Fairness and Machine Learning

Nishant Mehta

Lecture 31

nature machine intelligence

Article https://doi.org/10.1038/s42256-025-00986-z

Large language models that replace human participants can harmfully misportray and flatten identity groups

Received: 12 February 2024

Accepted: 7 January 2025

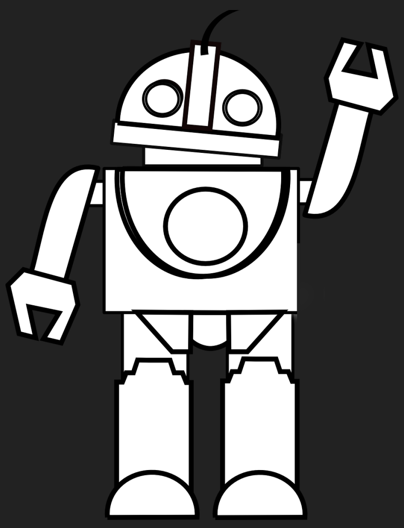
Published online: 17 February 2025

Check for updates

Angelina Wang[1], Jamie Morgenstern[2] & John P. Dickerson[3,4]

Large language models (LLMs) are increasing in capability and popularity, propelling their application in new domains—including as replacements for human participants in computational social science, user testing, annotation tasks and so on. In many settings, researchers seek to distribute their surveys to a sample of participants that are representative of the underlying human population of interest. This means that to be a suitable replacement, LLMs will need to be able to capture the influence of positionality (that is, the relevance of social identities like gender and race). However, we show that there are two inherent limitations in the way current LLMs are trained that prevent this. We argue analytically for why LLMs are likely to both misportray and flatten the representations of demographic groups, and then empirically show this on four LLMs through a series of human studies with 3,200 participants across 16 demographic identities. We also discuss a third limitation about how identity prompts can essentialize identities. Throughout, we connect each limitation to a pernicious history of epistemic injustice against the value of lived experiences that explains why replacement is harmful for marginalized demographic groups. Overall, we urge caution in use cases in which LLMs are intended to replace human participants whose identities are relevant to the task at hand. At the same time, in cases where the benefits of LLM replacement are determined to outweigh the harms (for example, engaging human participants may cause them harm, or the goal is to supplement rather than fully replace), we empirically demonstrate that our inference-time techniques reduce—but do not remove—these harms.

Can an algorithm be unethical?



Machine learning model class

Data

SGD

Fitted model

The New York Times

Facial Recognition Is Accurate, if You're a White Guy

By Steve Lohr

Feb. 9, 2018

Facial recognition technology is improving by leaps and bounds. Some commercial software can now tell the gender of a person in a photograph.

When the person in the photo is a white man, the software is right 99 percent of the time.

But the darker the skin, the more errors arise — up to nearly 35 percent for images of darker skinned women, according to a new study that breaks fresh ground by measuring how the technology works on people of different races and gender.

These disparate results, calculated by Joy Buolamwini, a researcher at the M.I.T. Media Lab, show how some of the biases in

Gender was misidentified in **up to 1 percent of lighter-skinned males** in a set of 385 photos.

Gender was misidentified in **up to 7 percent of lighter-skinned females** in a set of 296 photos.

Gender was misidentified in **up to 12 percent of darker-skinned males** in a set of 318 photos.

Gender was misidentified in **35 percent of darker-skinned females** in a set of 271 photos.

If societal harms aren’t always intentional, then what’s the solution?

