Paradoxes in Fair Machine Learning

Paul Gölz, Anson Kahng, and Ariel Procaccia

NeurIPS 2019

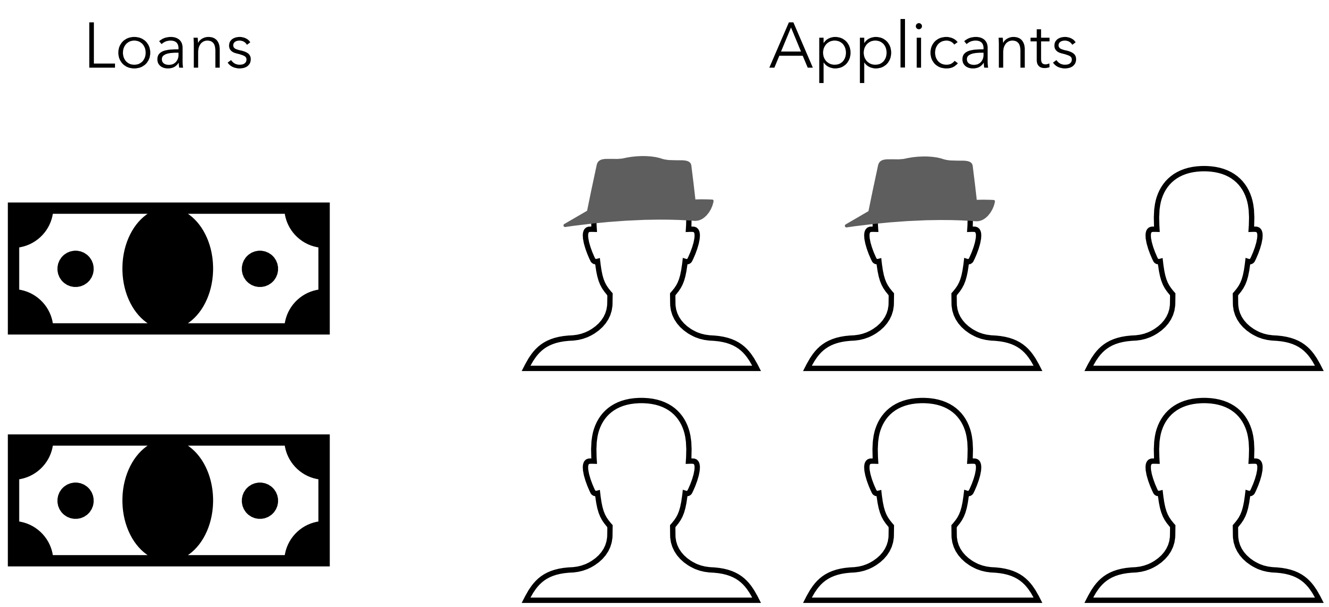
RESEARCH QUESTION

"Fairness in machine learning vs. fairness in fair division"

What is the relationship between **statistical notions of fairness** (in particular, equalized odds) and **axioms of fair division** (in particular, the axioms of resource monotonicity, population monotonicity, and consistency)?

CLASSIFICATION WITH CARDINALITY CONSTRAINTS

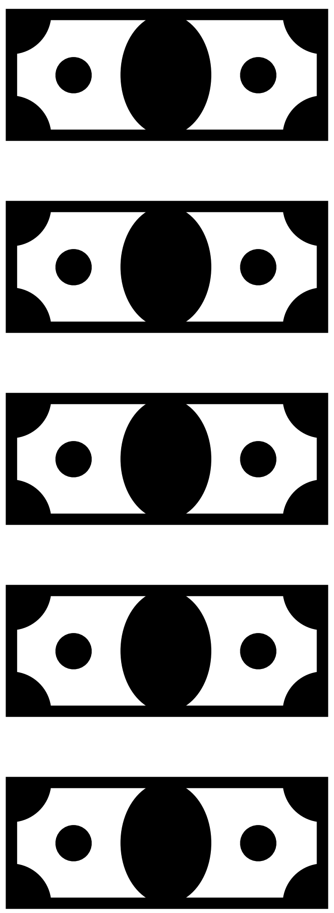
Classification problem with a fixed budget of available resources to distribute



