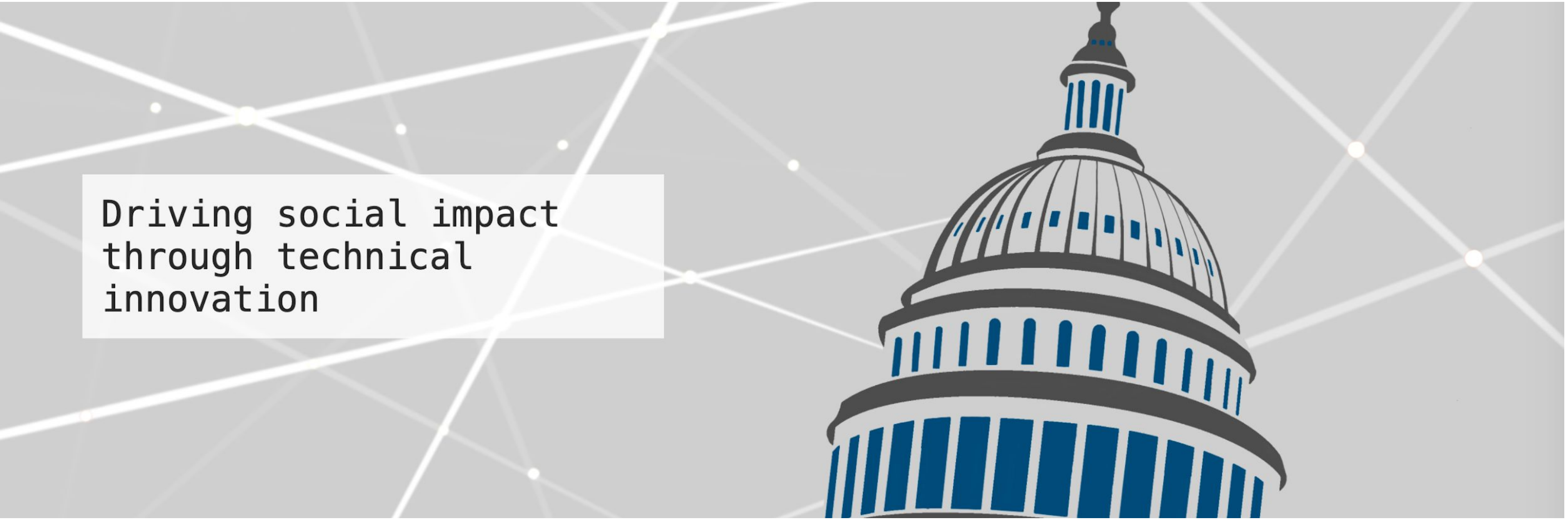
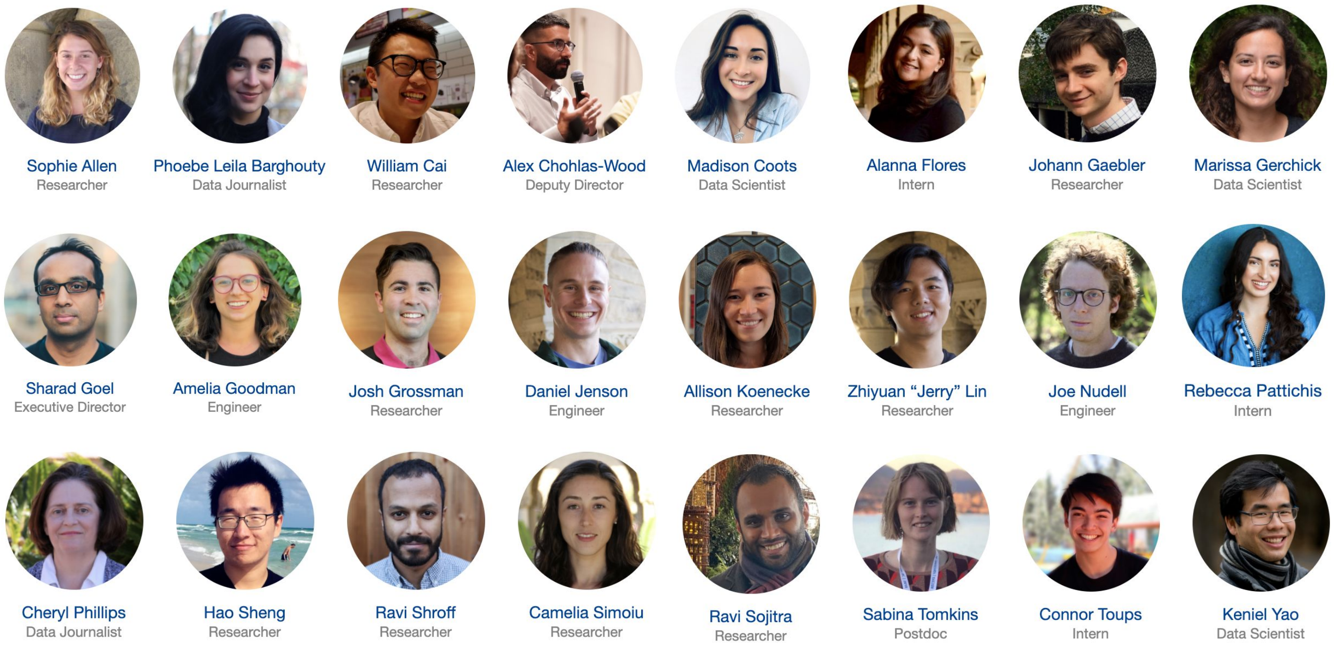
Designing for Equity

Sharad Goel
Stanford University

Stanford Computational Policy Lab

policylab.stanford.edu





We're an interdisciplinary team of researchers, engineers, and journalists that use technology to drive reform in criminal justice, education, healthcare, and beyond.

Part I: Racial disparities in automated speech recognition

References

Racial disparities in automated speech recognition
Proceedings of the National Academies of Science [2020]
Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R. Rickford, Dan Jurafsky, and Sharad Goel

Thanks to Allison Koenecke for help with the slides!

Automated speech recognition [ASR]

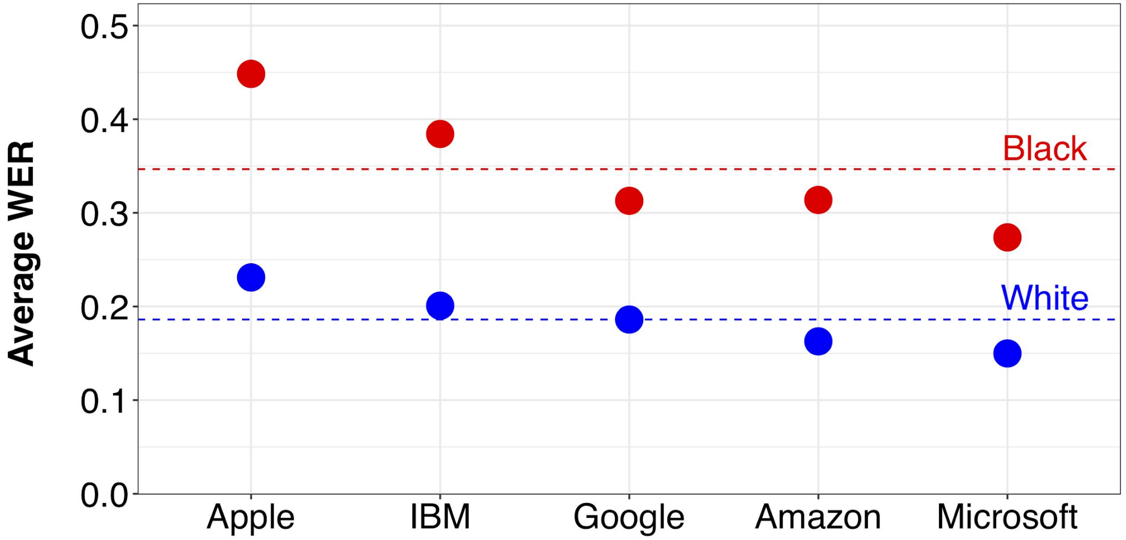
Automated speech-to-text systems are now widespread, powering **virtual assistants** [Siri, Alexa, Google Assistant], **dictation**, **translation**, **subtitling**, and **hands-free computing**.

Racial disparities in ASR systems

We audited five leading ASR providers [Amazon, Apple, Google, IBM, and Microsoft] by comparing human and machine-generated transcripts for 20 hours of audio from Black and white speakers.

Racial disparities in ASR systems

Error rates were twice as large for Black speakers



Racial disparities in ASR systems

~~Me~~ **I mean,** I ~~know I'm~~ **knew I was** kinda tall for ~~asking~~ **high school**. I didn't **wanna play center. I didn’t** because **center** ~~send it~~ **don’t** ~~on~~ have the ball that much. You get the ball occasionally when you in the post, I mean, but I didn't want to play it.

Error rates and AAVE

- African American Vernacular English is spoken by nearly 12% of all Americans

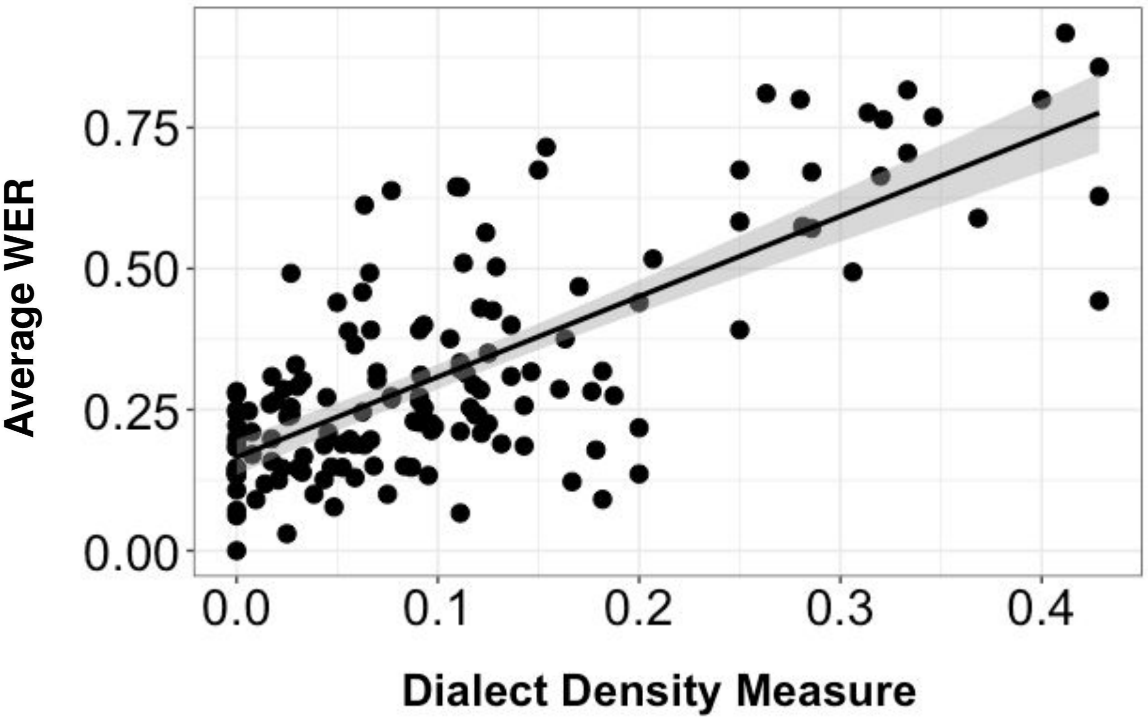
Error rates and AAVE

- African American Vernacular English is spoken by nearly 12% of all Americans
- We counted hand-coded AAVE linguistic features in random sample of audio snippets

Error rates and AAVE

- African American Vernacular English is spoken by nearly 12% of all Americans
- We counted hand-coded AAVE linguistic features in random sample of audio snippets
- Grammatical and phonological examples:
  - **Zero copula:** They gone
  - **Future *be*:** He be here tomorrow
  - **Final consonant cluster reduction:** band → ban’
  - **Hapology:** mississippi → misipi

Error rates and AAVE



The source of disparities

Modern ASR systems combine **language models** (that encode grammar) with **acoustic models**.

The source of disparities

Language models

The performance of a language model is often measured in terms of **perplexity**, which captures how well the model predicts the next word in a sequence.

The source of disparities

Language models

The performance of a language model is often measured in terms of **perplexity**, which captures how well the model predicts the next word in a sequence.

The dog jumped over the __________

`fence`	5%
`cup`	2%
`moon`	1%

The source of disparities

Language models

The performance of a language model is often measured in terms of **perplexity**, which captures how well the model predicts the next word in a sequence.

We find language models perform *better* on our sample of Black speakers than on our sample of white speakers.

