Algorithmic Fairness from an Economics POV

Annie Liang
(Northwestern)

prediction problems

many of you are familiar with prediction problems in machine learning

- there is an observable feature vector $x \in X$
- there is an outcome $y \in Y$ of interest

the goal is to predict the unknown y given the observed x

classic ML problems

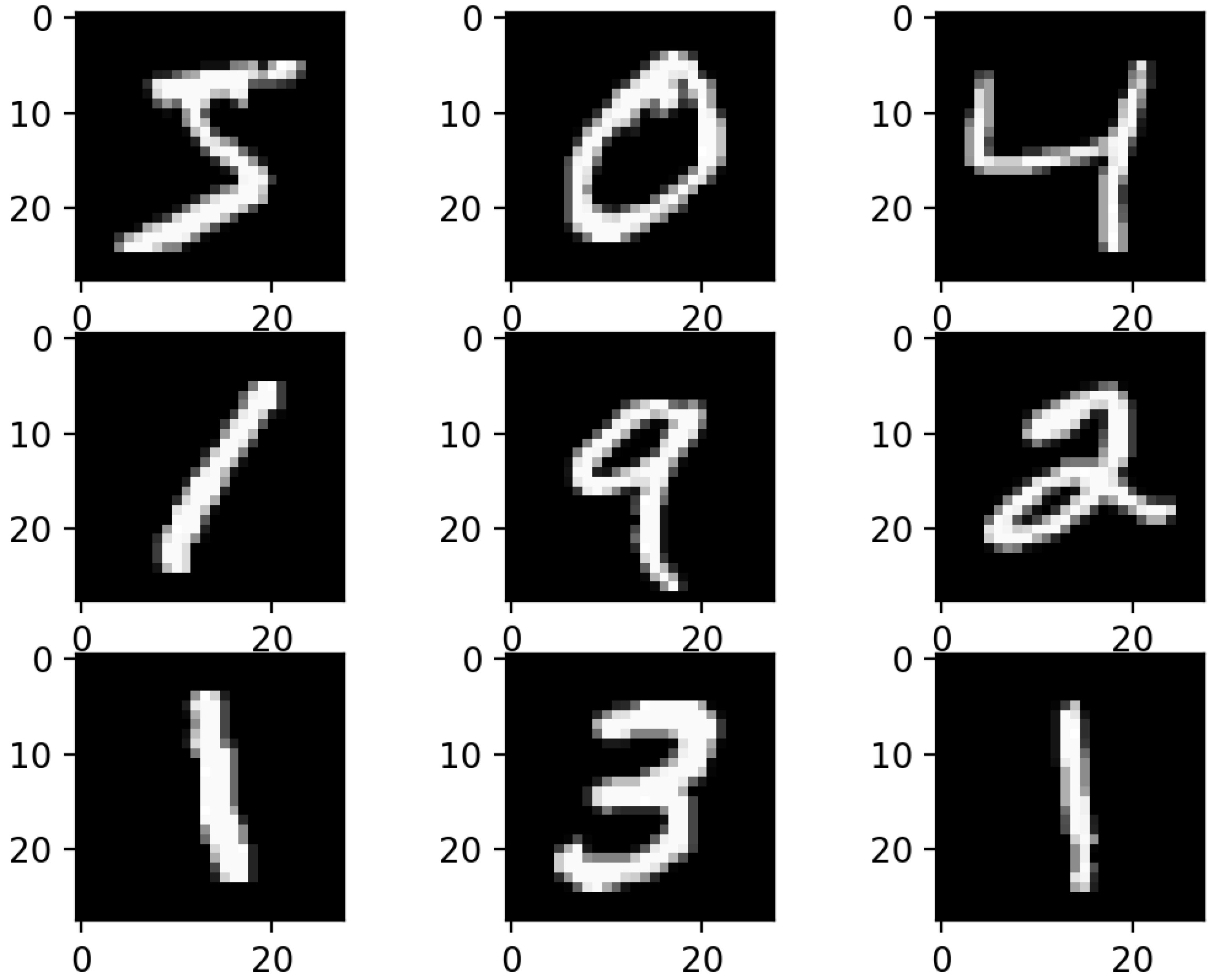


figure: predict the digits these images represent

classic ML problems

figure: is there a cat in this image?

algorithms and people

these algorithms are now being used to make predictions about people

- x is a description of a person
- y is an unobservable property of that person

algorithms and people

these algorithms are now being used to make predictions about people

- x is a description of a person
- y is an unobservable property of that person

original use cases revolved around digital marketing and product targeting:

- predict whether a web user will click on a particular ad
- predict whether a web user will purchase a particular good (and hence, whether to show them that ad)

in recent years, the scope of big data prediction problems has dramatically expanded

example 1: medical diagnosis

AI Now Diagnoses Disease Better Than Your Doctor, Study Finds

Peer-reviewed study says you'll soon consult Dr. Bot for a second opinion

HEALTH

Making the modern radiologist obsolete? How machine learning may revolutionize medicine

use cases:

- making medical diagnoses
- predicting which patients would benefit most from treatment

example 2: predicting who will be involved in crime

use cases:

- guiding judge decisions regarding whether to release a defendant on bail
- predicting places where crime is likely to occur

Predicting Recidivism Risk: New Tool in Philadelphia Shows Great Promise

by Nancy Ritter

Tool uses random forest modeling to identify probationers likely to reoffend within two years of returning to the community.

The tool — which has been successfully used in Philadelphia for four years — assesses each new probation case at its outset and assigns the probationer to a high-, moderate- or low-risk category. Although this is not a new concept, what is unique is that the tool uses “random forest modeling,” a sophisticated statistical approach that considers the nonlinear effects of a large number of variables with complex interactions (see sidebar, “What Is Random Forest Modeling?” on this page). Historically, corrections officials — in Philadelphia and elsewhere around the country — have used simpler statistical methods, such as linear regression models, to try to get a handle on the risk that a probationer may pose to the community.

example 3: predicting creditworthiness

Business

ZestFinance issues small, high-rate loans, uses big data to weed out deadbeats

use cases:

- setting credit limits
- guiding decisions about who should receive credit

what is different about these "social" prediction problems?

the criterion that we use to evaluate algorithms extend beyond accuracy

- it might not matter if an ML algorithm for digit recognition is twice as accurate for the digit 7 than for the digit 8
- but what if an ML algorithm is twice as accurate for assessing probability of committing a crime for one racial group than another?

enormous recent interest in the "fairness" of ML algorithms, defined as how the consequences of the ML algorithms vary across social groups

example 1: medical diagnosis

> Science. 2019 Oct 25;366(6464):447-453. doi: 10.1126/science.aax2342.