Goal: train a classifier to maximize **efficiency** (fraction of loans that will be repaid)

Classification with Cardinality Constraints

Loans

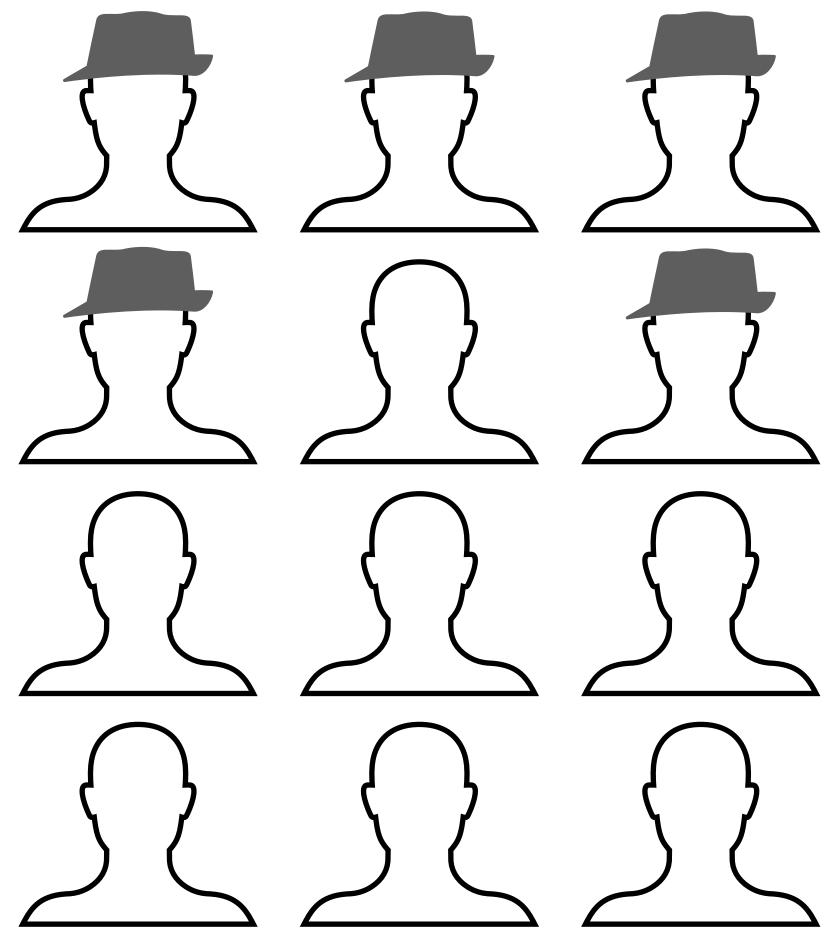


Two groups:
hats vs. no hats

Goal: Distribute loans to applicants in order to minimize the default rate.