The source of disparities
Acoustic models

- Find Black and white speakers saying identical phrases in our sample

The source of disparities

Acoustic models

- Find Black and white speakers saying identical phrases in our sample
- Match pairs of Black and white speakers (of the same gender and similar age) uttering 5 to 8 word n-grams
  - *“and then a lot of the”*
  - *“and my mother was a”*

The source of disparities

Acoustic models

- Find Black and white speakers saying identical phrases in our sample
- Match pairs of Black and white speakers (of the same gender and similar age) uttering 5 to 8 word n-grams
  - *“and then a lot of the”*
  - *“and my mother was a”*
- Compare error rates across the 206 matched phrases

The source of disparities

Acoustic models

On a subset of **identical phrases** spoken by Black and white individuals in our dataset, error rates were still about twice as large for Black speakers.

Call to action

- More diverse data should be collected: both of AAVE speech and of other varieties of English

Call to action

- More diverse data should be collected: both of AAVE speech and of other varieties of English
- The speech recognition community needs to invest resources to ensure ASR systems — and the institutions that build them — are broadly inclusive

Call to action

- More diverse data should be collected: both of AAVE speech and of other varieties of English
- The speech recognition community needs to invest resources to ensure ASR systems — and the institutions that build them — are broadly inclusive
- ASR developers should regularly assess and publicly report progress over time

Call to action

- More diverse data should be collected: both of AAVE speech and of other varieties of English
- The speech recognition community needs to invest resources to ensure ASR systems — and the institutions that build them — are broadly inclusive
- ASR developers should regularly assess and publicly report progress over time
- Learn from algorithmic & legislative progress made in other domains (e.g., computer vision)

Buolamwini, J., & Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. *FAT*.

Part II: A deontological approach to fairness

Risk assessment tools

Many high-stakes decisions are made by first estimating the likelihood of an adverse outcome based on the available information.

Lending is based on risk of default; pretrial detention is based on risk of recidivism.

Decisions guided by statistical risk assessments can, in theory, be more equitable than those made by intuition alone.

A mathematical definition of fairness

Classification parity

An algorithm is considered to be *fair* if error rates are [approximately] equal for white and Black defendants.

A mathematical definition of fairness

Proposed legislation in Idaho [2019]

“Pretrial risk assessment algorithms shall not be used … by the state until first shown to be **free of bias**, ...[meaning] that an algorithm has been formally tested and...the **rate of error is balanced** as between protected classes and those not in protected classes.”

[This requirement was removed from the final bill.]

A mathematical definition of fairness

False positive rate

A common mathematical definition of fairness is demanding equal false positive rates.

$$\text{False positive rate} = \frac{\text{Did not reoffend \& ``high risk''}}{\text{Did not reoffend}}$$

Error rate disparities in Broward County

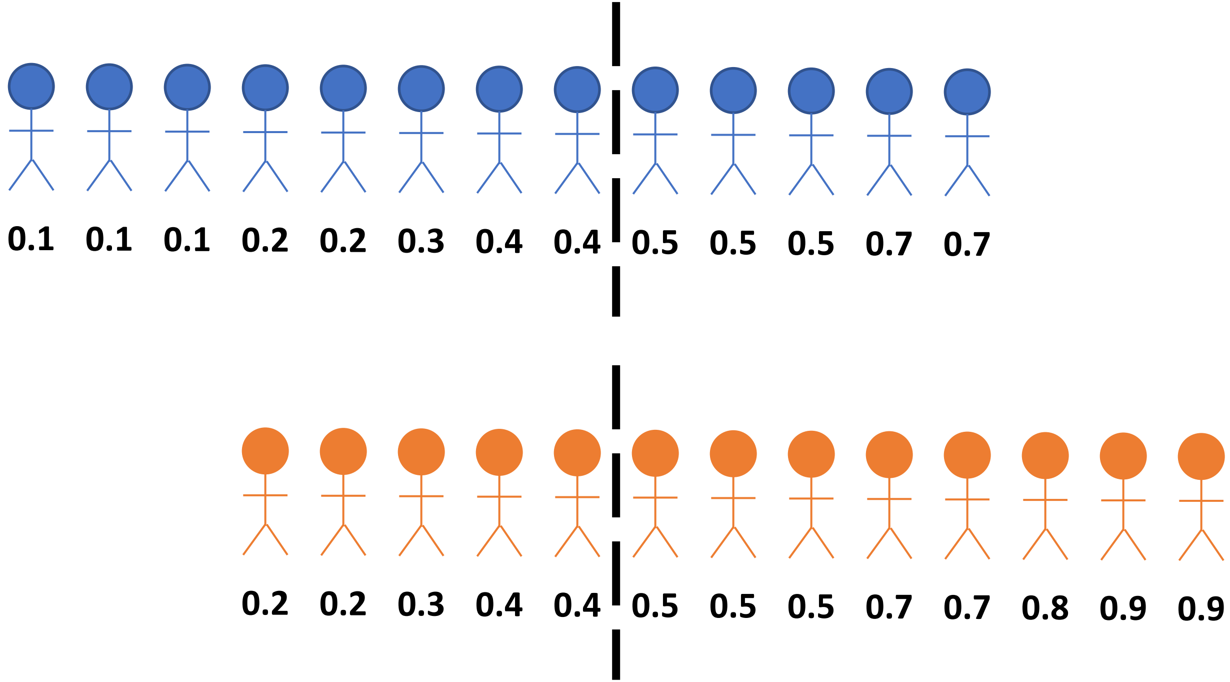
Via ProPublica

31%	vs.	15%
of Black defendants **who did not reoffend**		of white defendants **who did not reoffend**

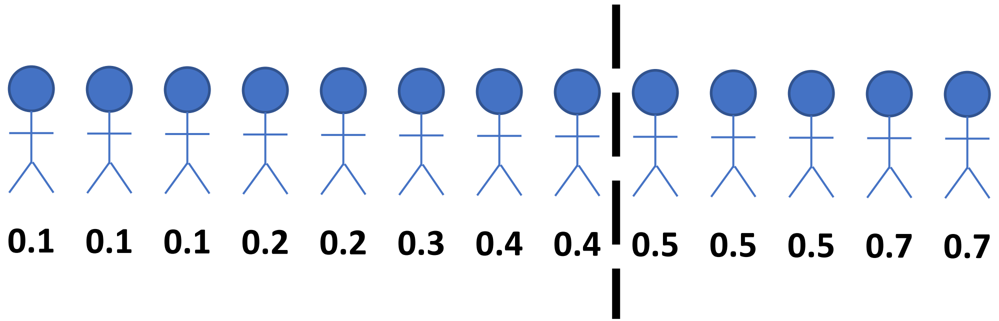
were deemed **high risk** of committing a violent crime

[Higher false positive rates for Black defendants]

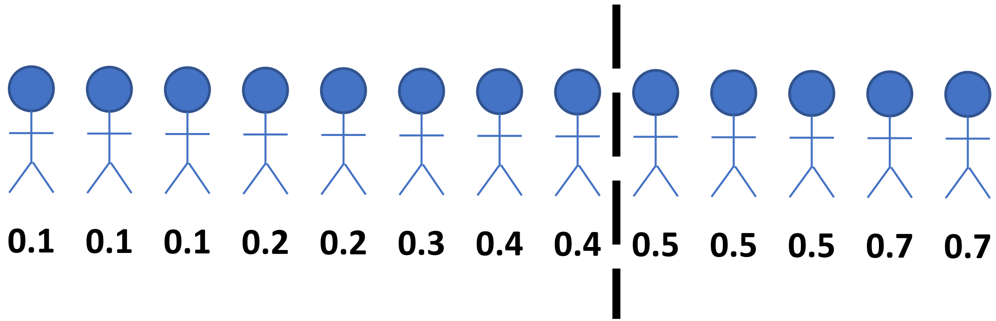
False positive rates



False positive rates

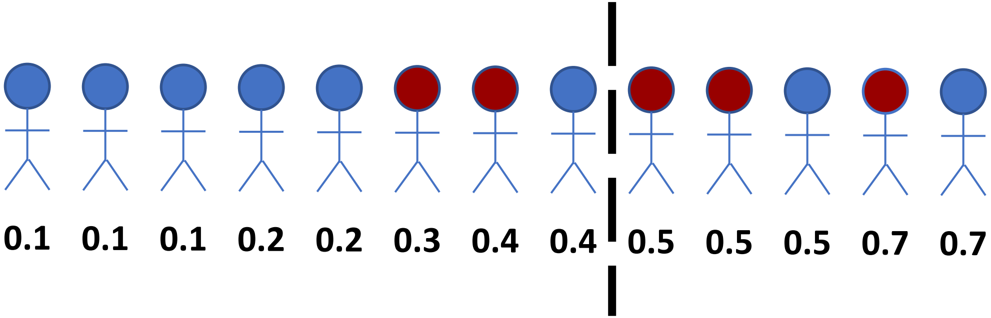


False positive rates



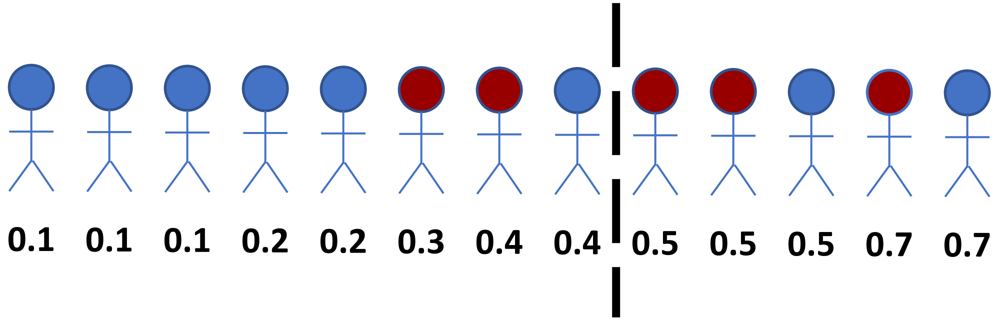
$$\frac{\text{Did not reoffend \& ``high risk''}}{\text{Did not reoffend}} =$$

False positive rates



$$\frac{\text{Did not reoffend \& "high risk"}}{\text{Did not reoffend}} =$$

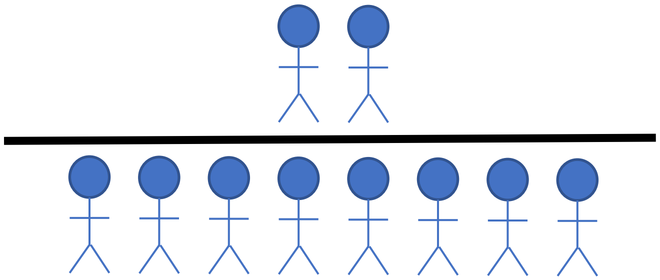
False positive rates



Did not reoffend & “high risk”