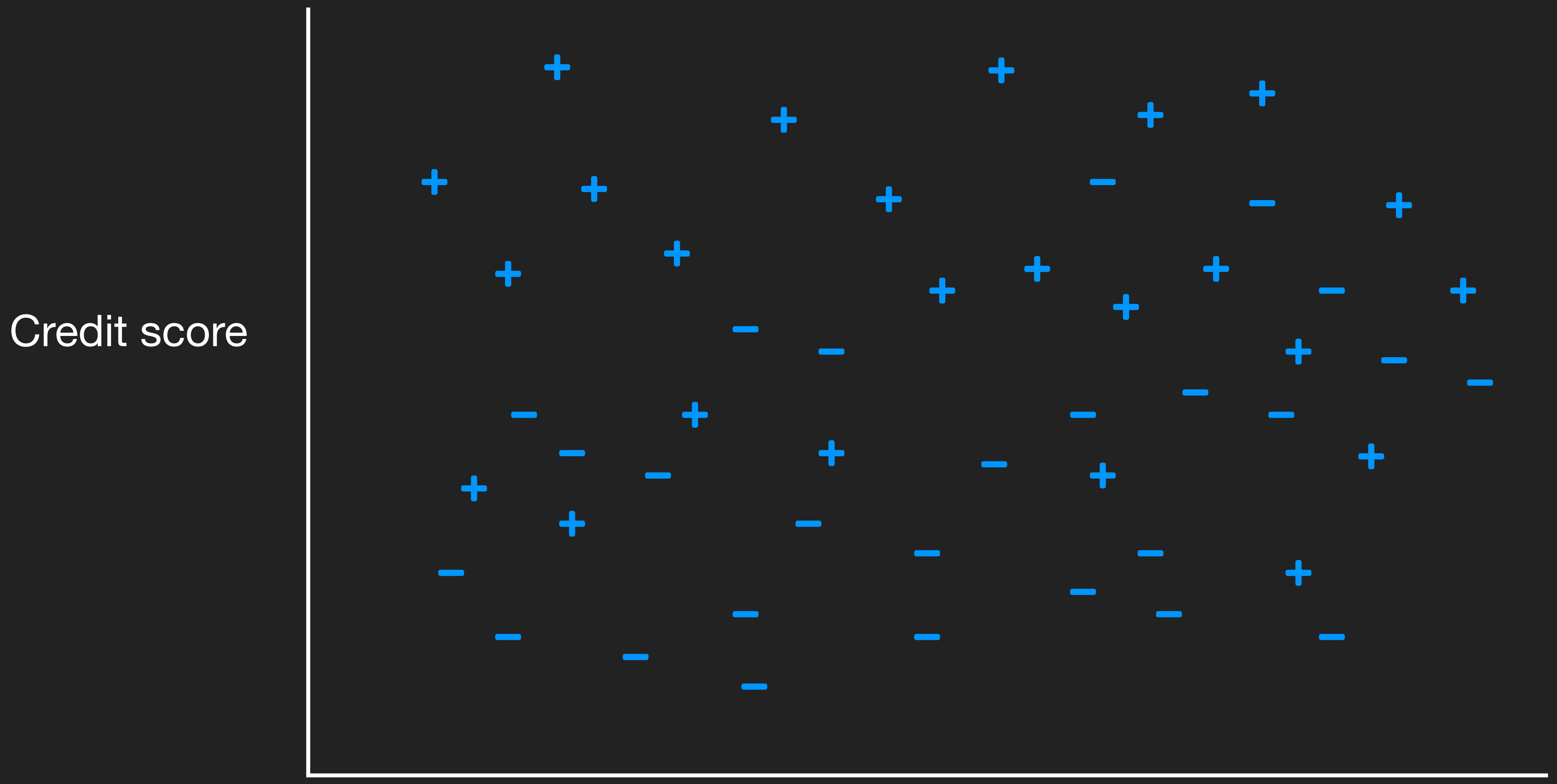
Fairness

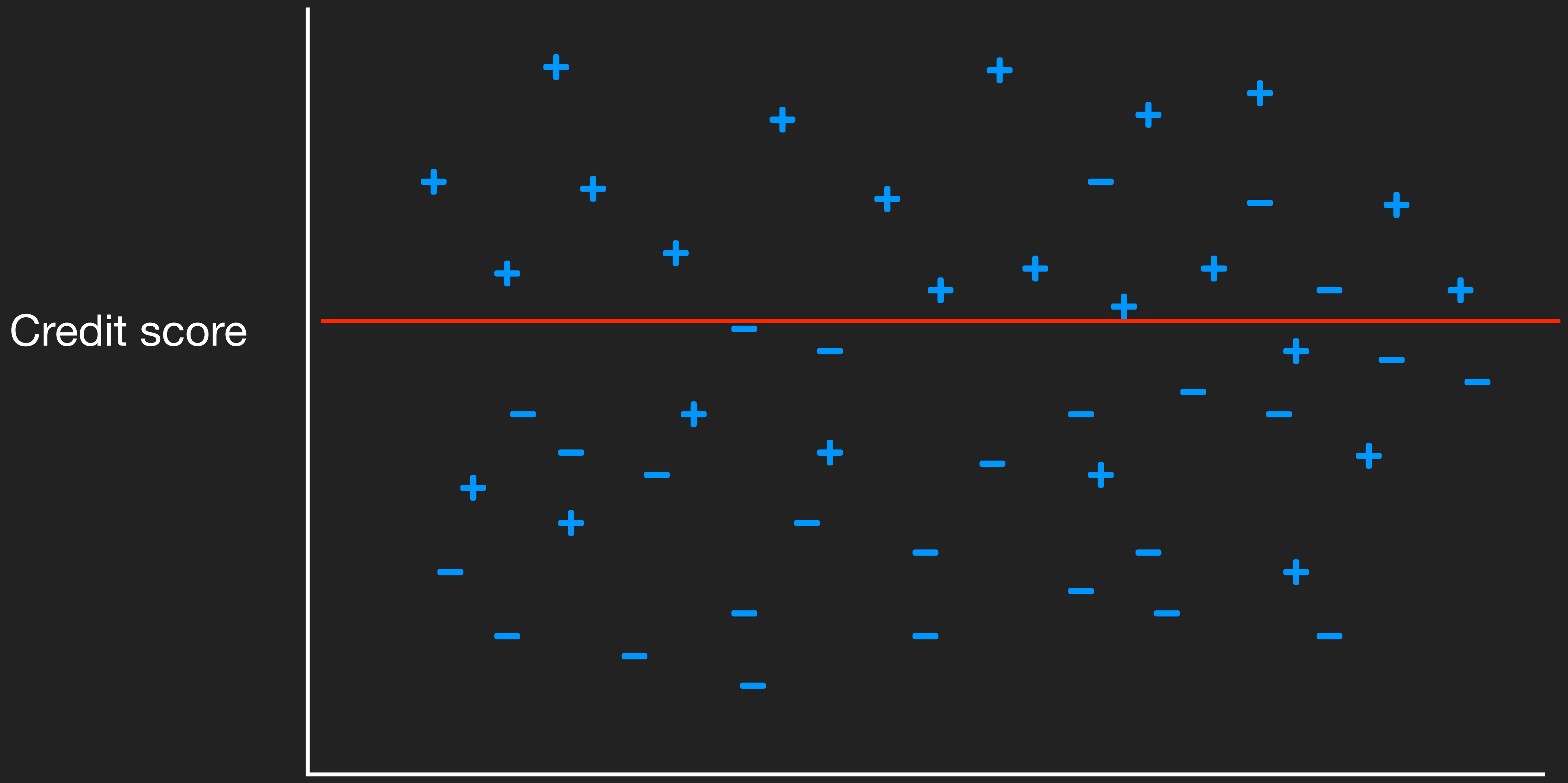
Bernard Parker, left, was rated high risk; Dylan Fugett was rated low risk. (Josh Ritchie for ProPublica)

Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

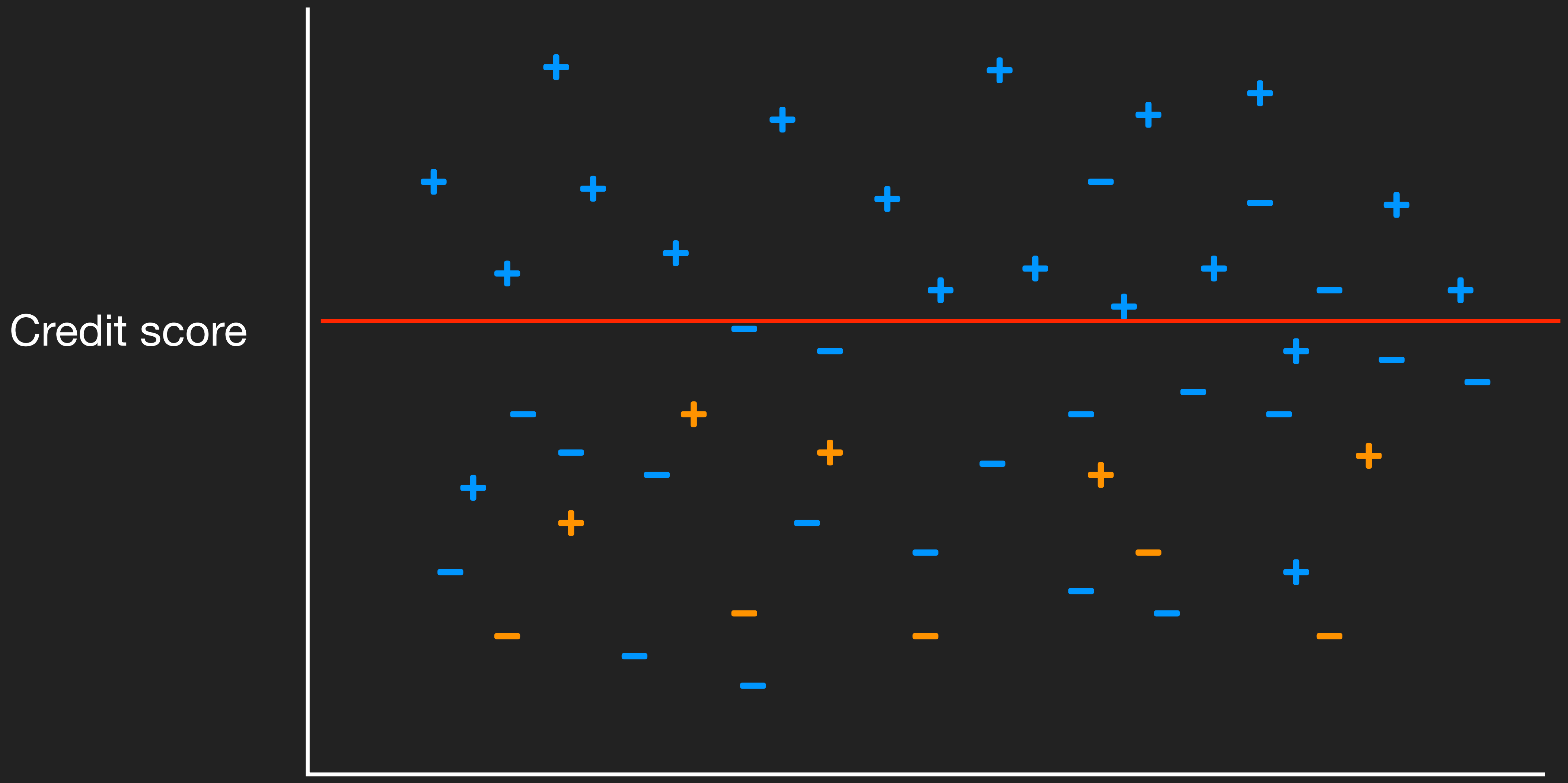
by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica
May 23, 2016