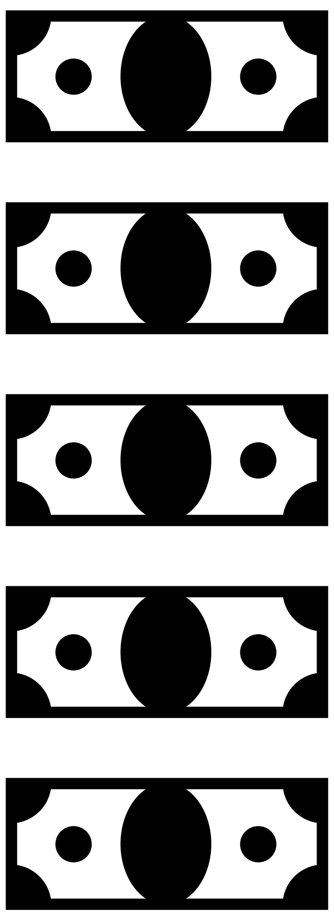
Metric: efficiency

Applicants



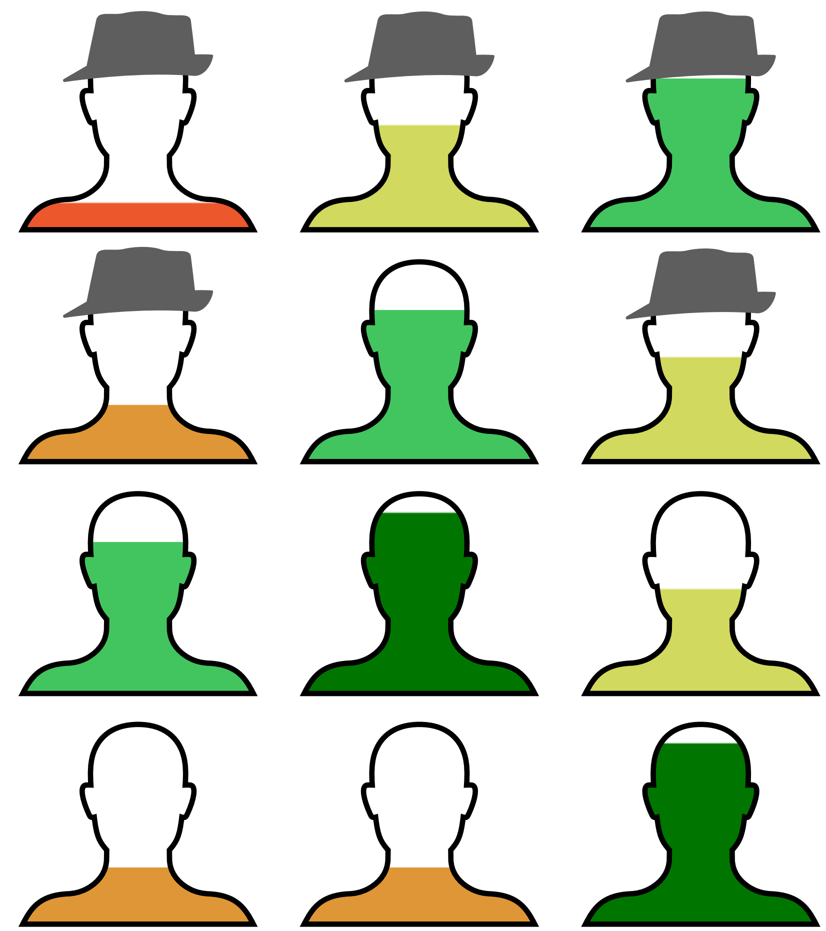
Classification with Cardinality Constraints