Dissecting racial bias in an algorithm used to manage the health of populations

Ziad Obermeyer [1] [2], Brian Powers [3], Christine Vogeli [4], Sendhil Mullainathan [5]

Affiliations + expand

PMID: 31649194 DOI: 10.1126/science.aax2342

Abstract

Health systems rely on commercial prediction algorithms to identify and help patients with complex health needs. We show that a widely used algorithm, typical of this industry-wide approach and affecting millions of patients, exhibits significant racial bias: At a given risk score, Black patients are considerably sicker than White patients, as evidenced by signs of uncontrolled illnesses. Remedying this disparity would increase the percentage of Black patients receiving additional help from 17.7 to 46.5%. The bias arises because the algorithm predicts health care costs rather than illness, but unequal access to care means that we spend less money caring for Black patients than for White patients. Thus, despite health care cost appearing to be an effective proxy for health by some measures of predictive accuracy, large racial biases arise. We suggest that the choice of convenient, seemingly effective proxies for ground truth can be an important source of algorithmic bias in many contexts.

example 2: predicting who will be involved in crime

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

ARTIFICIAL INTELLIGENCE

Predictive policing algorithms are racist. They need to be dismantled.

Lack of transparency and biased training data mean these tools are not fit for purpose. If we can't fix them, we should ditch them.

example 3: predicting creditworthiness

Apple Card Investigated After Gender Discrimination Complaints

A prominent software developer said on Twitter that the credit card was “sexist” against women applying for credit.

the response

algorithm designers increasingly optimize not only for accuracy but also "fairness" (maintain comparable error rates across groups)

max **accuracy**

subject to **unfairness** $\leq \varepsilon$

the response

algorithm designers increasingly optimize not only for accuracy but also "fairness" (maintain comparable error rates across groups)

max **accuracy**

subject to **disparity in errors across groups** $\leq \varepsilon$

plan for the talk

three papers on this topic:

1. **algorithm design: a fairness-accuracy frontier** (liang, lu, mu, and okumura)

2. **testing the fairness-accuracy improvability of algorithms** (auerbach, liang, tabord-meehan, okumura)

3. **algorithmic fairness and social welfare** (liang and lu)

Algorithm Design: A Fairness-Accuracy Frontier

Annie Liang (Northwestern)
Jay Lu (UCLA)
Xiaosheng Mu (Princeton)
Kyohei Okumura (Northwestern)

introduction

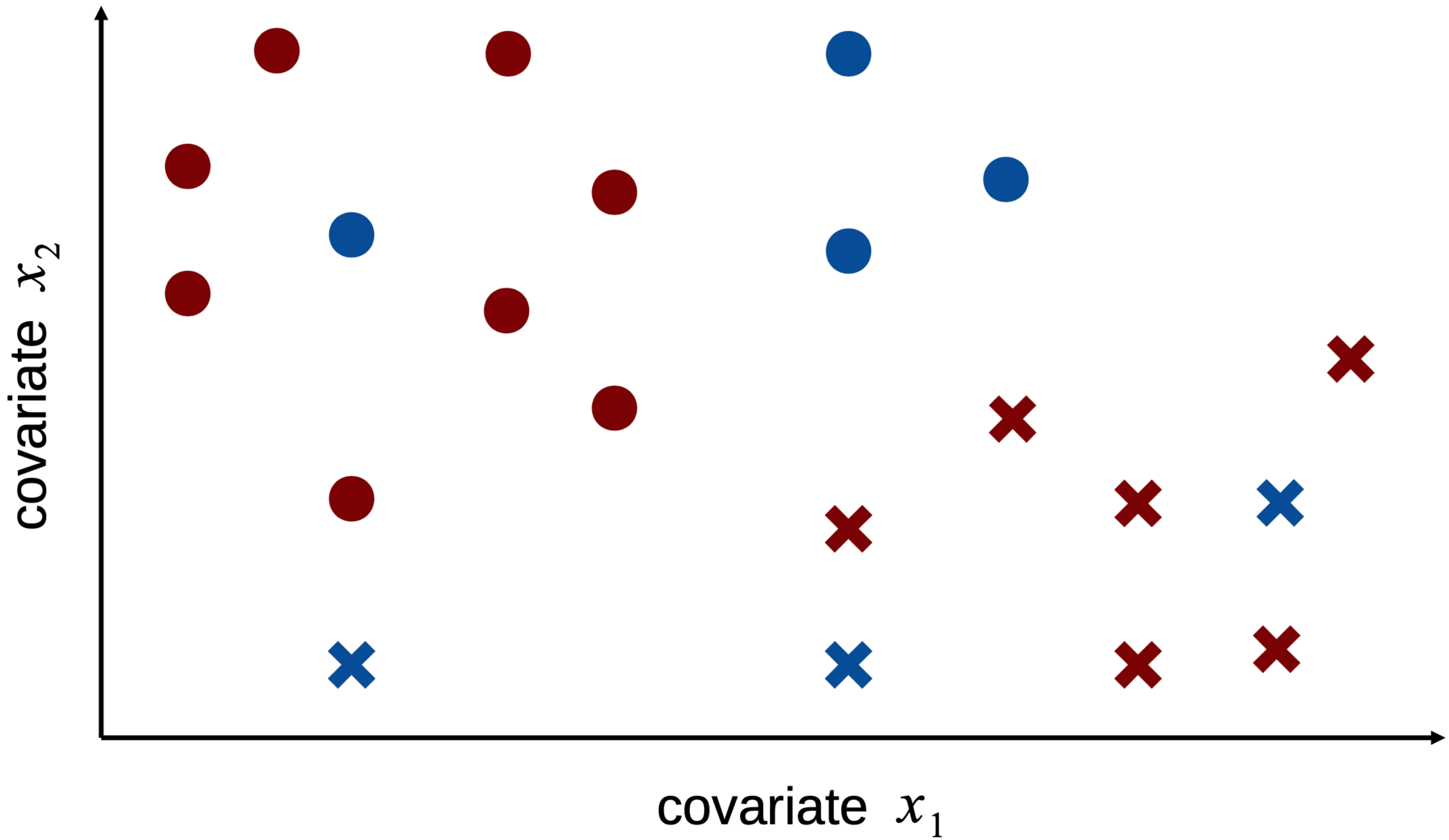
ideally the algorithm would be perfectly accurate and "fair" across groups, in practice there can be a conflict between these goals

