tracking, unlike the 10 influential criteria of fairness discussed last week, is not undermined by Hedden's counterexample.



Since the base rates for the two rooms are equal to the average risk scores assigned to the people in those rooms, base rate tracking is trivially satisfied by the optimal predictive algorithm.

Base Rate Tracking...

Base Rate Tracking...

1. ...is motivated by a simple and powerful philosophical intuition about the nature of fairness,

Base Rate Tracking...

1. ...is motivated by a simple and powerful philosophical intuition about the nature of fairness,

2. ...is not undermined by Hedden's coin flipping example or the insurance pricing example, and

Base Rate Tracking...

1. ...is motivated by a simple and powerful philosophical intuition about the nature of fairness,

2. ...is not undermined by Hedden's coin flipping example or the insurance pricing example, and

3. ...significantly expands the diagnostic scope of calibration within groups in some important cases.

A Possible Objection

Going back to Redlining 1, base rate tracking successfully identifies the fact that the algorithm is unfair to black applicants, because the difference between the average risk scores of white and black applicants is far greater than the difference between their base rates.

A Possible Objection

Going back to Redlining 1, base rate tracking successfully identifies the fact that the algorithm is unfair to black applicants, because the difference between the average risk scores of white and black applicants is far greater than the difference between their base rates. **However, base rate tracking still requires that white applicants should be assigned a lower average risk score than black applicants, since black applicants have a higher overall default rate.**

A Possible Objection

Going back to Redlining 1, base rate tracking successfully identifies the fact that the algorithm is unfair to black applicants, because the difference between the average risk scores of white and black applicants is far greater than the difference between their base rates. **However, base rate tracking still requires that white applicants should be assigned a lower average risk score than black applicants, since black applicants have a higher overall default rate.**

And one might plausibly object that this is obviously unfair, since black applicants have the same default rate as white applicants within any given zip code.

This in turn implies that base rate tracking is not a plausible statistical criterion of algorithmic fairness.

Response

If the algorithm was designed to disadvantage black applicants, or if the correlations upon which it relies are the product of unjust historical conditions, then those constitute independent sources of unfairness which need to be appropriately recognized and taken into account in the application of the algorithm.

Response

If the algorithm was designed to disadvantage black applicants, or if the correlations upon which it relies are the product of unjust historical conditions, then those constitute independent sources of unfairness which need to be appropriately recognized and taken into account in the application of the algorithm.

Of course, statistical criteria like base rate tracking are unable to directly diagnose these kinds of unfairness, since they concern the historical origins of the algorithm and the relevant correlations, rather than predictive properties of the algorithm itself.

One can recognize these sources of injustice without thinking that the algorithm and its predictions are themselves intrinsically unfair.

Sources of Unfairness

Base Rate Tracking measures the unfairness that is *intrinsic* to the algorithm, but there are other sources of unfairness of an algorithm:

Sources of Unfairness

Base Rate Tracking measures the unfairness that is *intrinsic* to the algorithm, but there are other sources of unfairness of an algorithm:

- Facts regarding the unjust historical conditions that gave rise to the correlations exploited by the algorithm.
- Facts about the unjust intentions of the algorithm's designers.

Concluding Remarks

While statistical criteria like Base Rate Tracking can play an important role in the fight against algorithmic unfairness, the hardest problem will be to develop mechanisms that properly identify and compensate for the way in which algorithms exploit correlations which themselves arise from unfair historical conditions.

It is important that we recognize this problem as distinct from the problem of diagnosing unfairness, that is, intrinsic to the way that a given algorithm makes predictions, since the tools we use to address the latter problem (statistical criteria of algorithmic fairness) are not well suited to addressing the former.

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

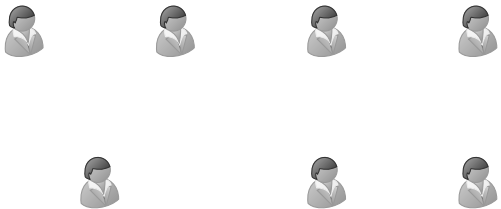
Research on algorithmic fairness studies the prospects of unbiased assessment. Bias in error rates is one form of bias, but not the only form and often considered not the most important form. Can bias in error rates and other important forms of bias be simultaneously eliminated?