Loans

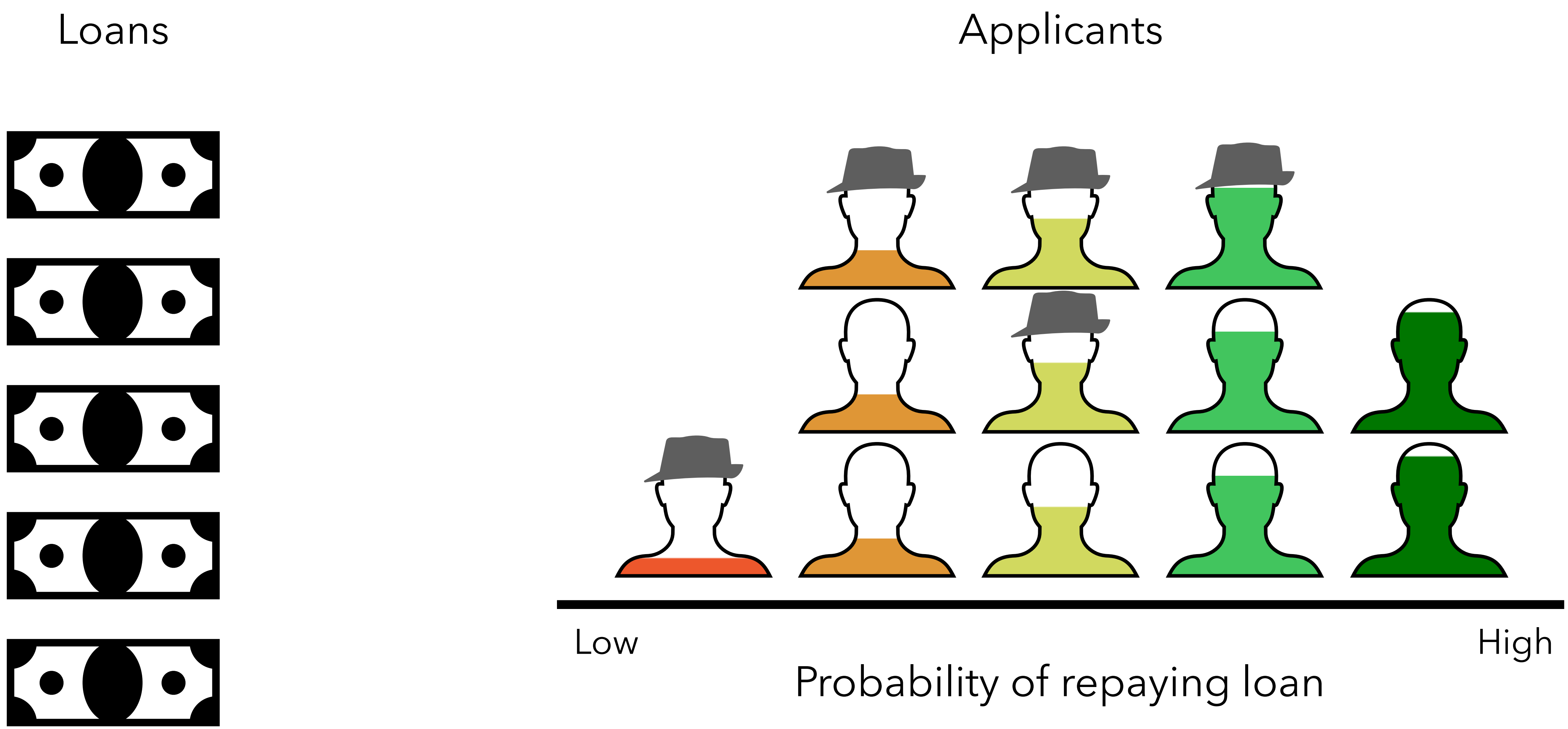


Calibrated classifier: If the classifier labels a set of people with probability p, then a p fraction of them are positive instances.