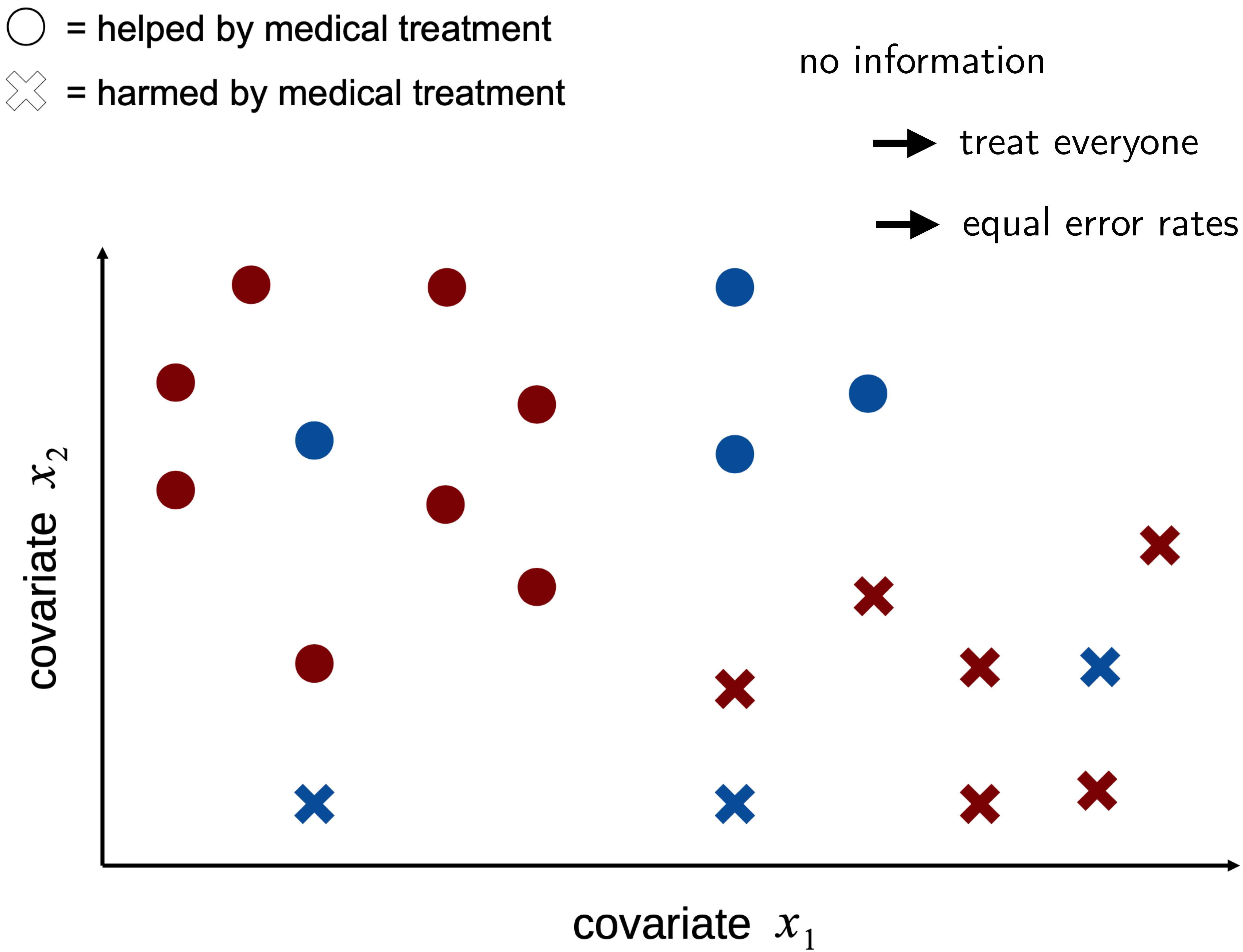
this paper: general framework for formalizing this tradeoff, and identification of simple properties of the algorithm's inputs that determine its shape