One lesson that emerges from some of these studies is that eliminating one form of bias can mean that it is impossible to eliminate another. Sometimes, then, we face a conflict between eliminating different forms of bias.

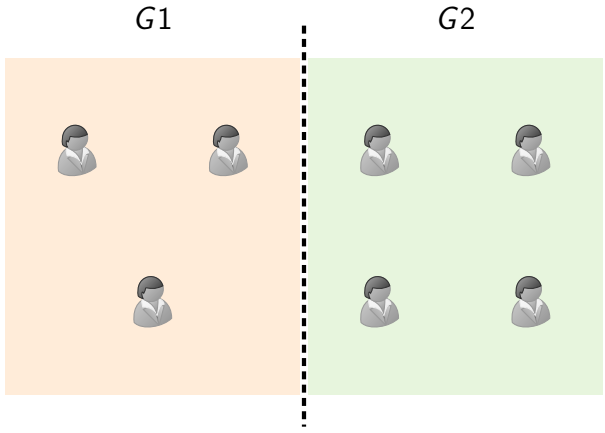
Not only do we face a conflict in eliminating different forms of bias, we also face a conflict in eliminating one form of bias across different groupings.

Eliminating a certain form of bias across groups for one way of categorizing people in a population can mean that it is impossible to eliminate that form of bias across groups for another way of classifying them.

Partitions

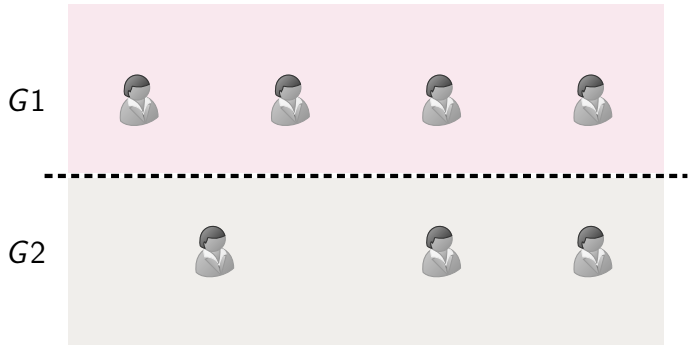


Partitions



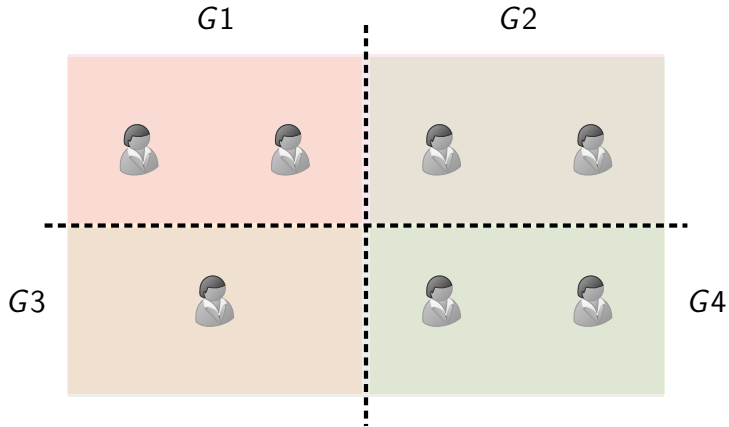
A partition is a way of carving up the population into non-overlapping groups.

Partitions



A partition is a way of carving up the population into non-overlapping groups.

Partitions



A partition is a way of carving up the population into non-overlapping groups.

Consider once again the bias found against black people in the COMPAS data.

In that same Broward County data set, there is a similar amount of bias in error rates against women compared to men, as a companion piece in ProPublica makes clear.

Consider once again the bias found against black people in the COMPAS data.

In that same Broward County data set, there is a similar amount of bias in error rates against women compared to men, as a companion piece in ProPublica makes clear.

Bias against either group is ethically relevant. Satisfying certain central fairness constraints for a race partition does not imply that those constraints are satisfied for a gender partition. Still other partitions could be pertinent.

Consider once again the bias found against black people in the COMPAS data.