Credit score


Credit score

Two groups: blue people and orange people

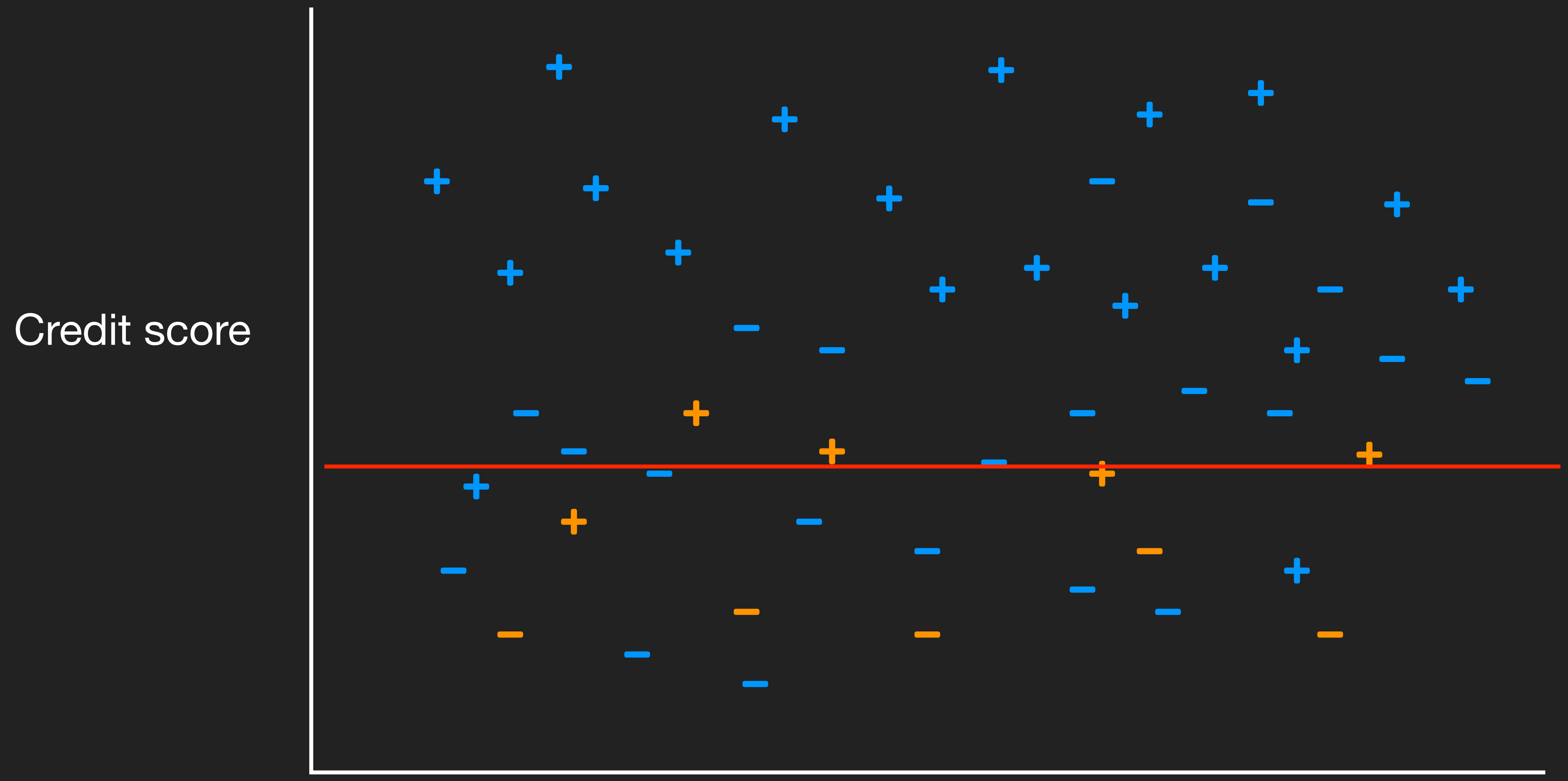
Is this is still a good choice of threshold?