○ = helped by medical treatment

✕ = harmed by medical treatment

two groups:
red and **blue**



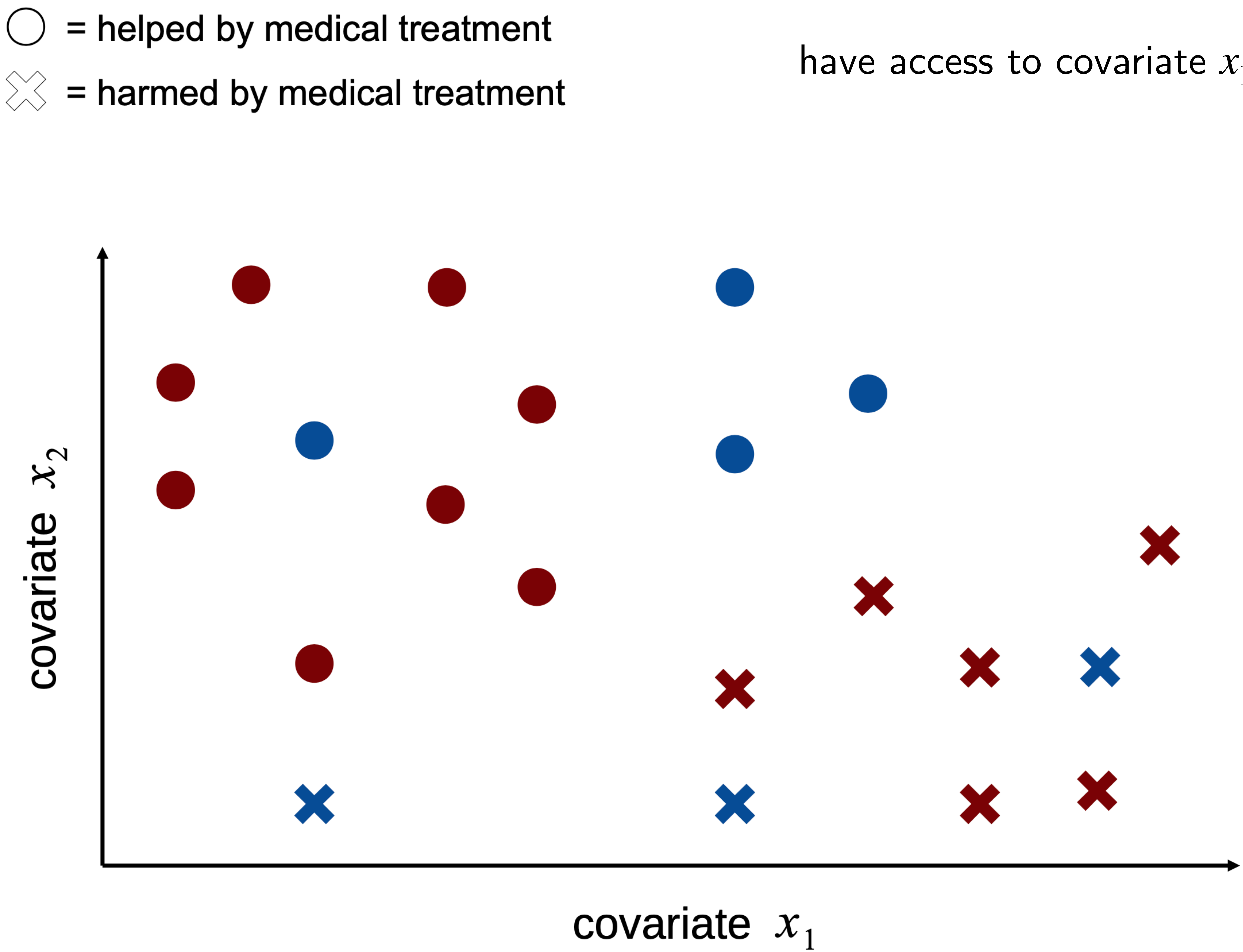


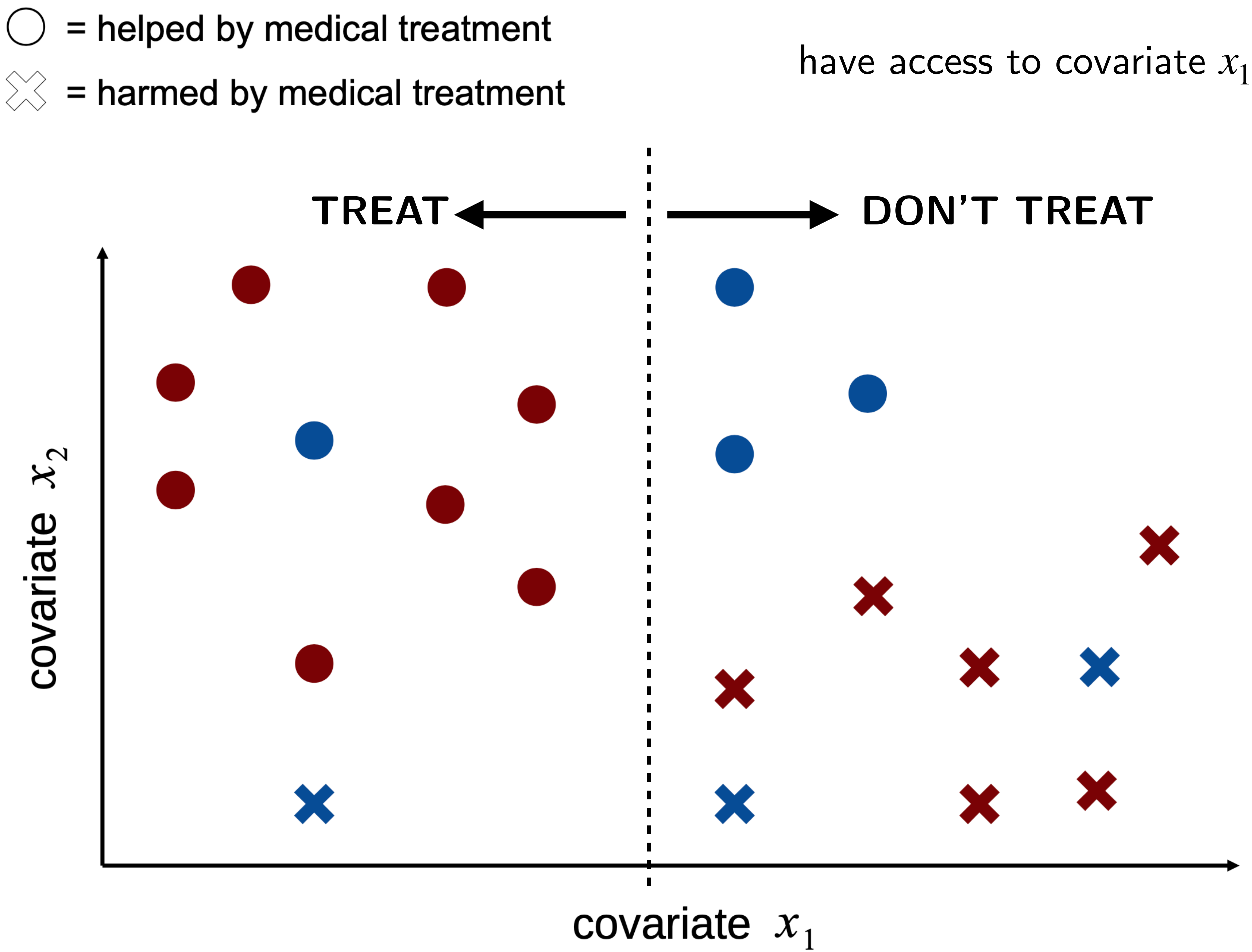
no information

→ treat everyone

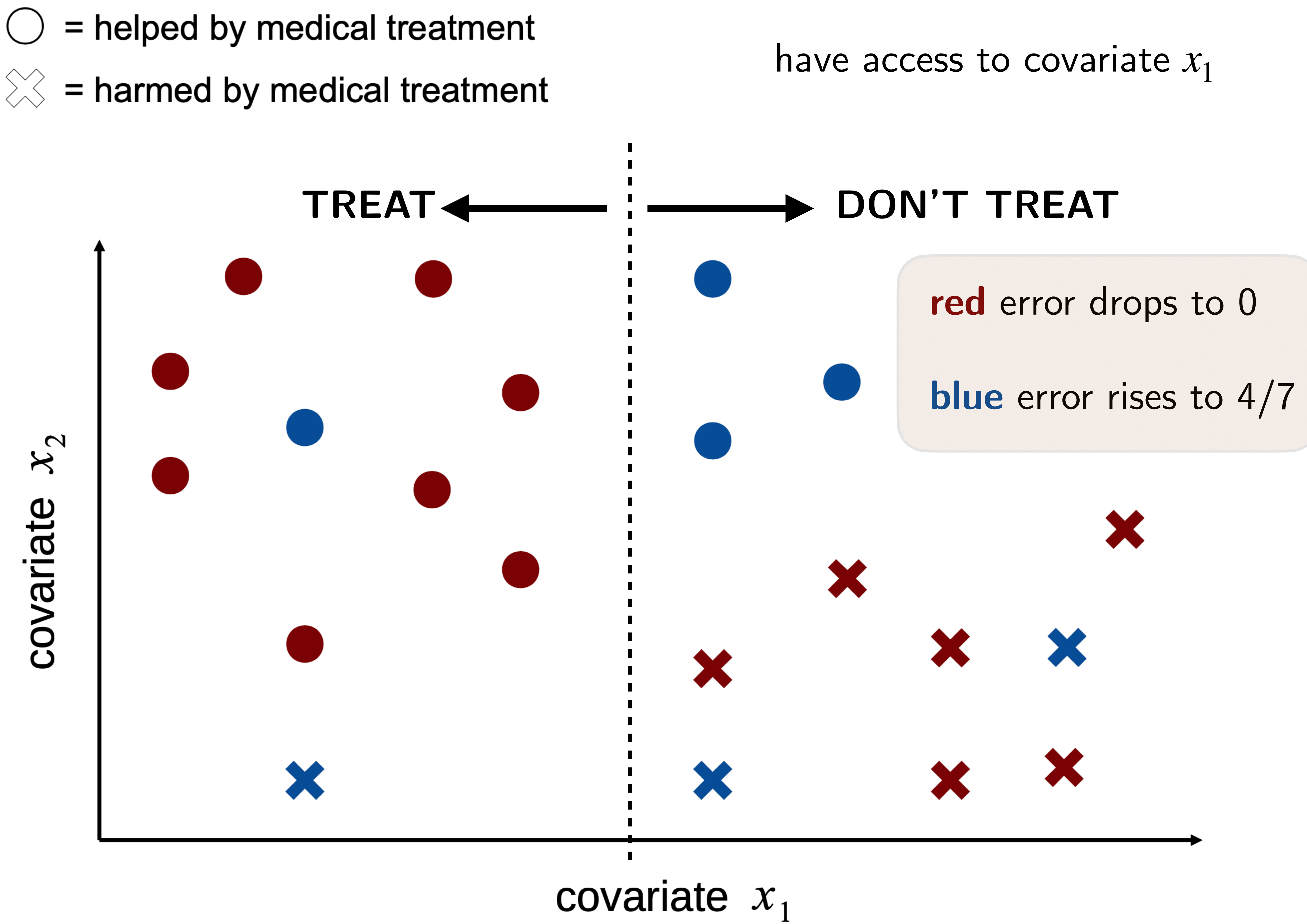
→ equal error rates (3/7)