In that same Broward County data set, there is a similar amount of bias in error rates against women compared to men, as a companion piece in ProPublica makes clear.

Bias against either group is ethically relevant. Satisfying certain central fairness constraints for a race partition does not imply that those constraints are satisfied for a gender partition. Still other partitions could be pertinent.

The relevant social identities cannot be decided a priori, without appeal to contingent social context and values.

Consider, for example, those who wear a size 8 shoe, or those born between nine and ten in the morning, local time. If size 8 shoes were to become extremely difficult to find then being someone who wears that shoe size may become an important part of one's identity and grounds for solidarity with those similarly unshod.

Consider, for example, those who wear a size 8 shoe, or those born between nine and ten in the morning, local time. If size 8 shoes were to become extremely difficult to find then being someone who wears that shoe size may become an important part of one's identity and grounds for solidarity with those similarly unshod.

Likewise, if an authoritarian ruler were to elect to severely curtail the freedoms of people born between nine and ten in the morning due to some supernatural belief or other, then the hour of one's birth and the persecution it entails for some is, again, likely to become an important aspect of one's identity and grounds for solidarity.

The priority of particular partitions in eliminating bias might reasonably depend not just on past history of discrimination, but also on current deprivation.

What groups suffer discrimination and deprivation is a matter to which we may frequently need to reattend.


2/3


2/3

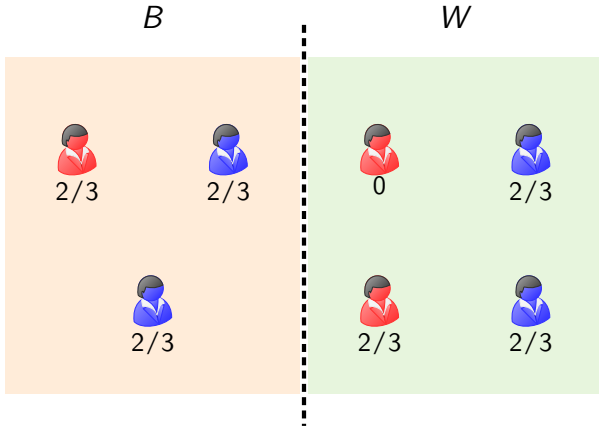

0


2/3

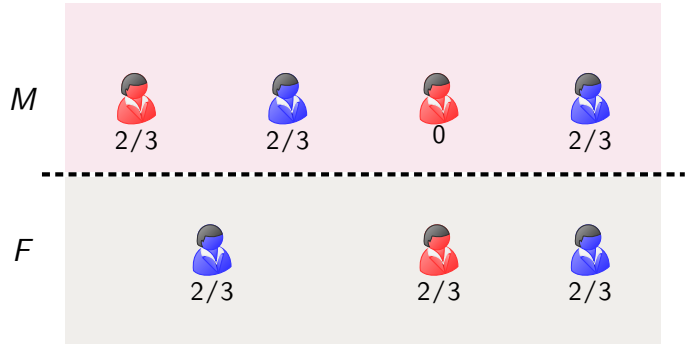

2/3


2/3

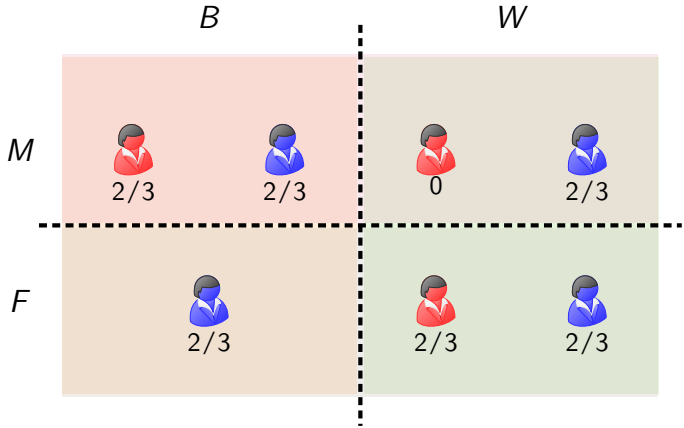

2/3



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 $W\&F$?