Applicants



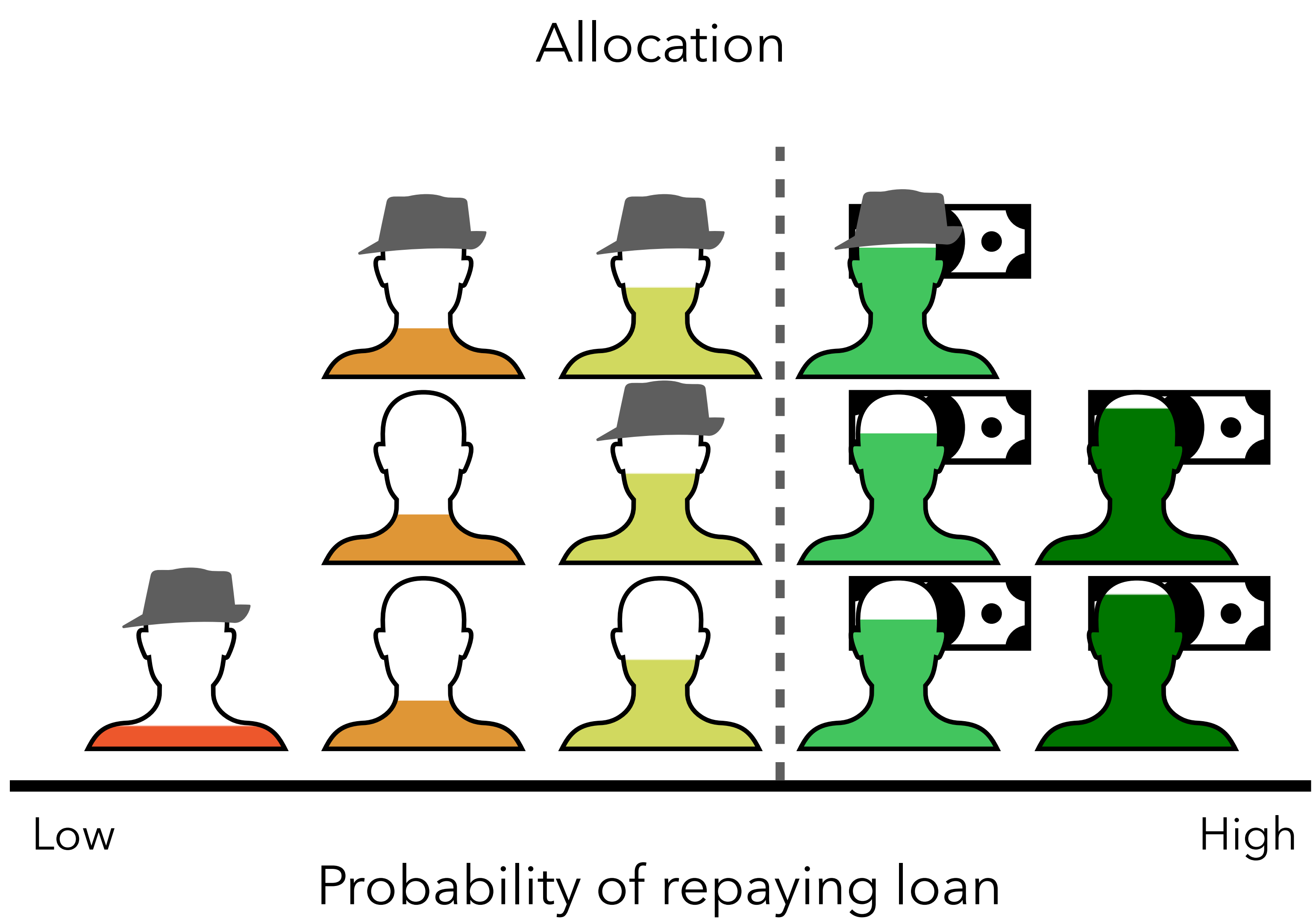
Classification with Cardinality Constraints



Classification with Cardinality Constraints

In this setting, the optimal allocation rule awards loans to the most qualified applicants.

But what about **fairness** between groups?



Fairness Concepts

Statistical Fairness	Fair Division Axioms
Equalized odds	Resource monotonicity
Demographic parity	Population monotonicity
	Consistency

How compatible are these notions of fairness?
How much does efficiency suffer if we have to satisfy both equalized odds and various fair division axioms?

Fairness Concepts

Statistical Fairness	Fair Division Axioms
$\boxed{\text{Equalized odds}}$	Resource monotonicity
Demographic parity	Population monotonicity
	Consistency

<u>Research question (rephrased)</u>:

How much does efficiency suffer if we must satisfy both equalized odds and various fair division axioms?

Statistical Fairness

Equalized Odds (EO):

"A predictor $\hat{Y}$ satisfies equalized odds with respect to a protected attribute A and outcome Y if $\hat{Y}$ and A are independent conditional on Y" (Hardt et al. 2016)

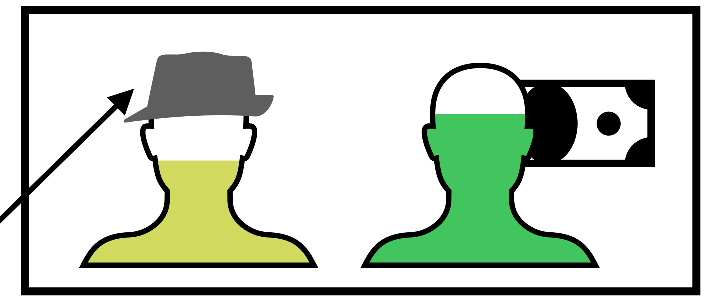
$$\Pr(\hat{Y} = 1 \,|\, A = 1, Y = 1) = \Pr(\hat{Y} = 1 \,|\, A = 0, Y = 1)$$

$$\Pr(\hat{Y} = 1 \,|\, A = 1, Y = 0) = \Pr(\hat{Y} = 1 \,|\, A = 0, Y = 0)$$

"True positive and false positive rates are equal across groups"