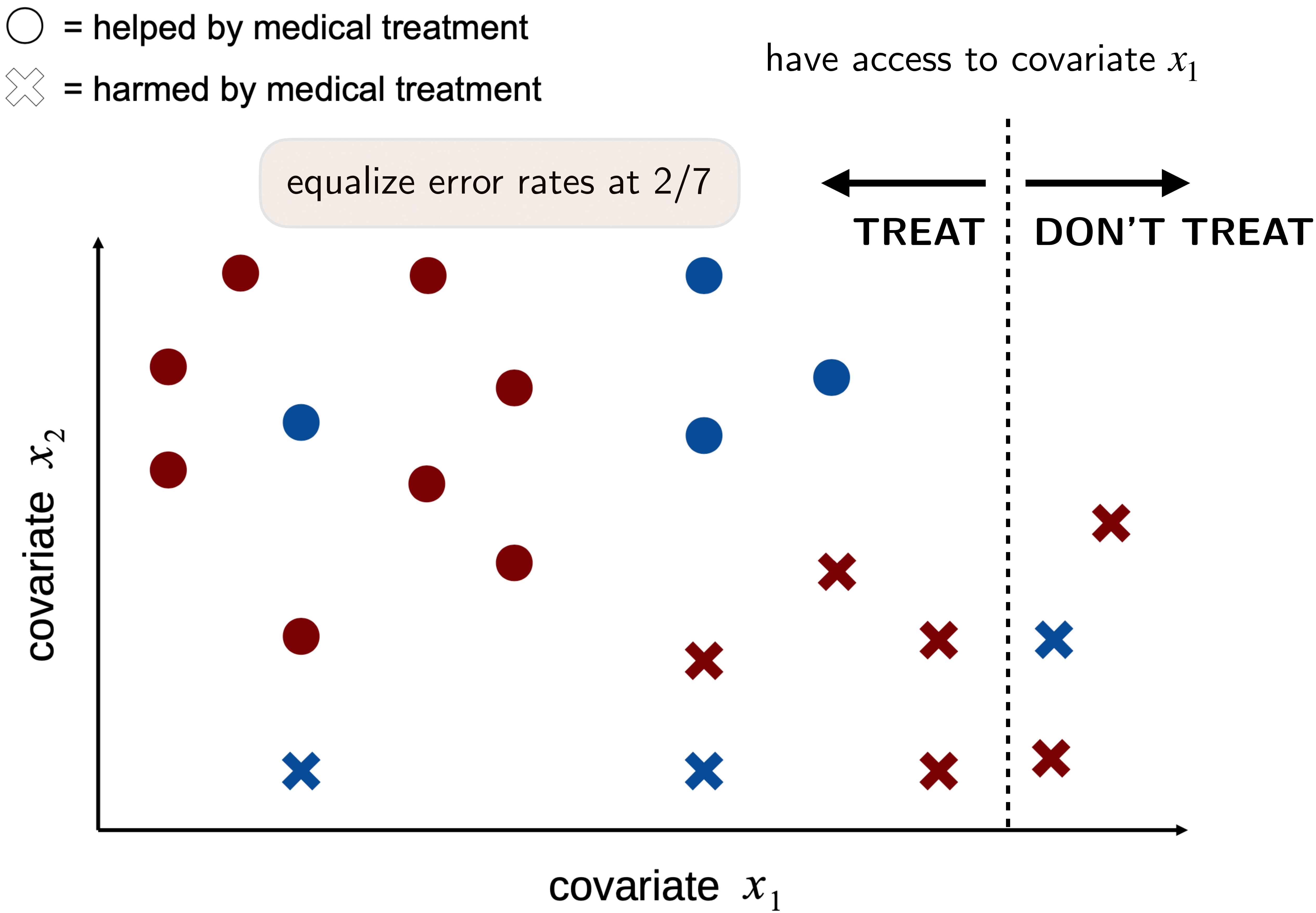
have access to covariate x_1





= helped by medical treatment
= harmed by medical treatment
have access to covariate x_1
TREAT
DON'T TREAT
covariate x_2
covariate x_1


= helped by medical treatment
= harmed by medical treatment
have access to covariate x_1
TREAT
DON'T TREAT
red error drops to 0
blue error rises to 4/7
covariate x_2
covariate x_1

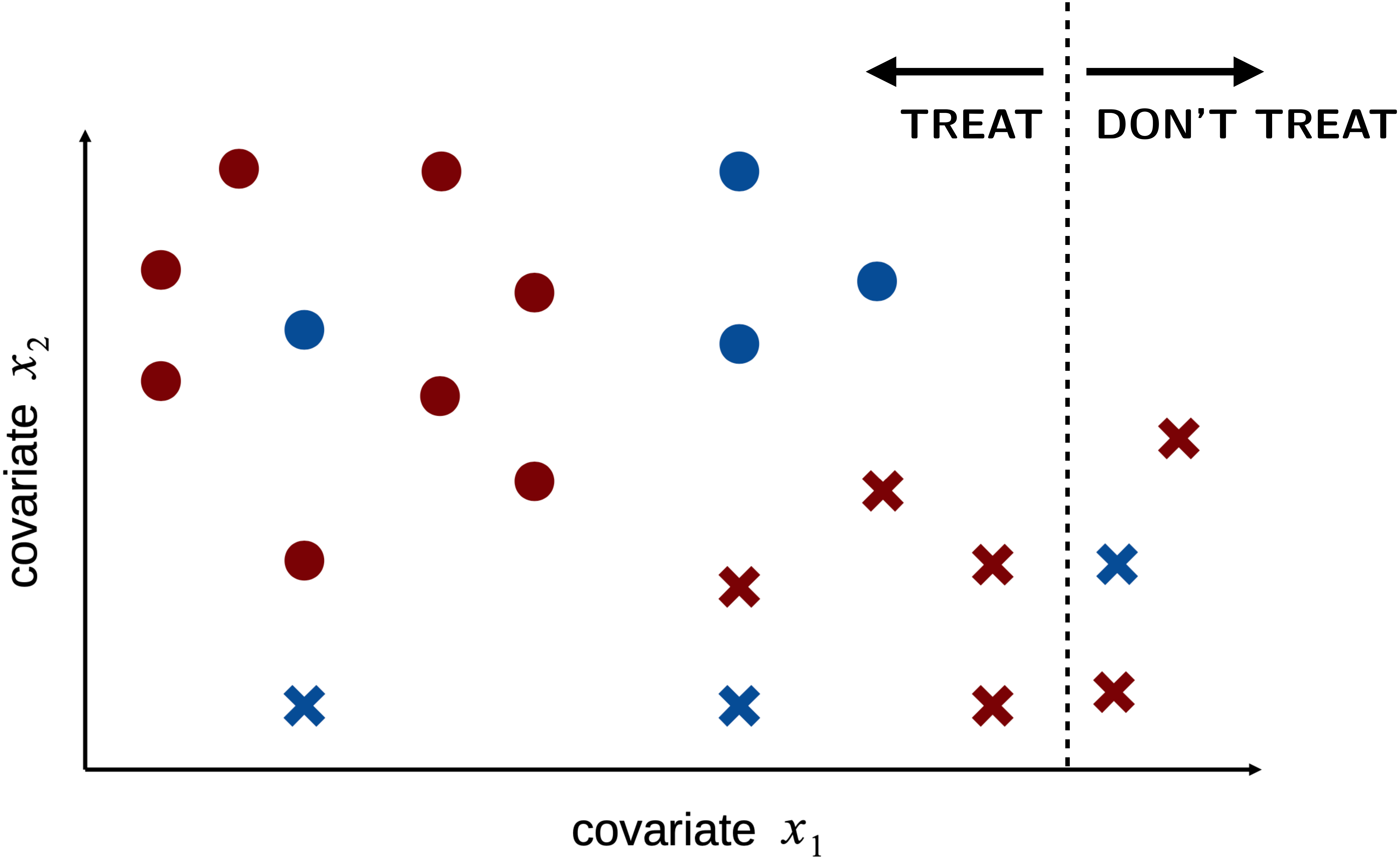


= helped by medical treatment
= harmed by medical treatment
have access to covariate x_1
equalize error rates at 2/7
TREAT
DON'T TREAT
covariate x_2
covariate x_1