Did not reoffend

=

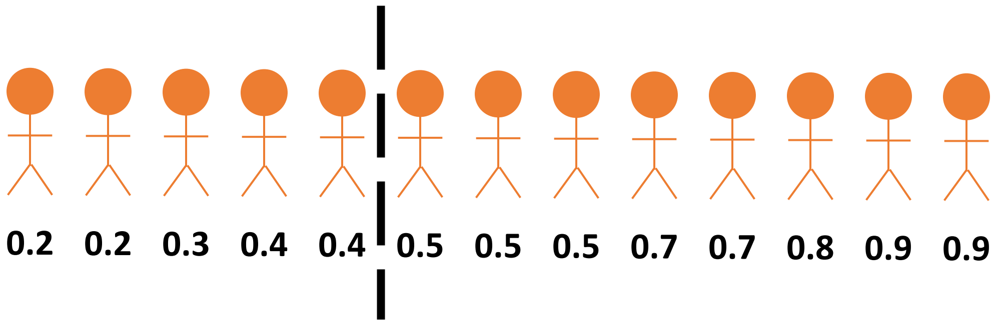


=

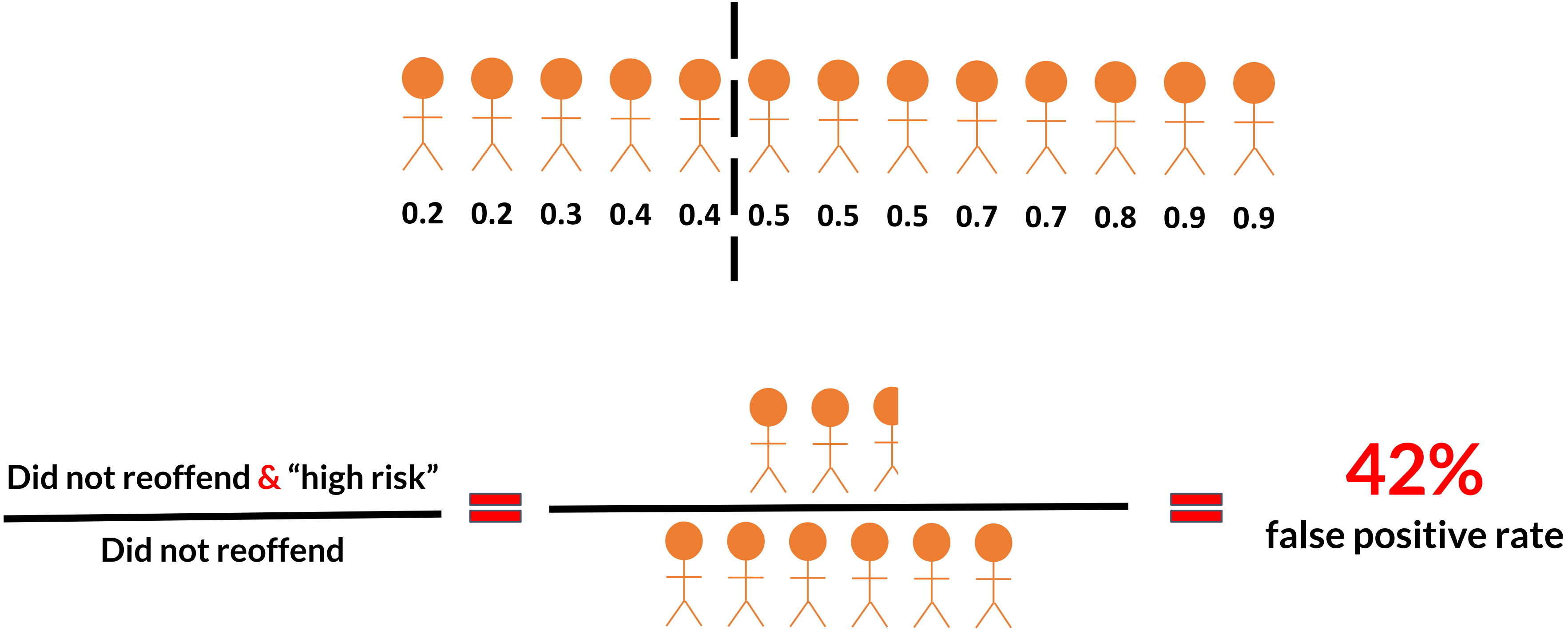
25%

false positive rate

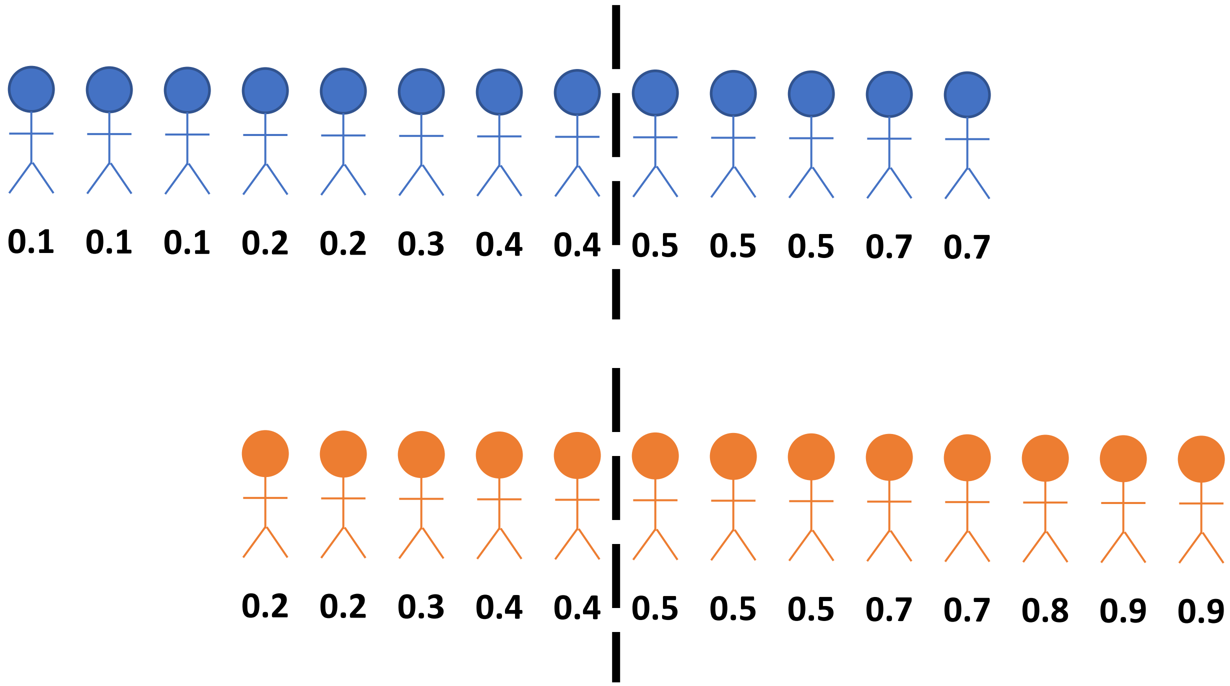
False positive rates



False positive rates



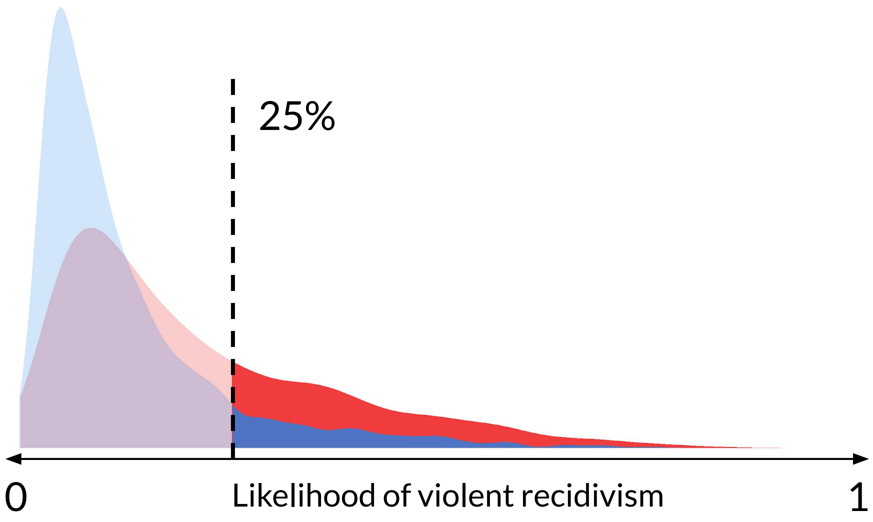
False positive rates



25%
false positive rate

42%
false positive rate

Broward County risk distributions

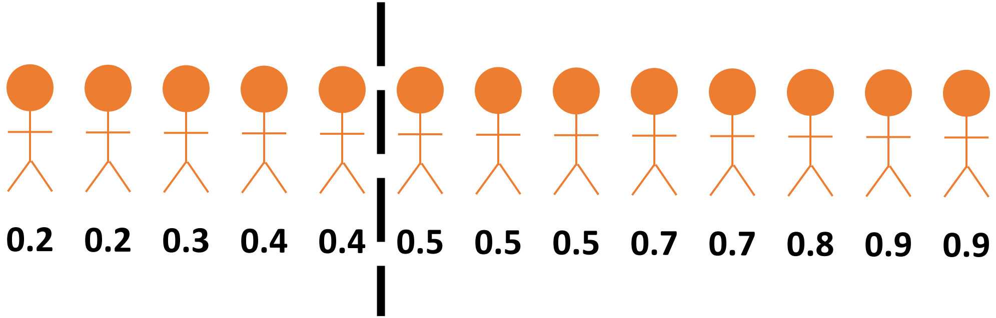


Black and white defendants have different risk distributions

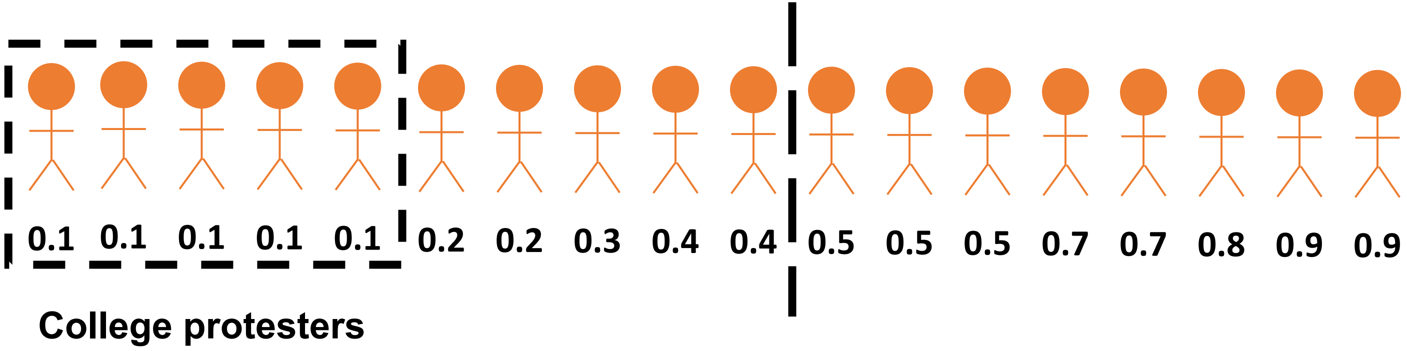
Infra-marginality

The false positive rate is an infra-marginal statistic—it depends not only on a group's threshold but on its distribution of risk.

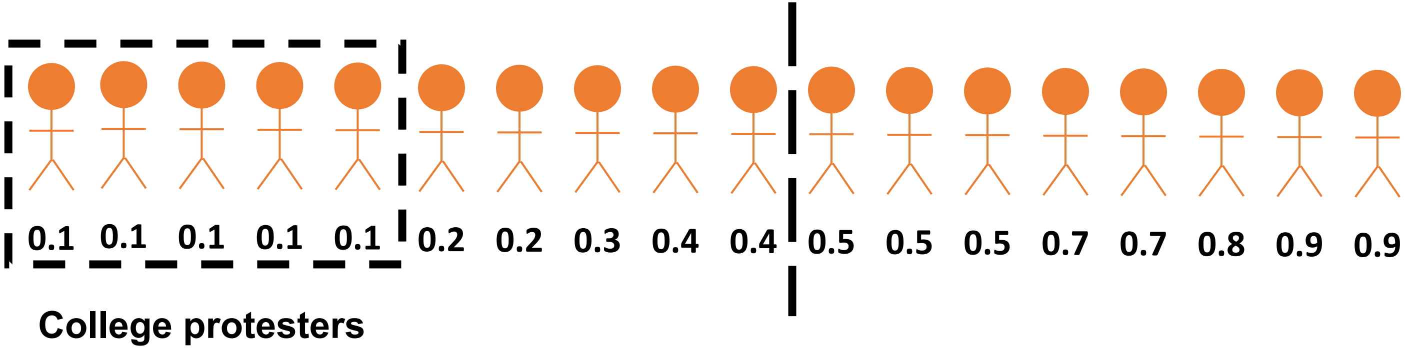
The problem with false positive rates



The problem with false positive rates



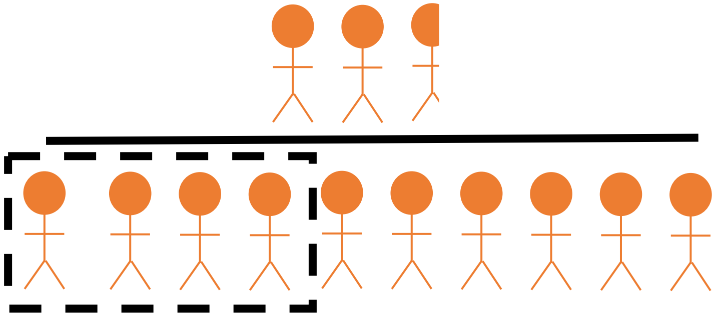
The problem with false positive rates



Did not reoffend & "high risk"