STATISTICAL FAIRNESS

Equalized Odds (EO):



"A predictor $\hat{Y}$ satisfies equalized odds with respect to a protected attribute A and outcome Y if $\hat{Y}$ and A are independent conditional on Y" (Hardt et al. 2016)

$$\Pr(\hat{Y} = 1 \,|\, A = 1, Y = 1) = \Pr(\hat{Y} = 1 \,|\, A = 0, Y = 1)$$

$$\Pr(\hat{Y} = 1 \,|\, A = 1, Y = 0) = \Pr(\hat{Y} = 1 \,|\, A = 0, Y = 0)$$

"True positive and false positive rates are equal across groups"

Fair Division Axioms

Resource monotonicity:

"Adding more resources makes everyone better off"

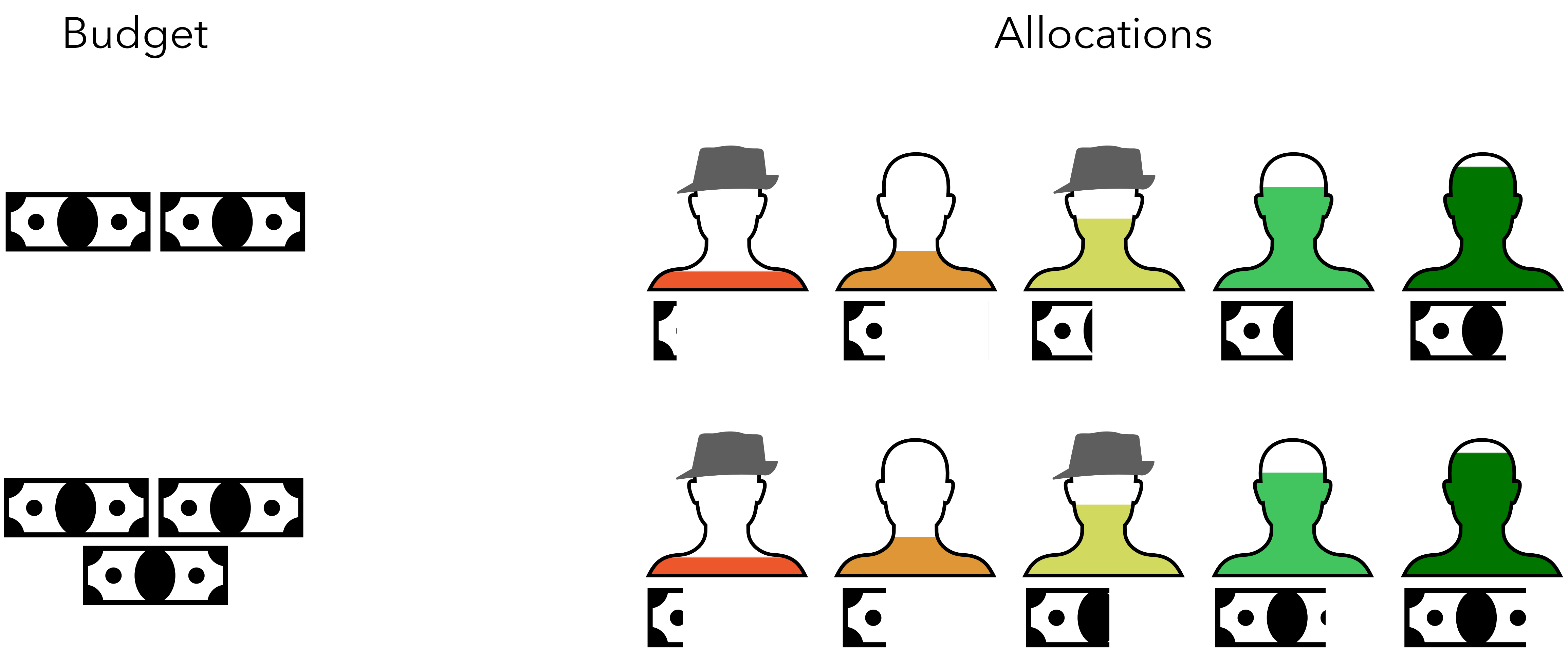
Population monotonicity:

"Adding more people makes everyone worse off"

Think of these axioms as preclusions of paradoxes.