◯ = helped by medical treatment

✕ = harmed by medical treatment

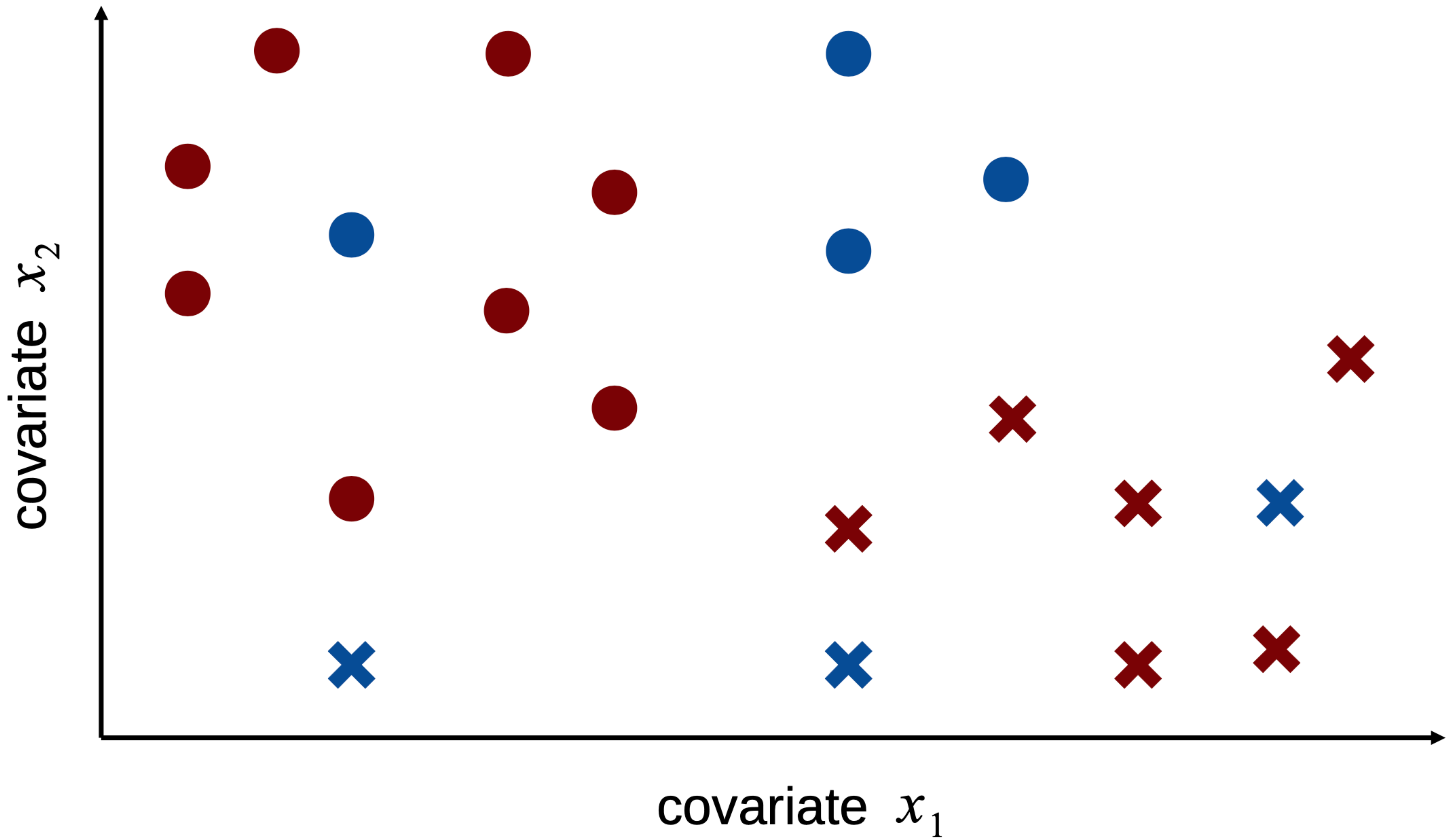
there is a fairness-accuracy tradeoff

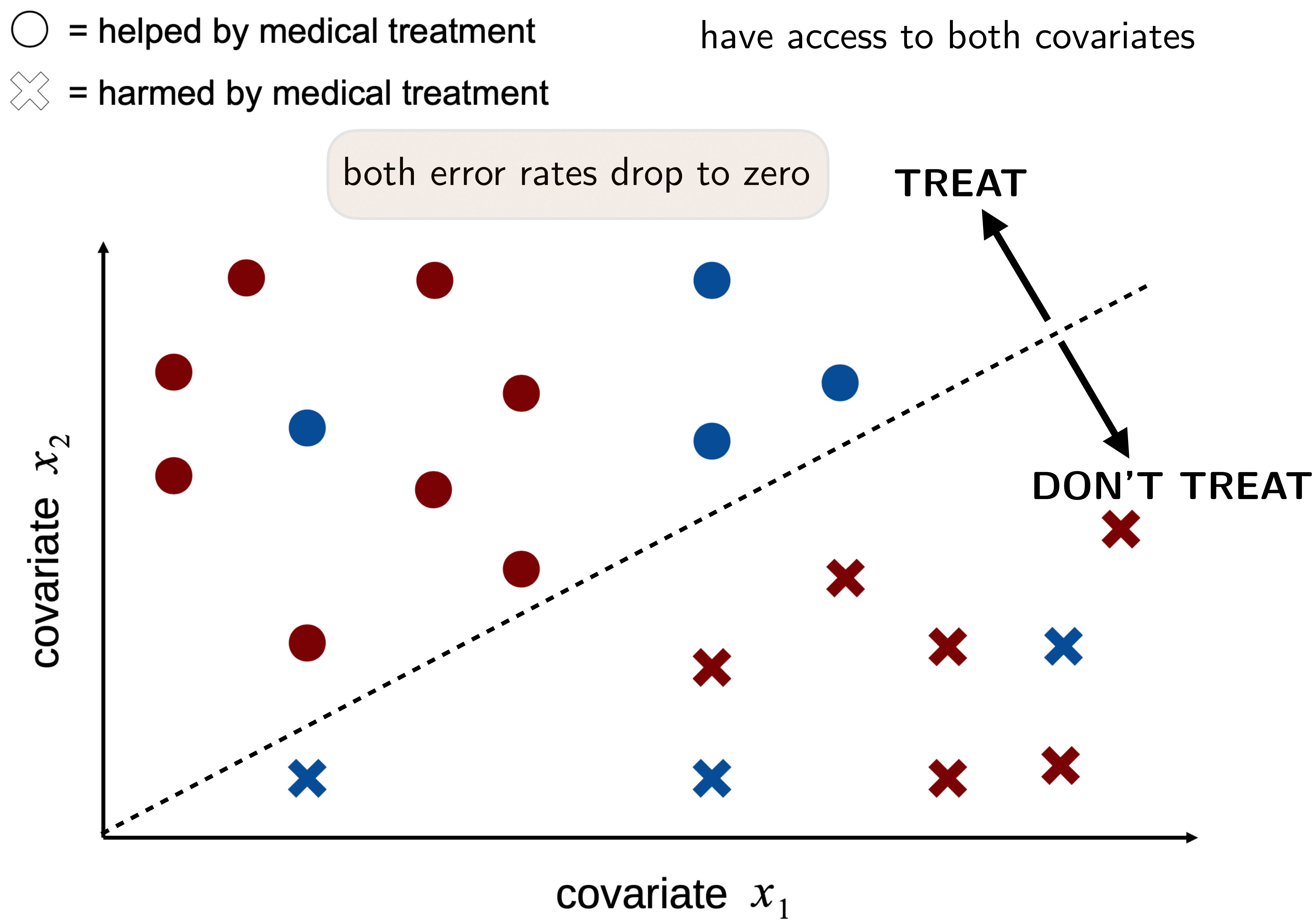


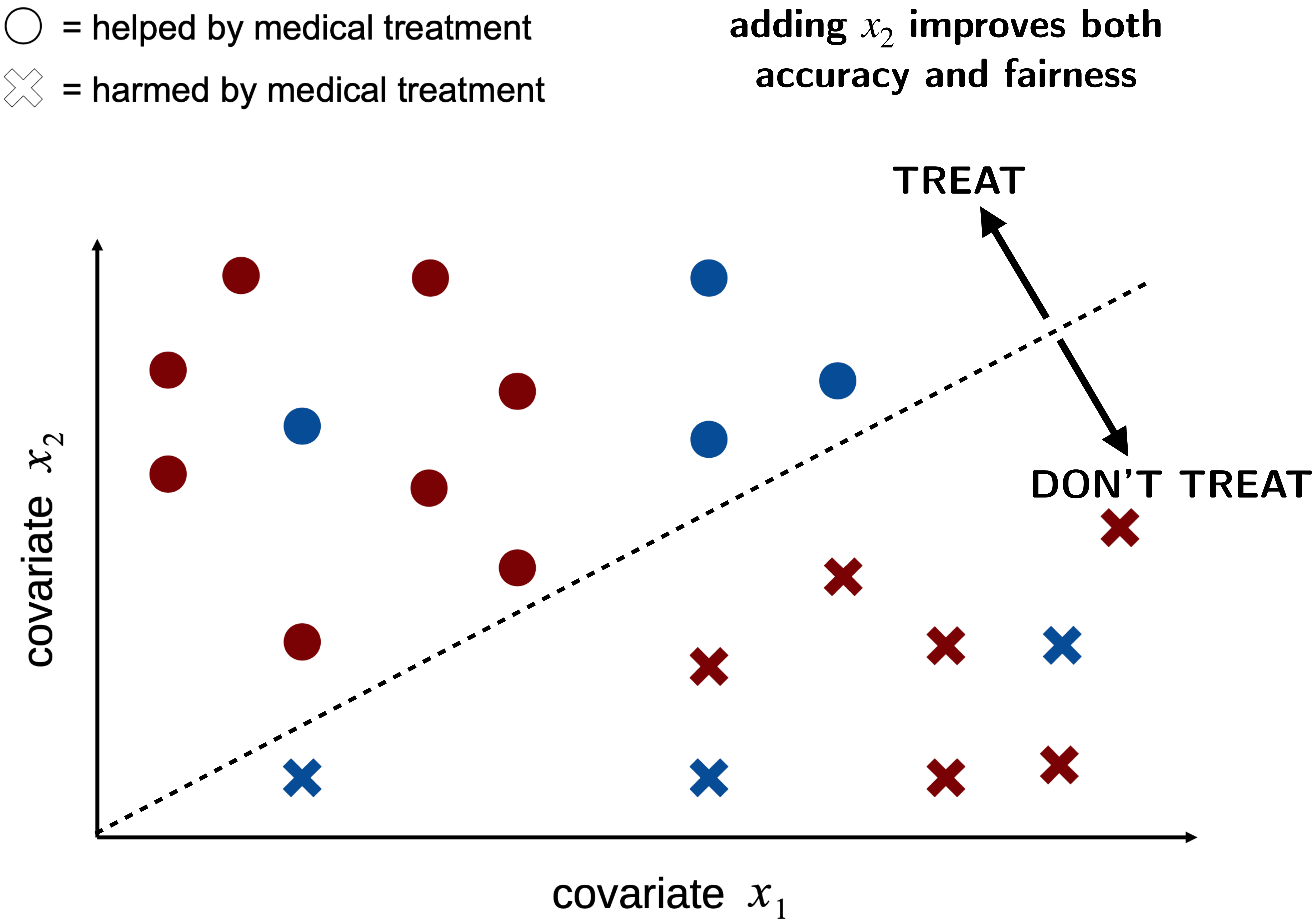
= helped by medical treatment

= harmed by medical treatment

have access to both covariates




= helped by medical treatment
= harmed by medical treatment
have access to both covariates
both error rates drop to zero
TREAT
DON'T TREAT
covariate x_2
covariate x_1


= helped by medical treatment
= harmed by medical treatment
adding x_2 improves both accuracy and fairness
TREAT
DON'T TREAT
covariate x_2
covariate x_1

model

setup

- each subject is described by three variables:
  - **type** Y taking values in $\mathscr{Y}$
    (e.g., need for medical procedure)
  - **group** $G \in \mathscr{G} = \{r, b\}$
    (e.g., race)
  - **covariate vector** X taking values in $\mathscr{X}$
    (e.g., image scans, number of past hospital visits, blood tests)

setup

- each subject is described by three variables:
  - **type** Y taking values in $\mathscr{Y}$
    (e.g., need for medical procedure)
  - **group** $G \in \mathscr{G} = \{r, b\}$
    (e.g., race)
  - **covariate vector** X taking values in $\mathscr{X}$
    (e.g., image scans, number of past hospital visits, blood tests)