Did not reoffend

=

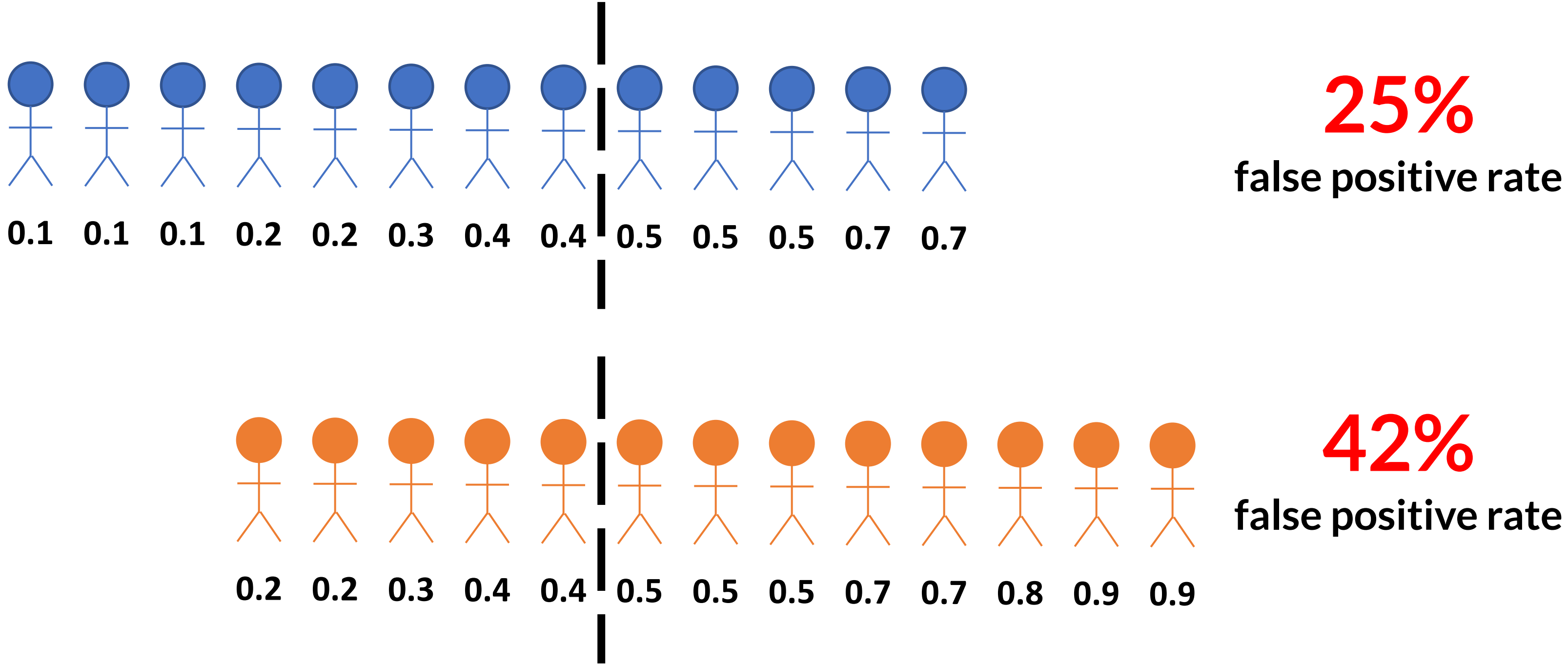


=

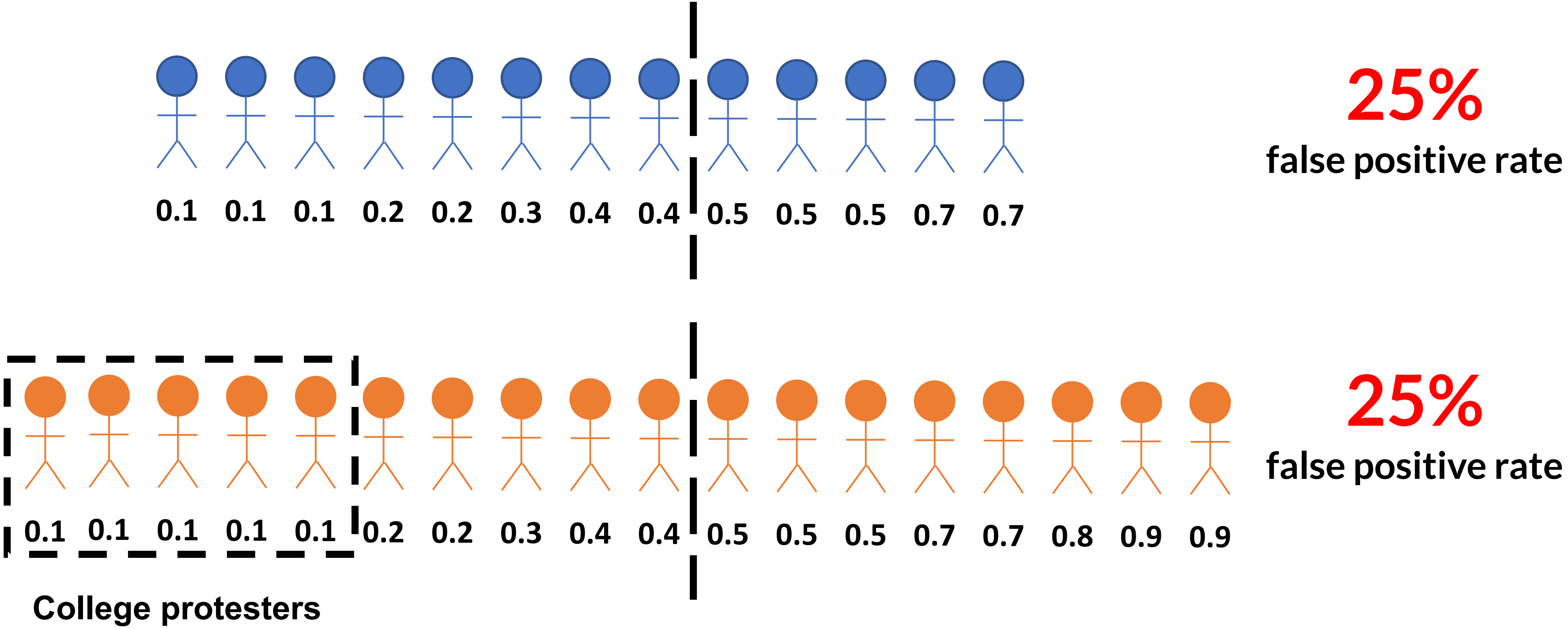
25%

false positive rate

The problem with false positive rates



The problem with false positive rates



Error rates in context

In traditional machine-learning settings, we compare multiple models on the **same dataset**.

Past *fair ML* work has often compared one model across **multiple datasets**, leading to hard-to-interpret results.

Error rates in context

In some settings, differences in error rates across groups can be a strong indicator of algorithmic problems and inequities.

In automated speech recognition — unlike for pretrial risk assessments — we have strong reason to believe that with more data and possibly better models, we should be able to obtain comparable error rates for Black and white speakers.

Part III: Consequentialist approach to fairness

Background

In many jurisdictions, people can be jailed for failing to appear at mandatory court dates.

As a result, it is possible to reduce incarceration by helping people appear in court.

One way to do this is to provide people with free door-to-door rideshare service to and from court.

Transportation assistance

Imagine we have enough money for 1,000 Lyft rides. Who should we give the rides to?

[We're preparing to give out rides starting this summer.]

Fairness in algorithms

Optimize appearance
1,000 new appearances
30% of one group gets rides, 10% of the other

Equal allocation across groups
800 new appearances
20% of each group gets rides

Which approach do you prefer?
Maybe somewhere in the middle?

Reference

Learning to be Fair: A Consequentialist Approach to Equitable Decision-Making

Alex Chohlas-Wood, Madison Coots, Emma Brunskill, Sharad Goel

Thanks to Alex Chohlas-Wood for help with slides!

The consequentialist approach

Traditional, deontological approaches do not consider the potential impacts of decisions on outcomes, and as a result, likely end in an allocation **misaligned** with **stakeholder preferences**.

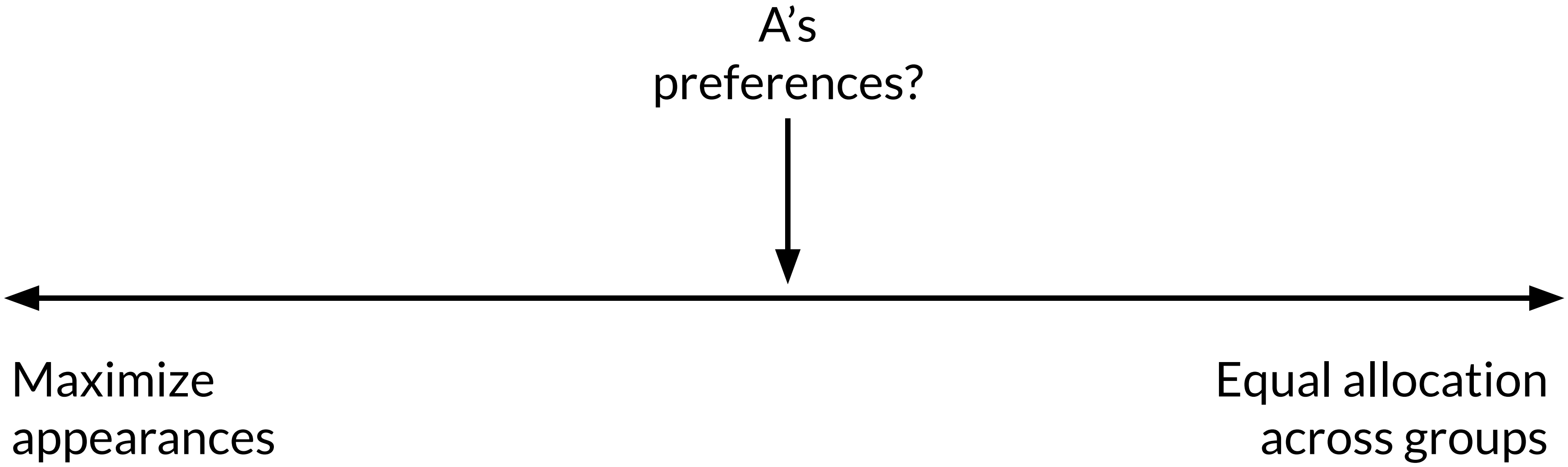
We take a different approach: we aim for **decisions** that **maximize stakeholder utility**, including one's preferences for parity.

Competing priorities

Maximize appearances

Equal allocation across groups

Competing priorities



Competing priorities

Competing priorities

The technical problem

In real-world settings, we want to **quickly learn and use** a policy that maximizes our **utility** subject to **budget constraints**.

But our utility depends on both **individual-level outcomes** (e.g., appearances) and **policy-level outcomes** (e.g., parity)**.**

Maximizing utility in practice

We use **multi-armed bandit algorithms**—including UCB and Thompson sampling—to make estimates of potential outcomes.

Maximizing utility in practice

We use **multi-armed bandit algorithms**—including UCB and Thompson sampling—to make estimates of potential outcomes.

We use a **linear program** to identify the optimal policy, according to these estimates, our utility, and our budget.

Algorithm outline

Step 1. Randomly treat a small warm-up population.

Algorithm outline

Step 1. Randomly treat a small warm-up population.

Step 2. Use the already-treated population to train a model that predicts outcomes for all available treatments.

Algorithm outline

Step 1. Randomly treat a small warm-up population.

Step 2. Use the already-treated population to train a model that predicts outcomes for all available treatments.

Step 3. Generate optimistic estimates for the potential outcomes under all actions.
[These two steps are the bandit.]

Algorithm outline

Step 1. Randomly treat a small warm-up population.

Step 2. Use the already-treated population to train a model that predicts outcomes for all available treatments.

Step 3. Generate optimistic estimates for the potential outcomes under all actions.
[These two steps are the bandit.]

Step 4. Using these estimates, solve for the policy that maximizes utility.
[This is the linear program.]

Algorithm outline

Step 1. Randomly treat a small warm-up population.

Step 2. Use the already-treated population to train a model that predicts outcomes for all available treatments.

Step 3. Generate optimistic estimates for the potential outcomes under all actions.
[These two steps are the bandit.]

Step 4. Using these estimates, solve for the policy that maximizes utility.
[This is the linear program.]

Step 5. Act according to this policy.

Algorithm outline

Step 1. Randomly treat a small warm-up population.

Step 2. Use the already-treated population to train a model that predicts outcomes for all available treatments.

Step 3. Generate optimistic estimates for the potential outcomes under all actions.
[These two steps are the bandit.]

Step 4. Using these estimates, solve for the policy that maximizes utility.
[This is the linear program.]

Step 5. Act according to this policy.

Repeat from Step 2.