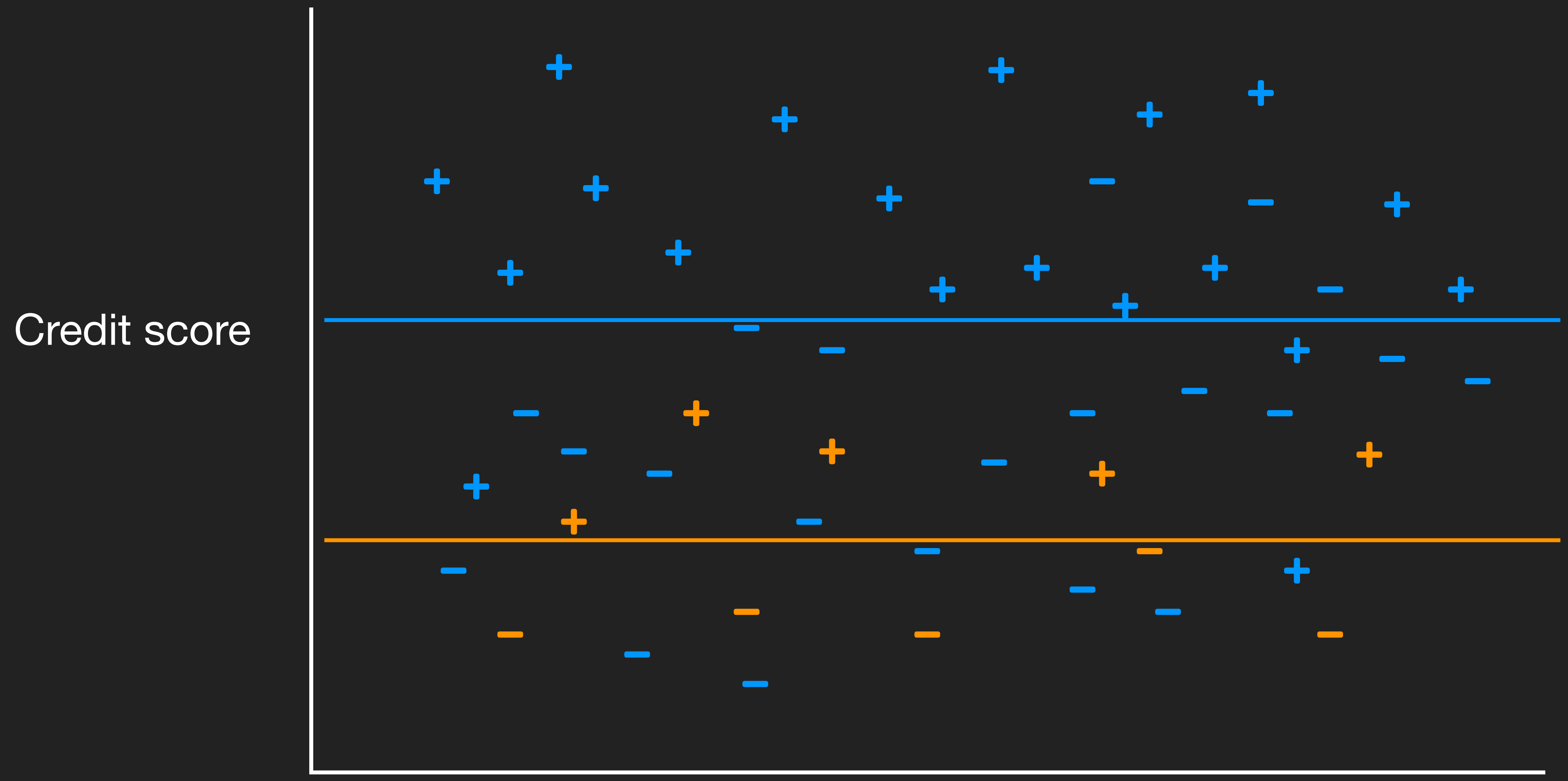
Maximin Principle: Minimize the maximum risk

$\Updownarrow$

Select threshold such that the worst-off group (group with maximum risk) has as small risk as possible



Or use separate thresholds, one per group



Fairness through unawareness?

OK, so it looks like having separate thresholds for blue and orange people boosts accuracy for both groups… but is this fair? Can a fair classifier use someone's color as input?

"Fairness is unawareness" - a fair predictor should not use protected attributes (like color)