- in the population, $(X, Y, G) \sim P$ (with no restrictions on P)
- an algorithm is a map $a : \mathscr{X} \to \{0,1\}$ from covariate vectors into a decision $d \in \{0,1\}$

setup

- each subject is described by three variables:
  - **type** Y taking values in $\mathscr{Y}$
    (e.g., need for medical procedure)
  - **group** $G \in \mathscr{G} = \{r, b\}$
    (e.g., race)
  - **covariate vector** X taking values in $\mathscr{X}$
    (e.g., image scans, number of past hospital visits, blood tests)
- in the population, $(X, Y, G) \sim P$ (with no restrictions on P)
- an algorithm is a map $a : \mathscr{X} \to \{0,1\}$ from covariate vectors into a decision $d \in \{0,1\}$
- a policymaker chooses from a set of of algorithms $\mathscr{A}$ (e.g., linear rules) and their randomizations — for most of the talk, let $\mathscr{A}$ be unconstrained

how algorithms are evaluated

- primitive loss function $\ell(y, d)$ expresses the "inaccuracy" of decision d for an individual with type y
  - e.g., $\ell(y, d) = 1(y \neq d)$ if d is a prediction of y
  - e.g., a convex combination of Type I and Type II errors

how algorithms are evaluated

- primitive loss function $\ell(y, d)$ expresses the "inaccuracy" of decision d for an individual with type y
  - e.g., $\ell(y, d) = 1(y \neq d)$ if d is a prediction of y
  - e.g., a convex combination of Type I and Type II errors