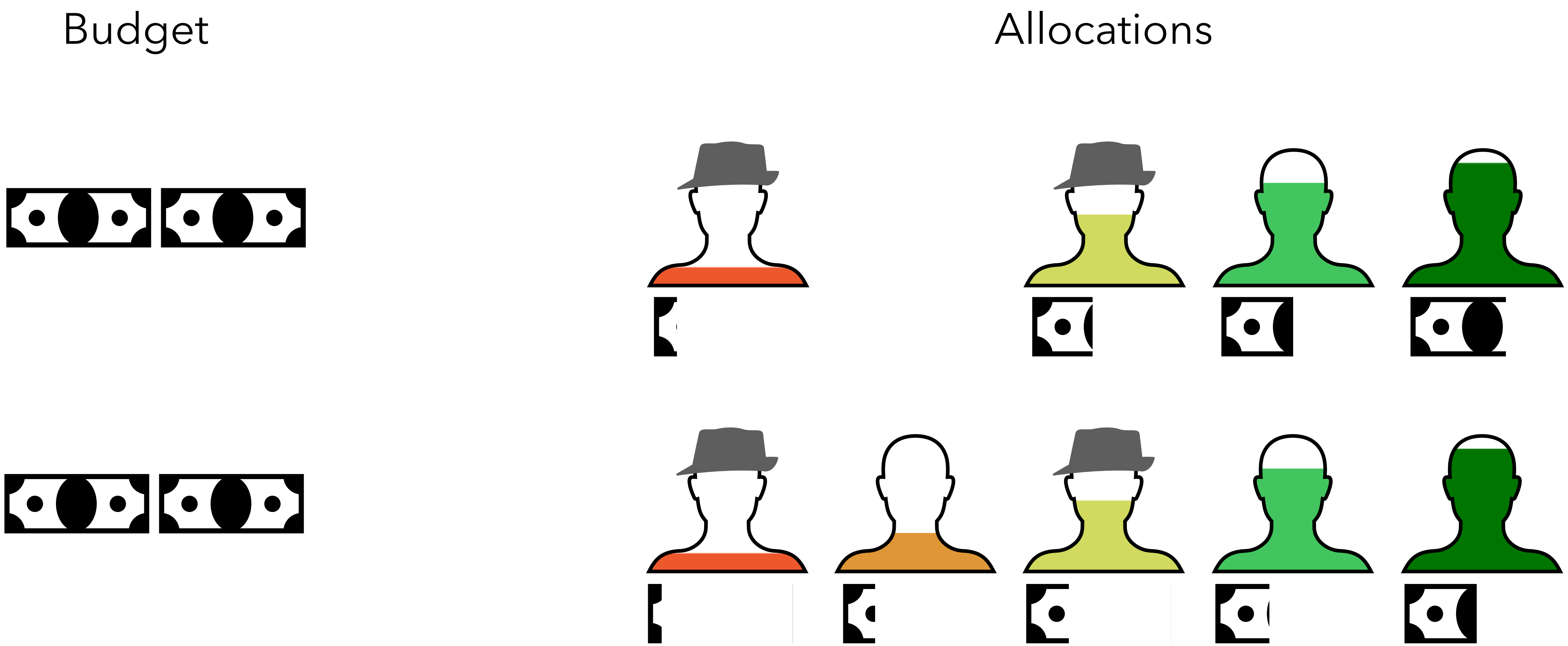
RESOURCE MONOTONICITY

"Adding more resources makes everyone weakly better off"



Population Monotonicity

"Adding more people makes everyone weakly worse off"



RESULTS

1. In the cardinality-constrained model, we characterize the optimal allocation rule that satisfies equalized odds

2. Equalized odds and **resource monotonicity** are achievable with no loss to optimal EO efficiency

3. Any rule that satisfies equalized odds and **population monotonicity** cannot achieve a constant-factor approximation to optimal EO efficiency

4. The only rule that satisfies equalized odds and **consistency** is uniformly random allocation