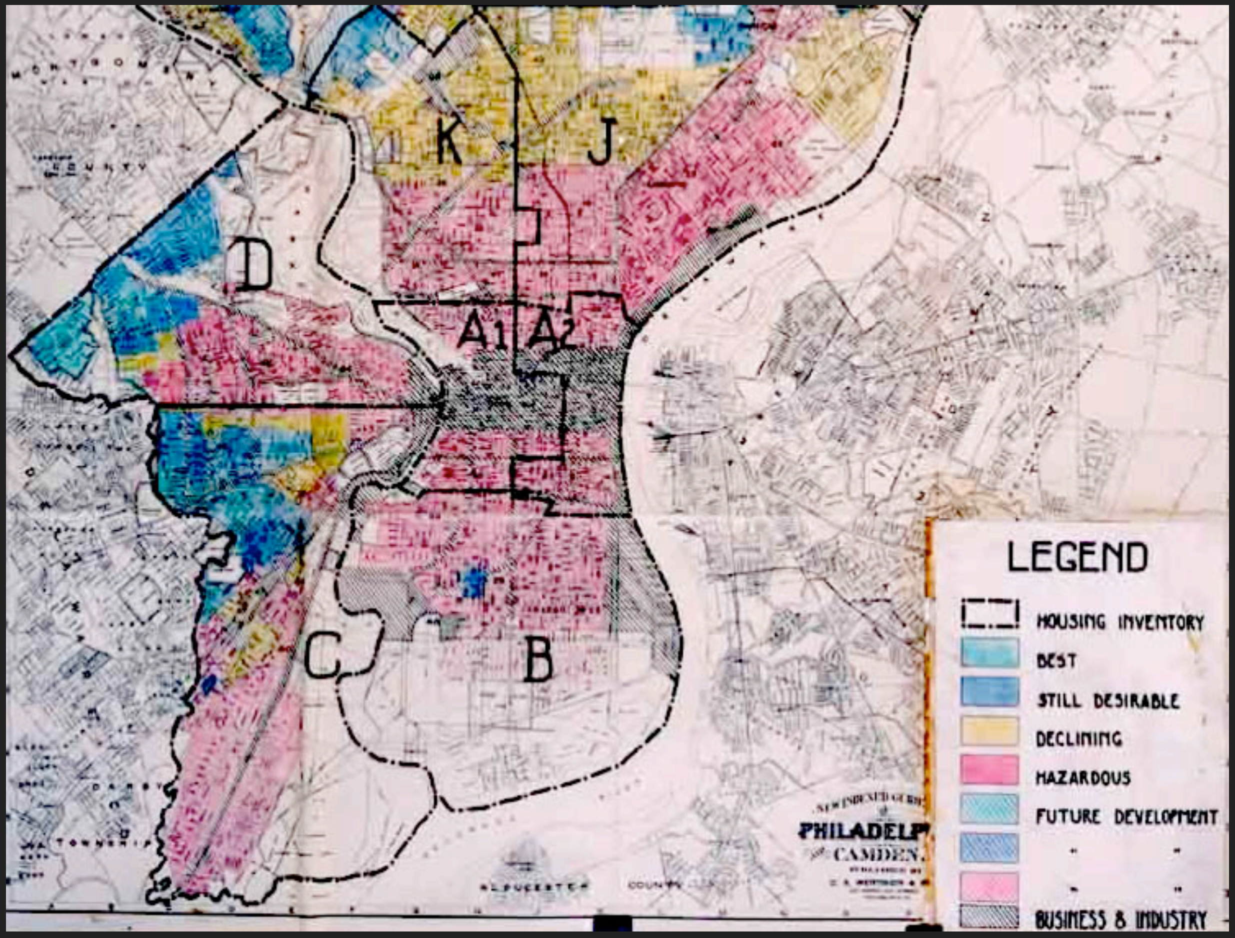
This is problematic though, due to *redundant encoding problem*

Redlining

But does the redundant encoding occur in the real world?

> "Redlining is the practice of arbitrarily denying or limiting financial services to specific neighborhoods, generally because its residents are people of color or are poor."

—D. Bradford Hunt (Redlining. Encyclopedia of Chicago, 2005)



Fairness is awareness

Much recent research has converged onto the point that *fairness is awareness*:

> If you want to avoid discrimination, you actually may need to form risk scores or classifications in a way that *considers the group to which a person belongs.*