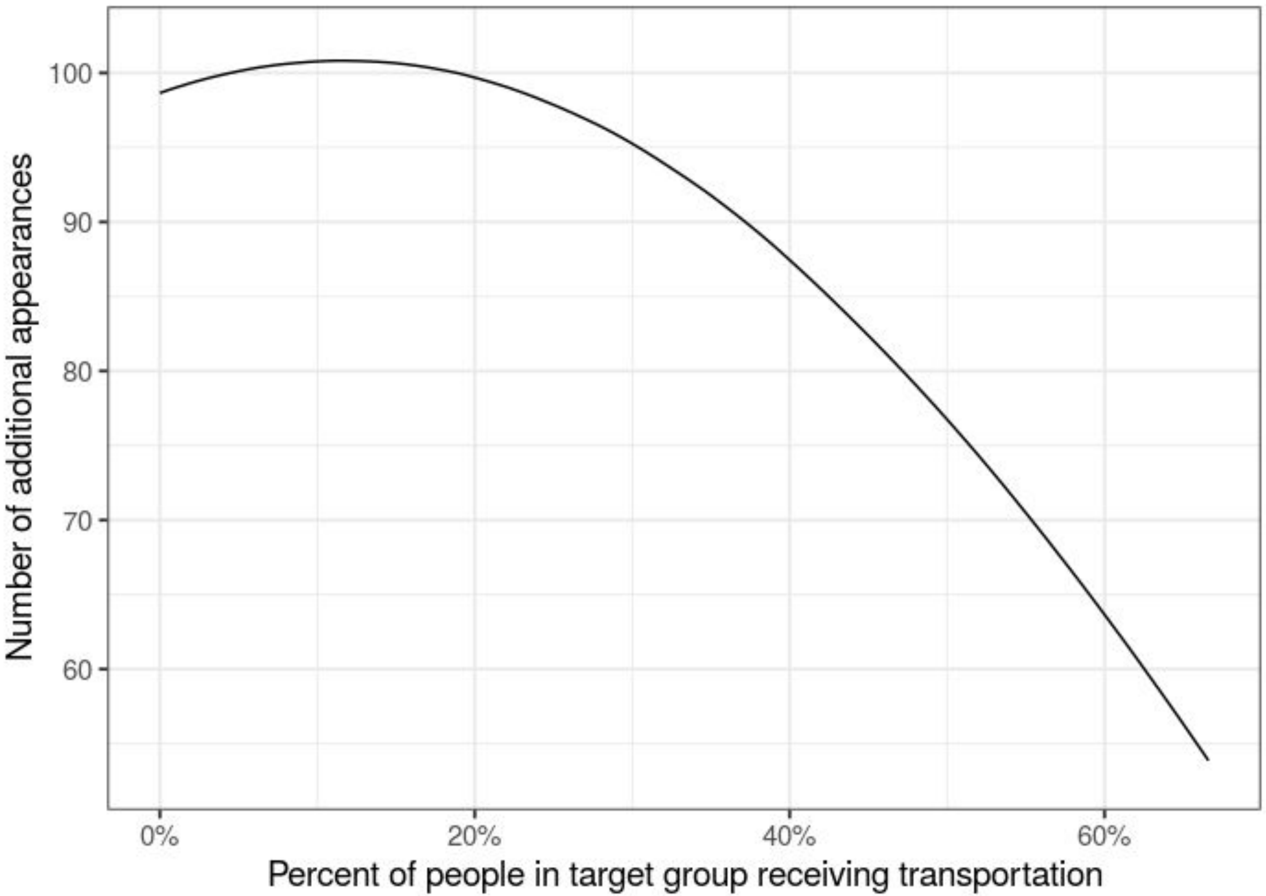
Traditional fairness approaches vs. our approach

What happens when we force our policy to satisfy a particular mathematical fairness constraint, rather than directly deciding which outcomes we prefer?

For example, what happens when we insist on demographic parity or classification parity? Satisfying these mathematical constraints will result in *sub-optimal* outcomes.

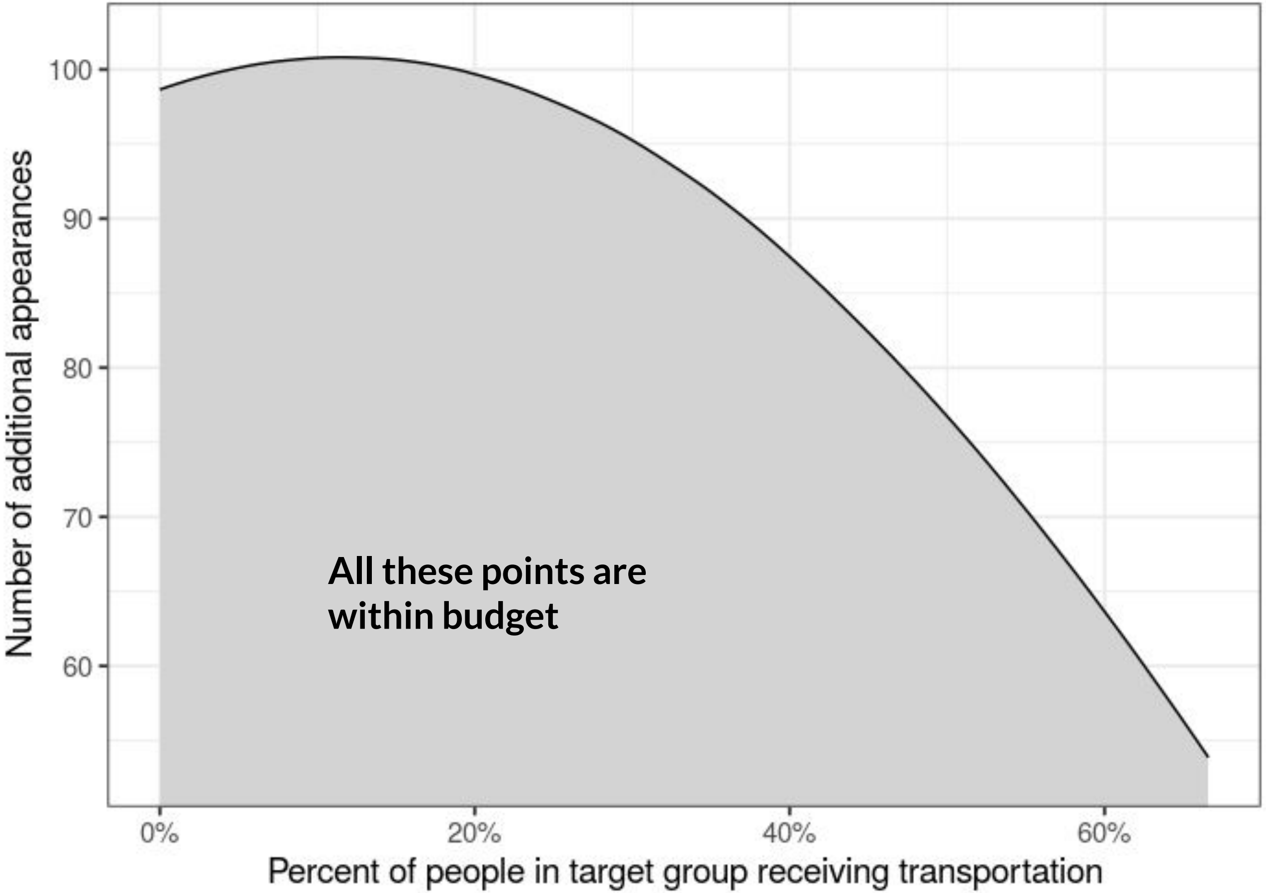
Principled trade-offs:

Different outcomes on the Pareto frontier



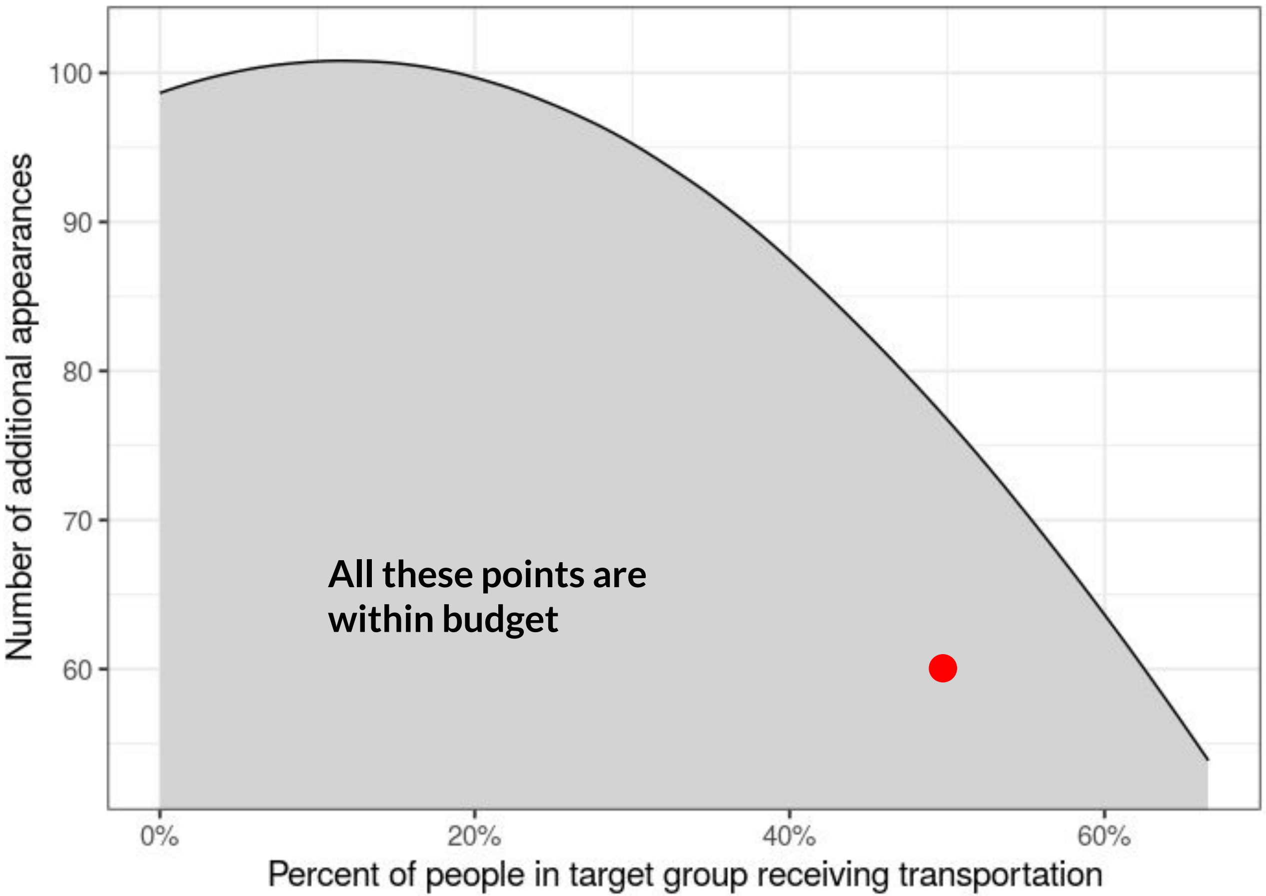
Principled trade-offs:

Different outcomes on the Pareto frontier



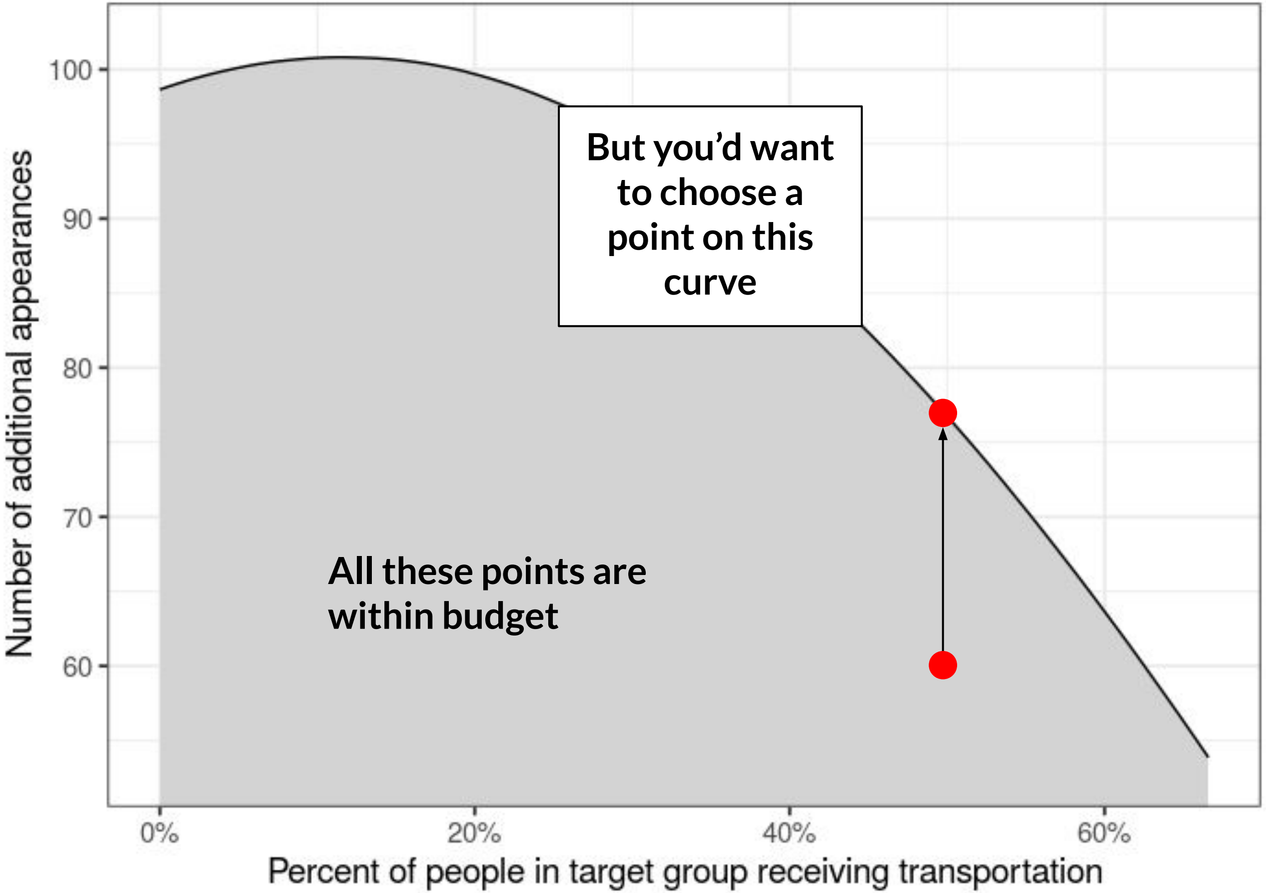
Principled trade-offs:

Different outcomes on the Pareto frontier



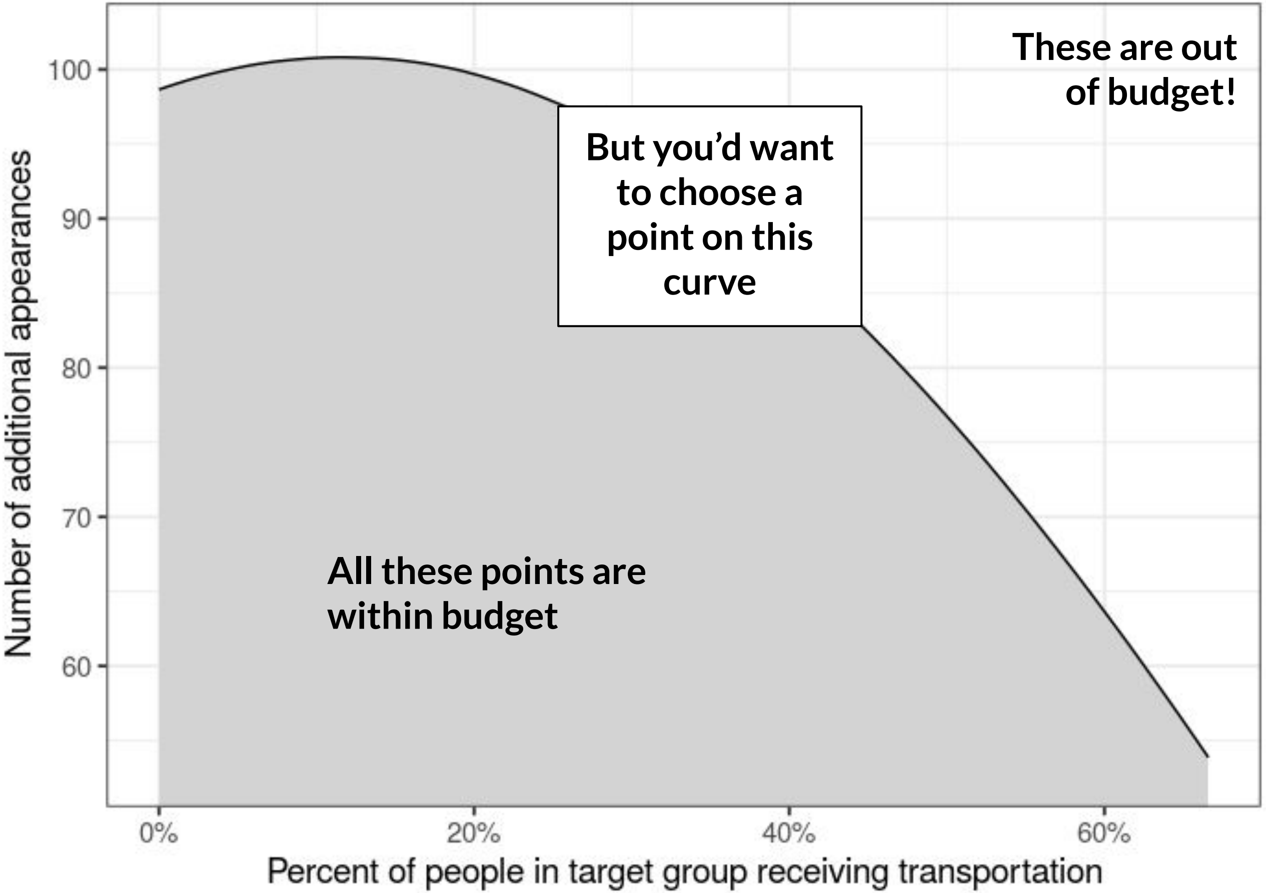
Principled trade-offs:

Different outcomes on the Pareto frontier



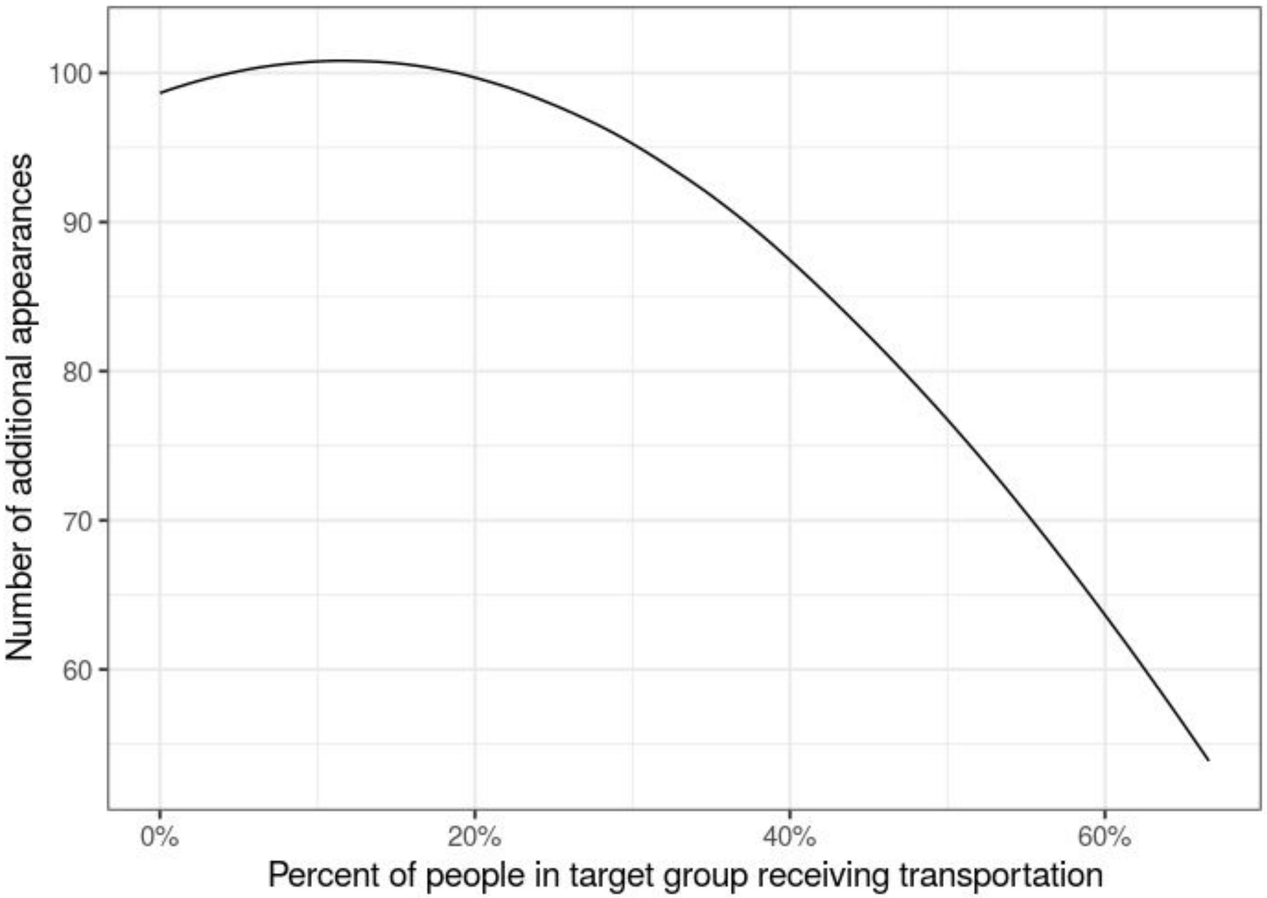
Principled trade-offs:

Different outcomes on the Pareto frontier



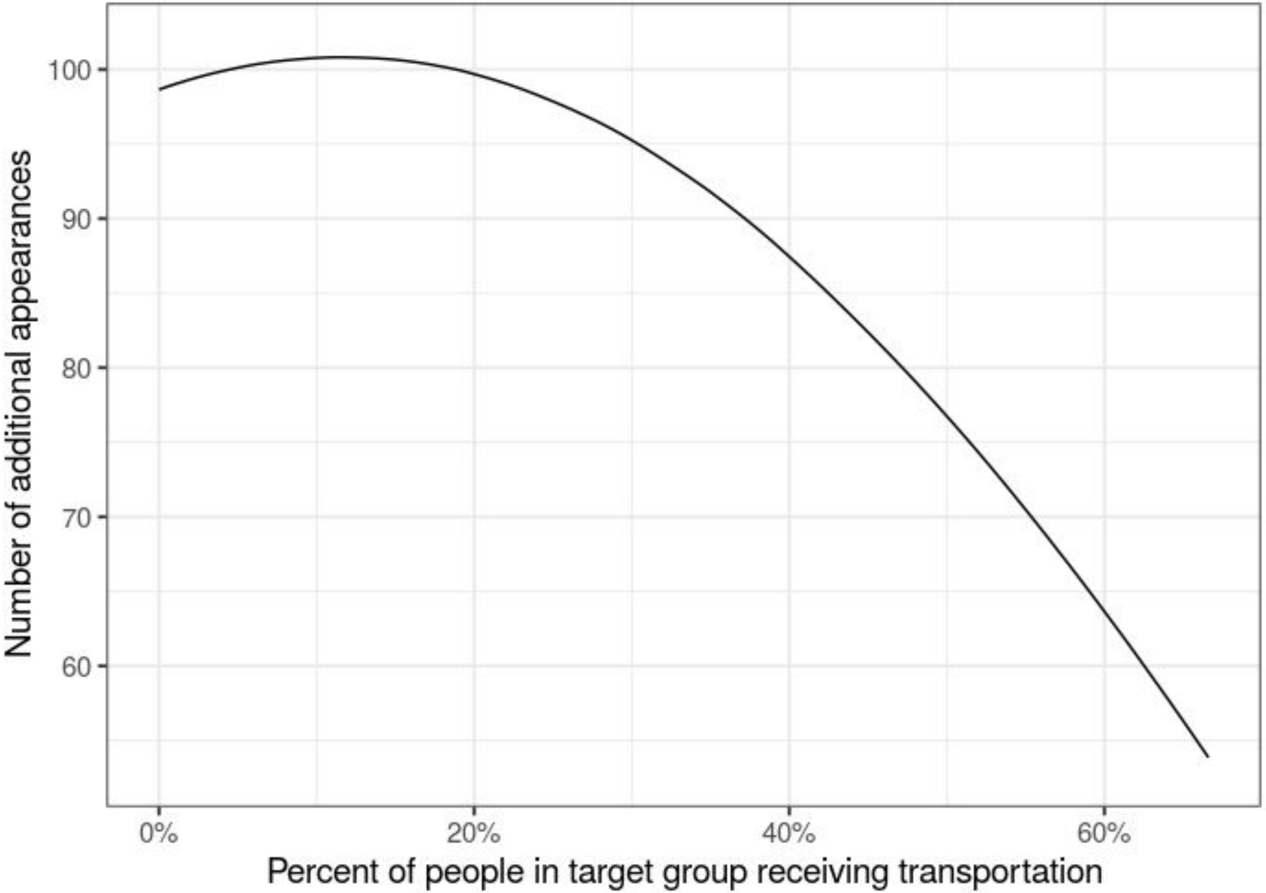
Principled trade-offs:

Different outcomes on the Pareto frontier



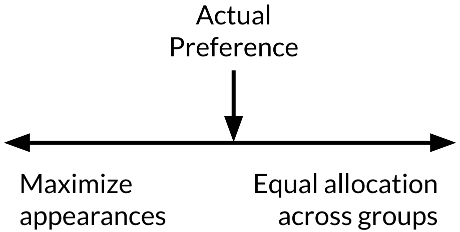
Principled trade-offs:

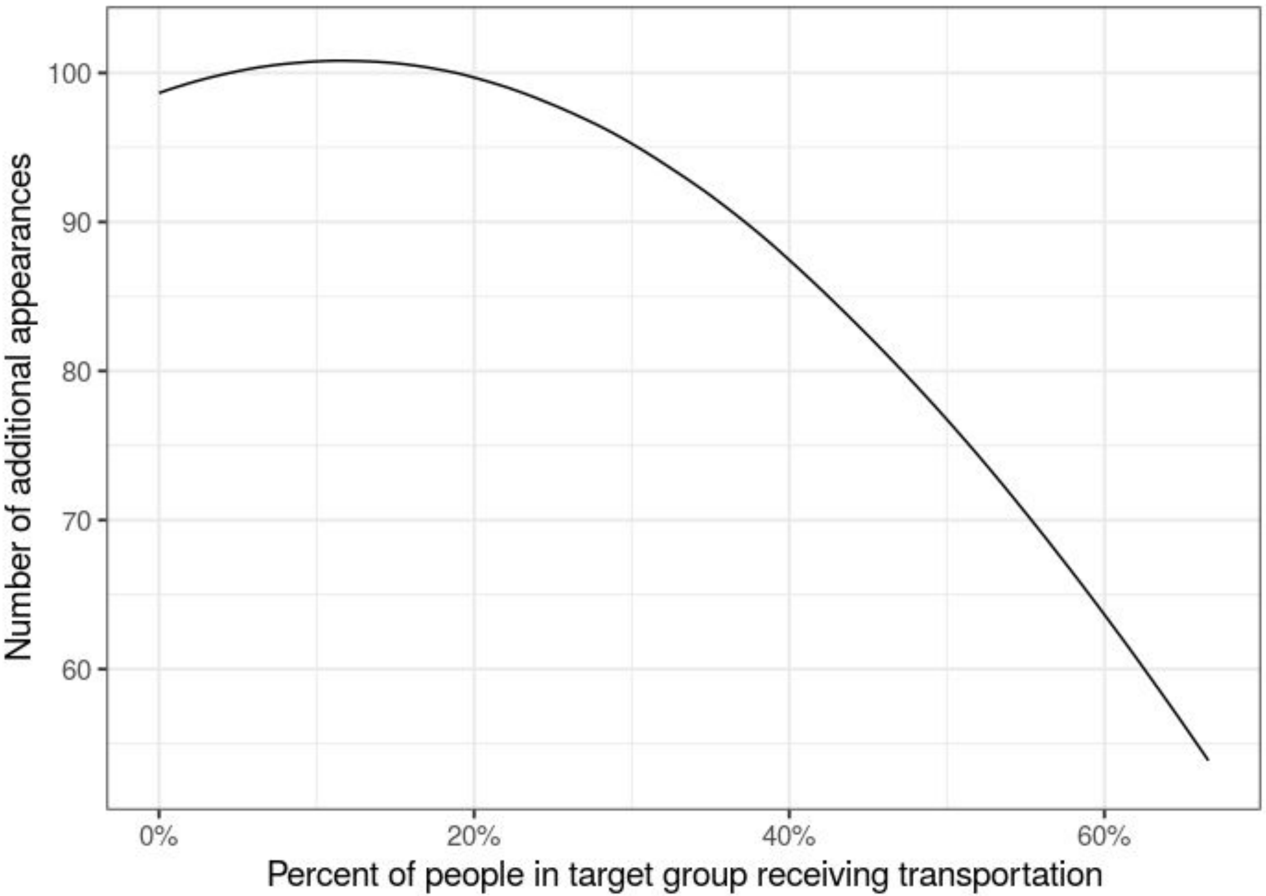
Different outcomes on the Pareto frontier



Principled trade-offs:

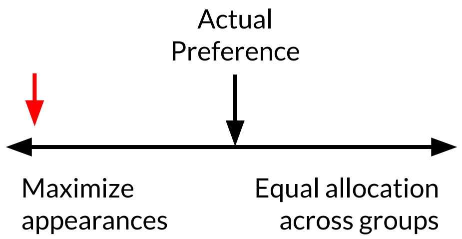
Different outcomes on the Pareto frontier

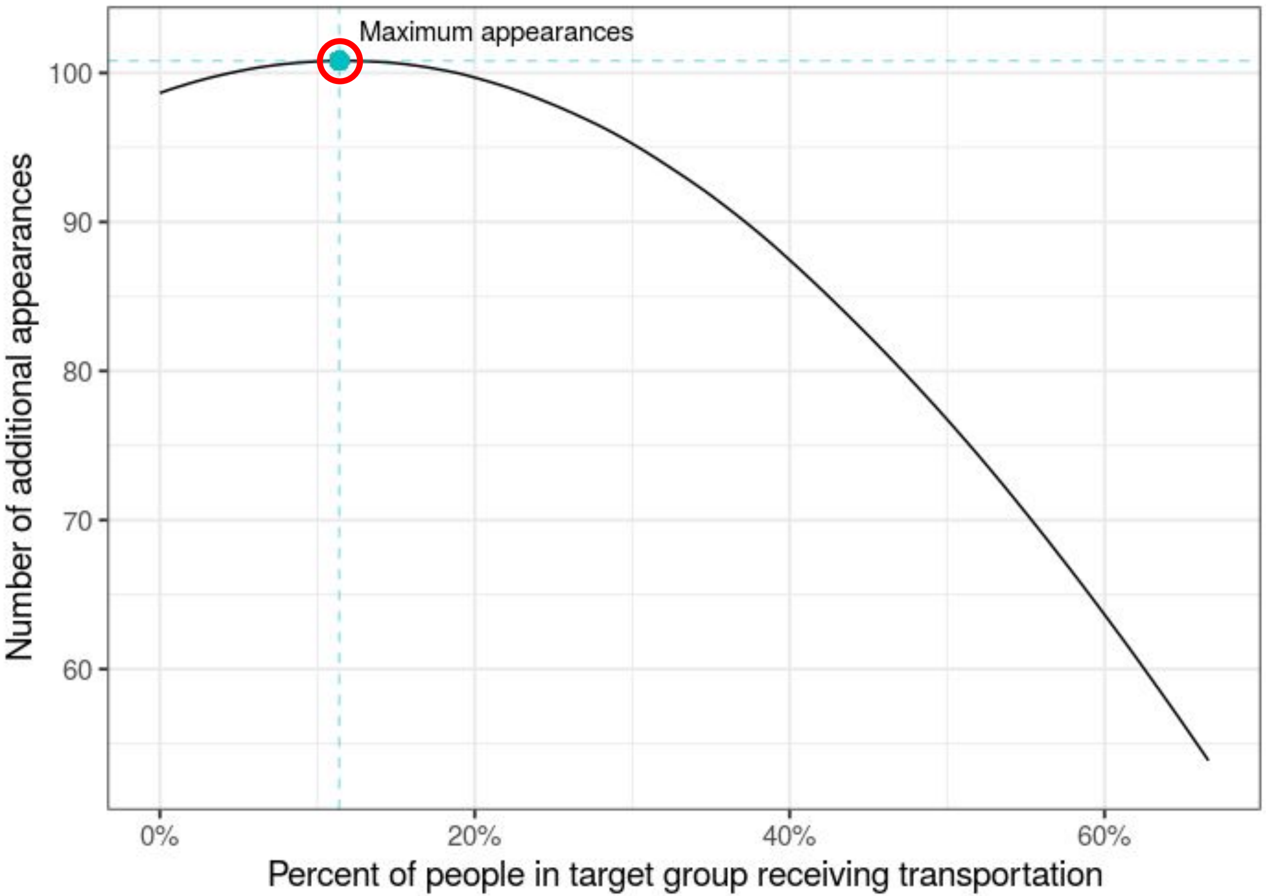




Principled trade-offs:

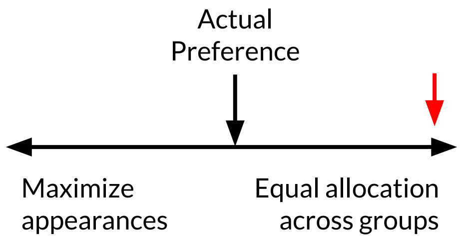
Different outcomes on the Pareto frontier

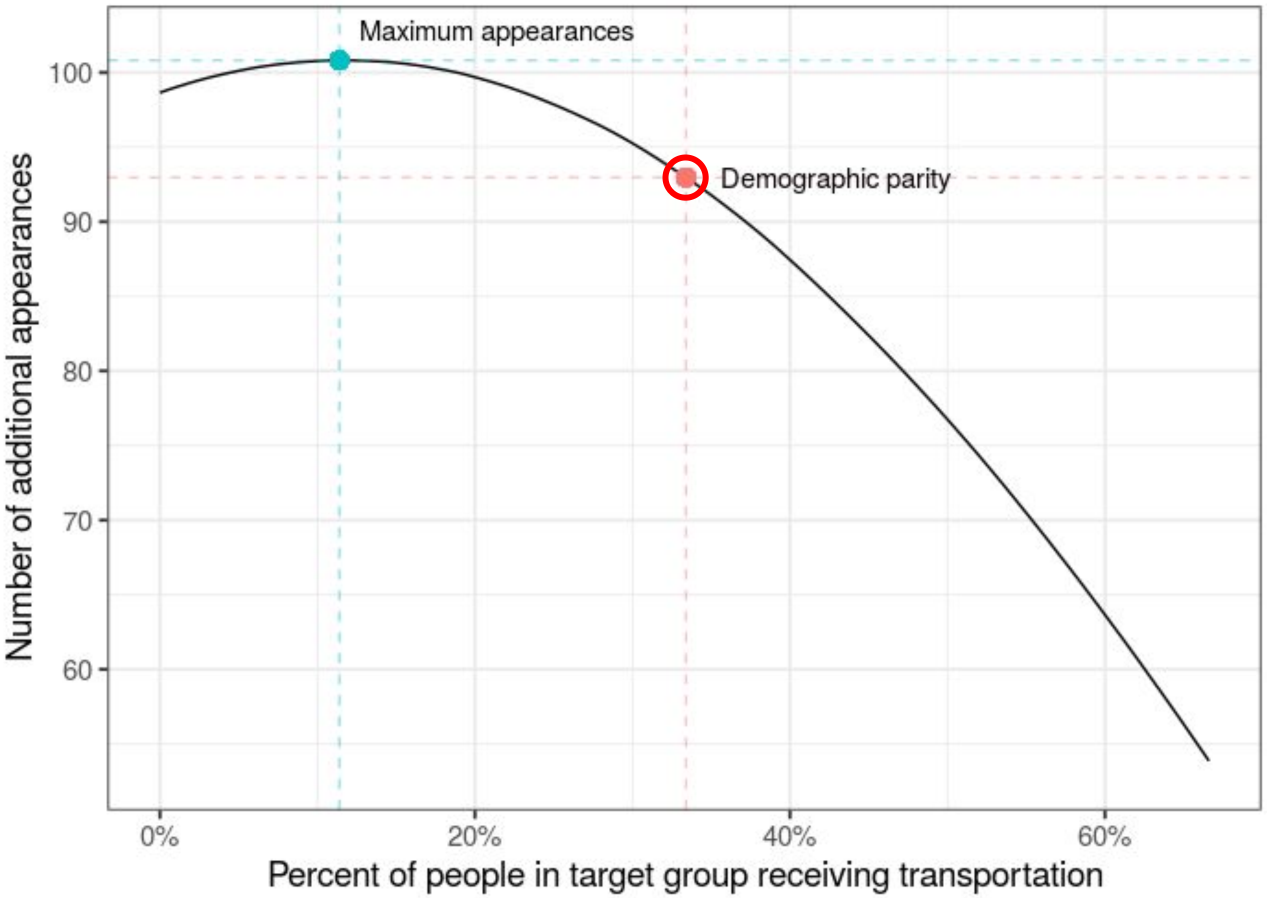




Principled trade-offs:

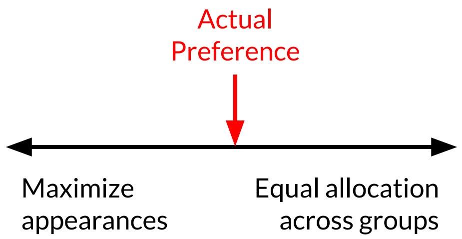
Different outcomes on the Pareto frontier

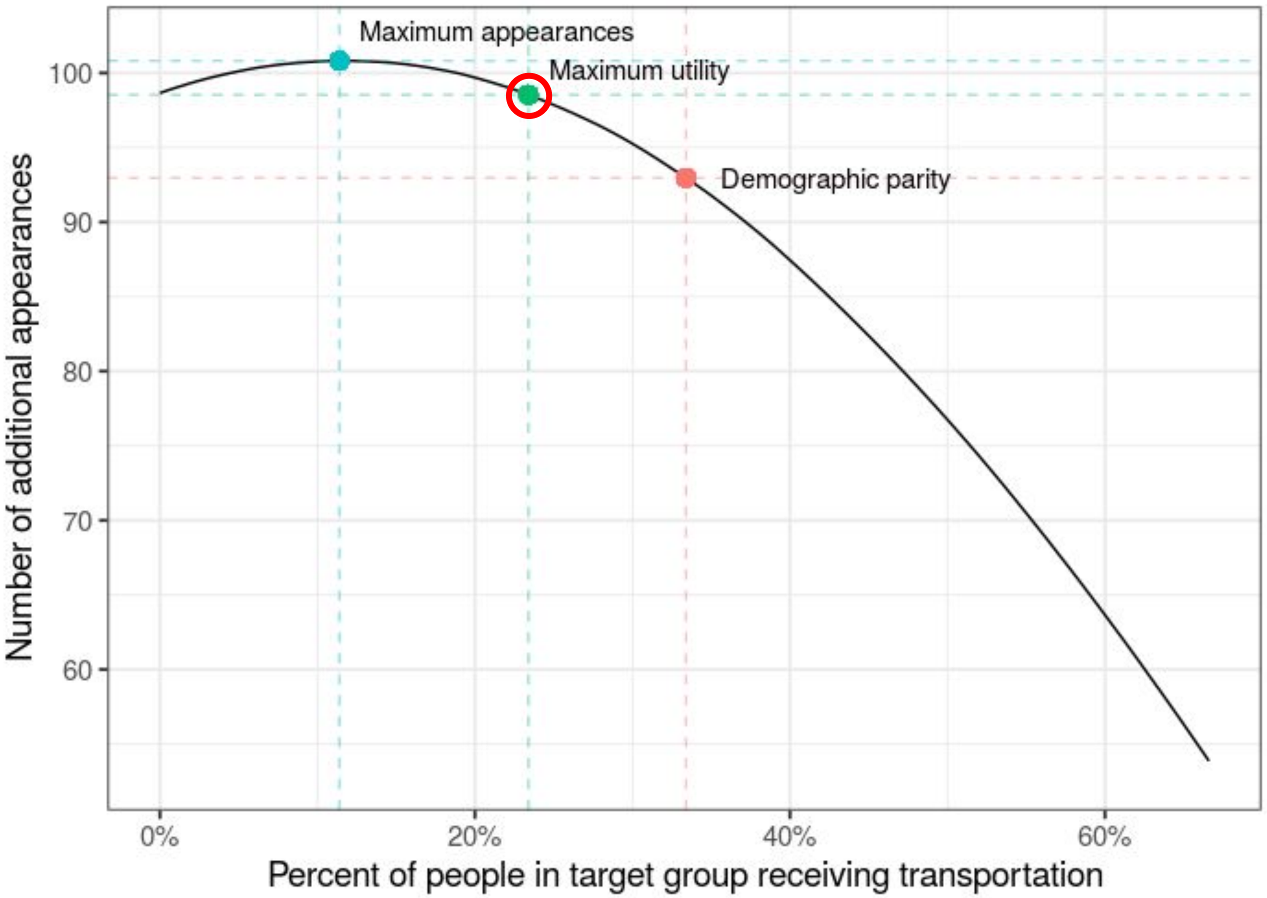




Principled trade-offs:

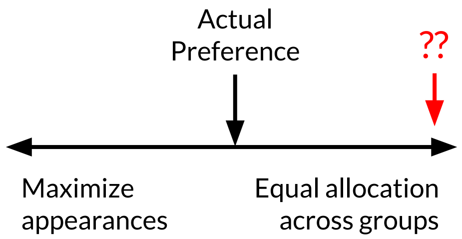
Different outcomes on the Pareto frontier

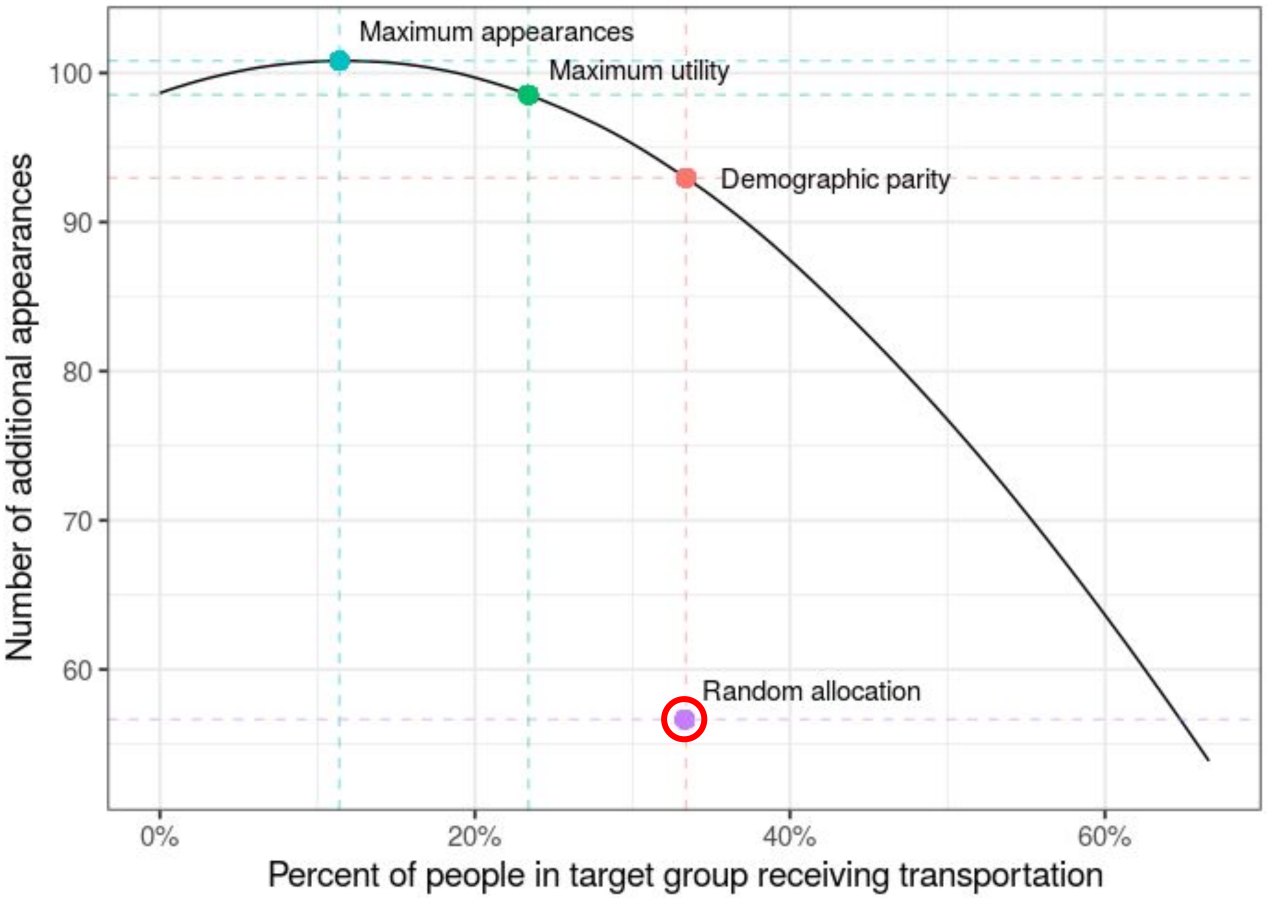




Principled trade-offs:

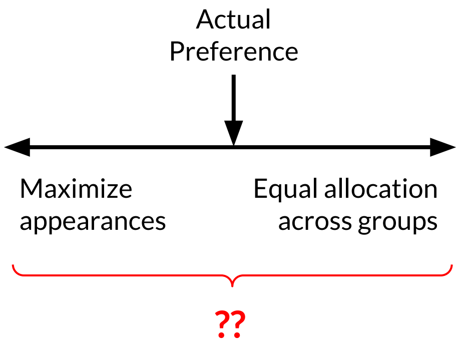
Different outcomes on the Pareto frontier

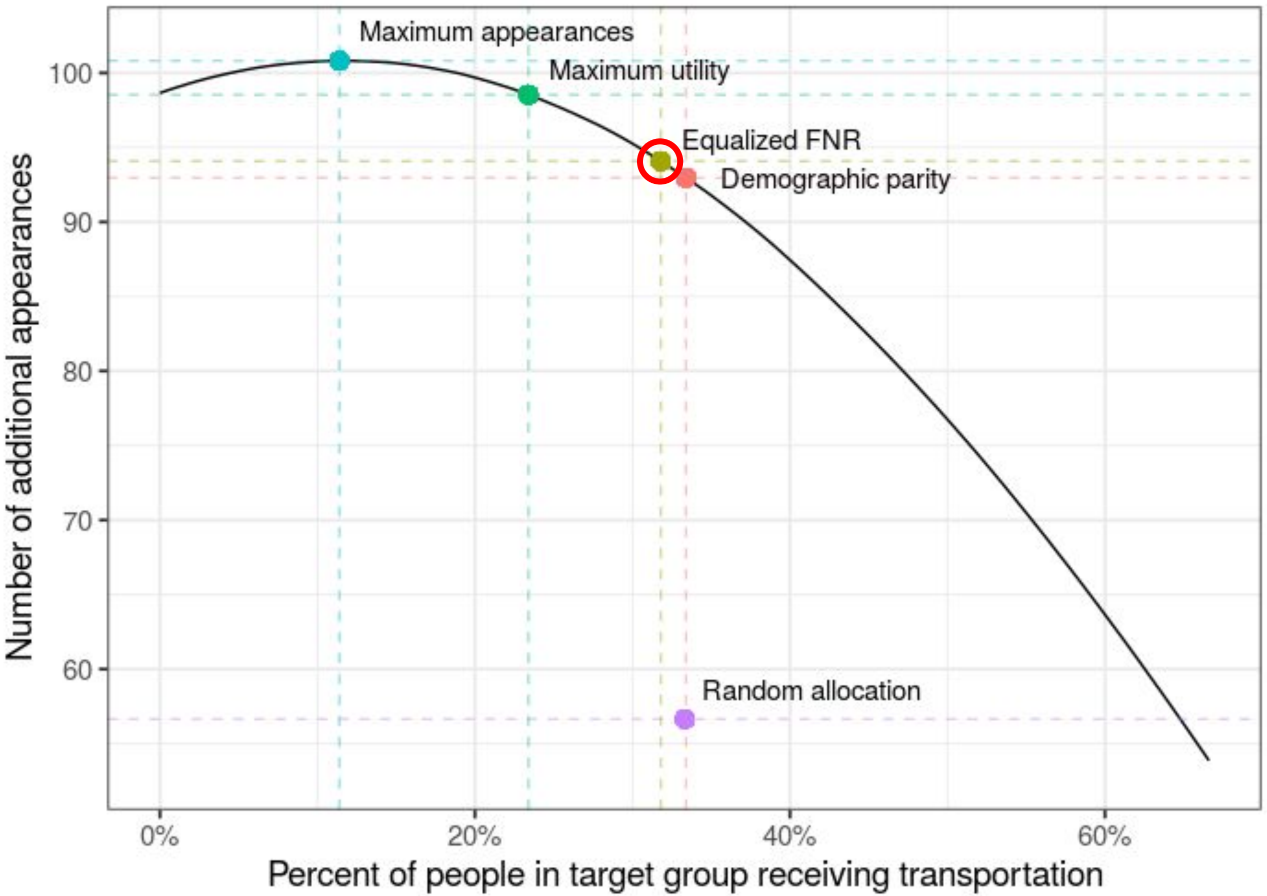




Principled trade-offs:

Different outcomes on the Pareto frontier





Stanford Computational Policy Lab

policylab.stanford.edu