Equality of opportunity

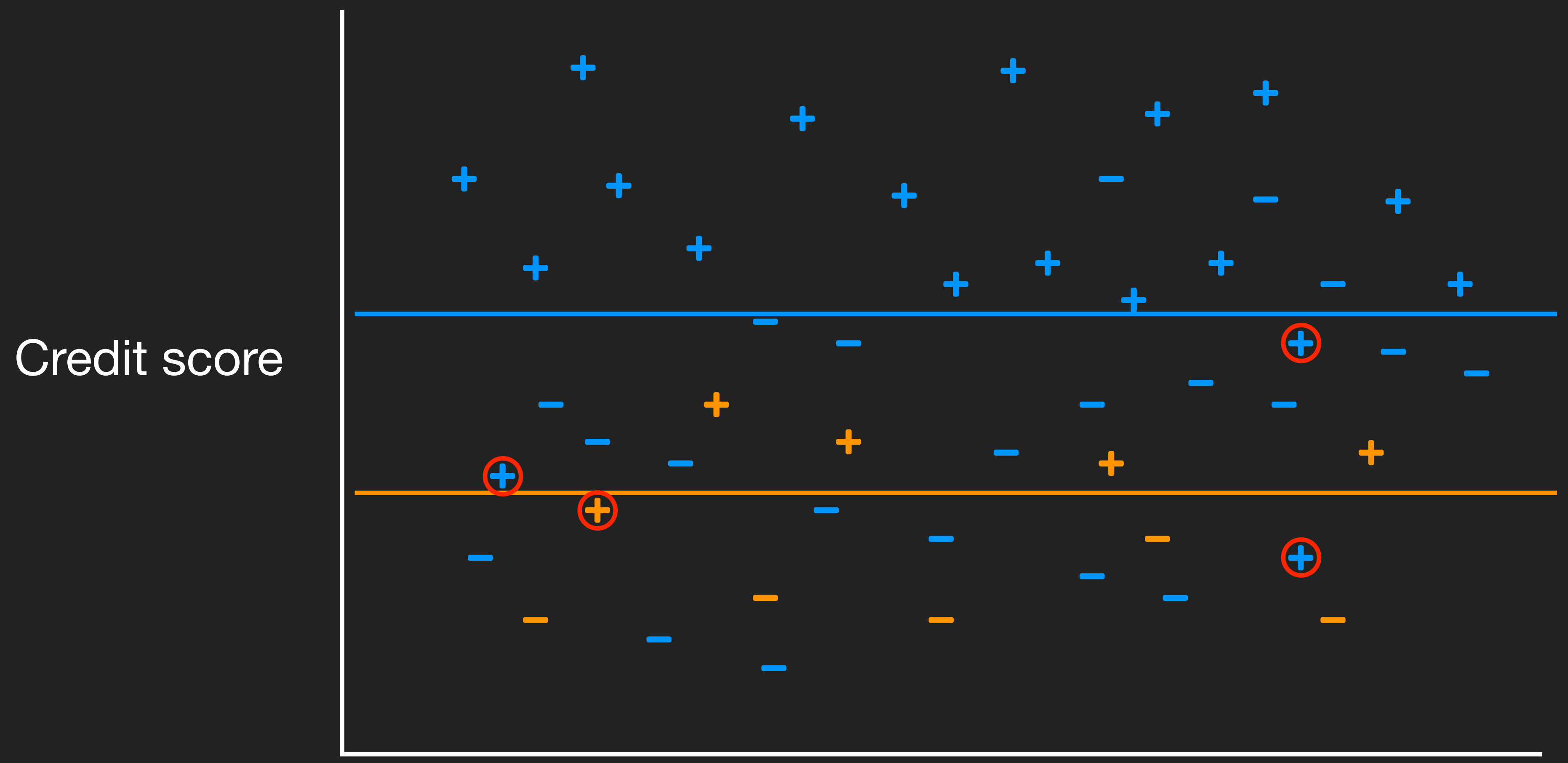
Positive label	Negative label
Will repay loan *(approved for a loan)* Will not reoffend *(granted parole)*	Will default on loan *(denied for a loan)* Will reoffend *(rejected for parole)*
Positive predictions benefit person	Negative predictions harm person

- False negative rate: among those individuals who belong to the positive class, what fraction of them do we label as negative (and so deprive them of opportunity)
- Equality of opportunity - strive to equalize the false negative rate across groups

How can we achieve equality of opportunity?

In many cases, we first compute risk scores for each group (probability of defaulting on loan, probability of reoffending)

We can then use separate thresholds for each group to equalize the false negative rate



Considering how a classification will be used

What happens when the perception of a prediction is different for different groups?

Example:

Among blue people that are predicted to reoffend, 70% of them actually reoffend. Among green people that are predicted to reoffend, 90% of them actually reoffend.

If you are a judge and see two people, a blue person and a green person, both predicted to reoffend, who will you be more lenient towards?

Solution: equalize negative predictive value (above, these were 70% and 90%), so that the same prediction has the same meaning to the user, regardless of group identity

Impossibility

Unless either perfect prediction is possible or groups have the same base rate, it is impossible to satisfy all 3 of the below fairness conditions:

(1) Equalized <u>Positive Predictive Value</u>

Among people to whom you gave loans, what proportion will repay the loan?

(2) Equalized <u>False Positive Rate</u>

Think of this as equality of "bank error in your favor." Among people that would default on a loan, to what proportion do you give a loan?

(3) Equalized <u>False Negative Rate</u>

Think of this as equality of opportunity. Among people that would repay a loan, to what proportion do you not give a loan?

Further reading: [Miconi. 2017. The impossibility of "fairness": a generalized impossibility result for decisions.]

Inherent trade-offs

Often, achieving fairness requires that predictions are intentionally made *less accurate* for one group if its labels are easier to predict than another group's labels

Fair Recidivism Prediction

- **Goal:** Release individuals if and only if, were they released, they will not commit a crime in the next 2 years

- Fairness constraint: Strive for
    - Similar false positive rates across the groups
      (released a reoffender)
    - Similar false negative rates across the groups
      (did not release a non-reoffender)

Fair Recidivism Prediction

- **Goal:** Release individuals if and only if, were they released, they will not commit a crime in the next 2 years

- Fairness constraint: Strive for
  - Similar false positive rates across the groups
    (released a reoffender)
  - Similar false negative rates across the groups
    (did not release a non-reoffender)

But… unreleased individuals won't reveal their true label!

If we do not even know the false negative rates,
how can we balance them across the groups?!

Fair Recidivism Prediction

- **Goal:** Release individuals if and only if, were they released, they will not commit a crime in the next 2 years

- Fairness constraint: Strive for

Key Difficulty:
Decisions made by the algorithm affect what it learns!

(did not release a non-reoffender)

But… unreleased individuals won't reveal their true label!

If we do not even know the false negative rates,
how can we balance them across the groups?!

The Fair Apple Tasting Problem

This type of feedback is already captured by an old framework, known as **Apple Tasting**

You encounter a sequence of apples; each apple can be good or bad

When an apple arrives, you decide whether or not to taste the apple.

Mistakes involve either tasting a bad apple or not tasting a good apple
(releasing a reoffender) (not releasing a non-reoffender)

Not tasting a good apple is a mistake the learner does not know about!

The Fair Apple Tasting Problem

This type of feedback is already captured by an old framework, known as **Apple Tasting**

You encounter a sequence of apples; each apple can be good or bad

When an apple arrives, you decide whether or not to taste the apple.

taste
label: good

Mistakes involve either tasting a bad apple or not tasting a good apple
(releasing a reoffender) (not releasing a non-reoffender)

Not tasting a good apple is a mistake the learner does not know about!

The Fair Apple Tasting Problem

This type of feedback is already captured by an old framework, known as **Apple Tasting**

You encounter a sequence of apples; each apple can be good or bad

When an apple arrives, you decide whether or not to taste the apple.

taste
label: good

Mistakes involve either tasting a bad apple or not tasting a good apple
(releasing a reoffender) (not releasing a non-reoffender)

Not tasting a good apple is a mistake the learner does not know about!

The Fair Apple Tasting Problem

This type of feedback is already captured by an old framework, known as **Apple Tasting**

You encounter a sequence of apples; each apple can be good or bad

When an apple arrives, you decide whether or not to taste the apple.

Mistakes involve either tasting a bad apple or not tasting a good apple
(releasing a reoffender) (not releasing a non-reoffender)

Not tasting a good apple is a mistake the learner does not know about!

The Fair Apple Tasting Problem

This type of feedback is already captured by an old framework, known as **Apple Tasting**

You encounter a sequence of apples; each apple can be good or bad

When an apple arrives, you decide whether or not to taste the apple.

Mistakes involve either tasting a bad apple or not tasting a good apple

(releasing a reoffender) (not releasing a non-reoffender)

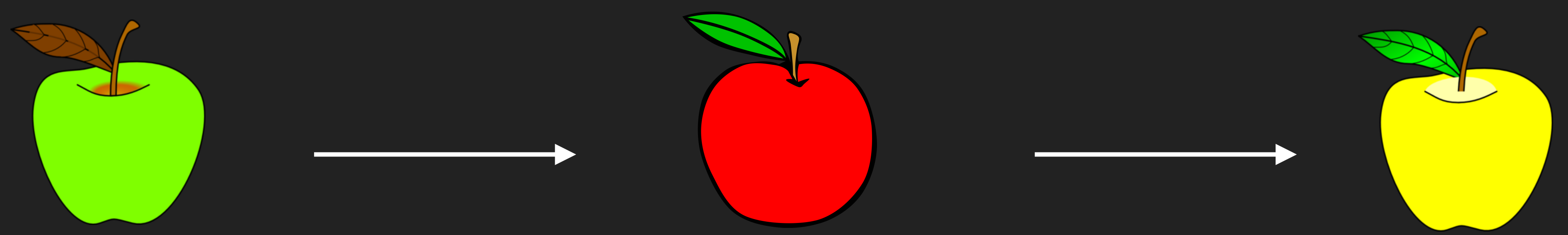
Not tasting a good apple is a mistake the learner does not know about!

The Fair Apple Tasting Problem

This type of feedback is already captured by an old framework, known as **Apple Tasting**

You encounter a sequence of apples; each apple can be good or bad

When an apple arrives, you decide whether or not to taste the apple.



taste label: good	taste label: bad ✗	don't taste label: unknown ✗?

Mistakes involve either tasting a bad apple or not tasting a good apple
(releasing a reoffender) (not releasing a non-reoffender)

Not tasting a good apple is a mistake the learner does not know about!

Feedback loops

Predictive Policing

Given historical crime incident data for a collection of regions, decide how to allocate patrol officers to areas to detect crime.

- Police predict more crime will happen in Area X
- So they put more police in Area X
- They measure more crime in Area X
- So they predict more crime will happen in Area X