- **definition:** the *group error* $e_g(a) \equiv \mathbb{E}[\ell(Y, a(X)) \mid G = g]$ is the **average loss** for members of group g under algorithm a
  - for the first loss function, $e_g(a)$ is the fraction of incorrect predictions ("misclassification rate") for group g members

how algorithms are evaluated

- primitive loss function $\ell(y, d)$ expresses the "inaccuracy" of decision d for an individual with type y
  - e.g., $\ell(y, d) = 1(y \neq d)$ if d is a prediction of y
  - e.g., a convex combination of Type I and Type II errors

- **definition:** the *group error* $e_g(a) \equiv \mathbb{E}[\ell(Y, a(X)) \mid G = g]$ is the **average loss** for members of group g under algorithm a
  - for the first loss function, $e_g(a)$ is the fraction of incorrect predictions ("misclassification rate") for group g members

- the policymaker evaluates algorithm a based on the induced group errors $(e_r(a), e_b(a)) \in \mathbb{R}^2$
  - improving **accuracy**: lowering e_r and e_b
  - improving **fairness**: lowering $|e_r - e_b|$

example preferences

- **utilitarian:** minimize $p_r e_r + p_b e_b$ where p_r and p_b are the proportions of either group (or **generalized utilitarian**: minimize $\alpha_r r e_r + \alpha_b e_b$**)**
- **egalitarian:** minimize $\left| e_r - e_b \right|$ (break ties using utilitarian rule)
- **rawlsian:** minimize $\max \left\{ e_r, e_b \right\}$ (break ties using utilitarian rule)
- **constrained optimization:** (e.g., Hardt et al, 2016)

$$\min_{a:\mathcal{X}\to\Delta(\mathcal{D})} \quad p_r e_r(a) + p_b e_b(a) \quad \text{s.t. } |e_r(a) - e_b(a)| \leq \varepsilon$$

broad class of fairness-accuracy preferences

a **fairness-accuracy preference** is any preference over group error pairs (e_r, e_b) consistent with the following partial order:

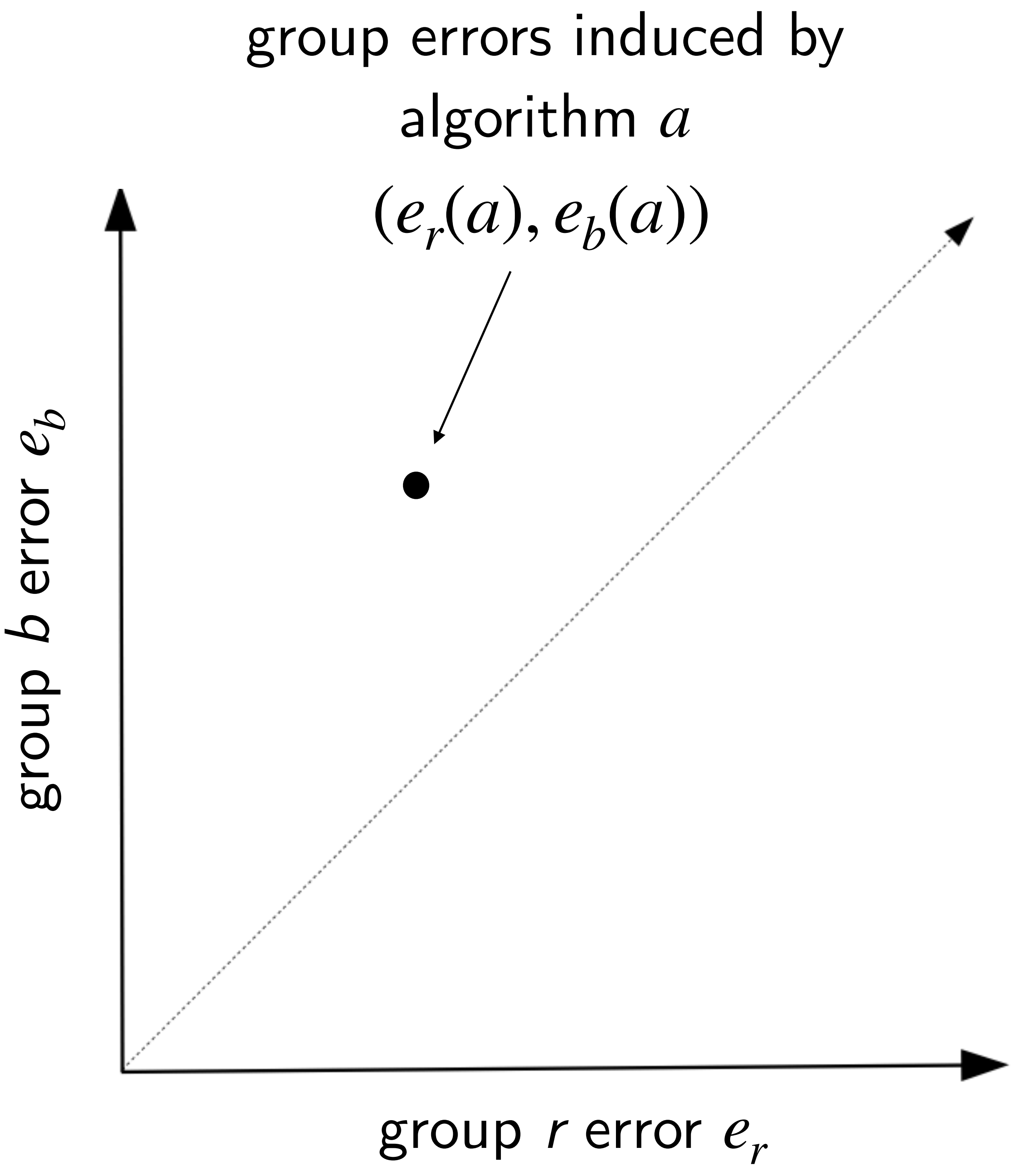
definition: $(e_r, e_b) >_{FA} (e'_r, e'_b)$ (in words: (e_r, e_b) FA-dominates (e'_r, e'_b)) if

$$\underbrace{e_r \leq e'_r, \quad e_b \leq e'_b,}_{\text{higher accuracy}} \quad \text{and } \underbrace{|e_r - e_b| \leq |e'_r - e'_b|}_{\text{higher fairness}}$$

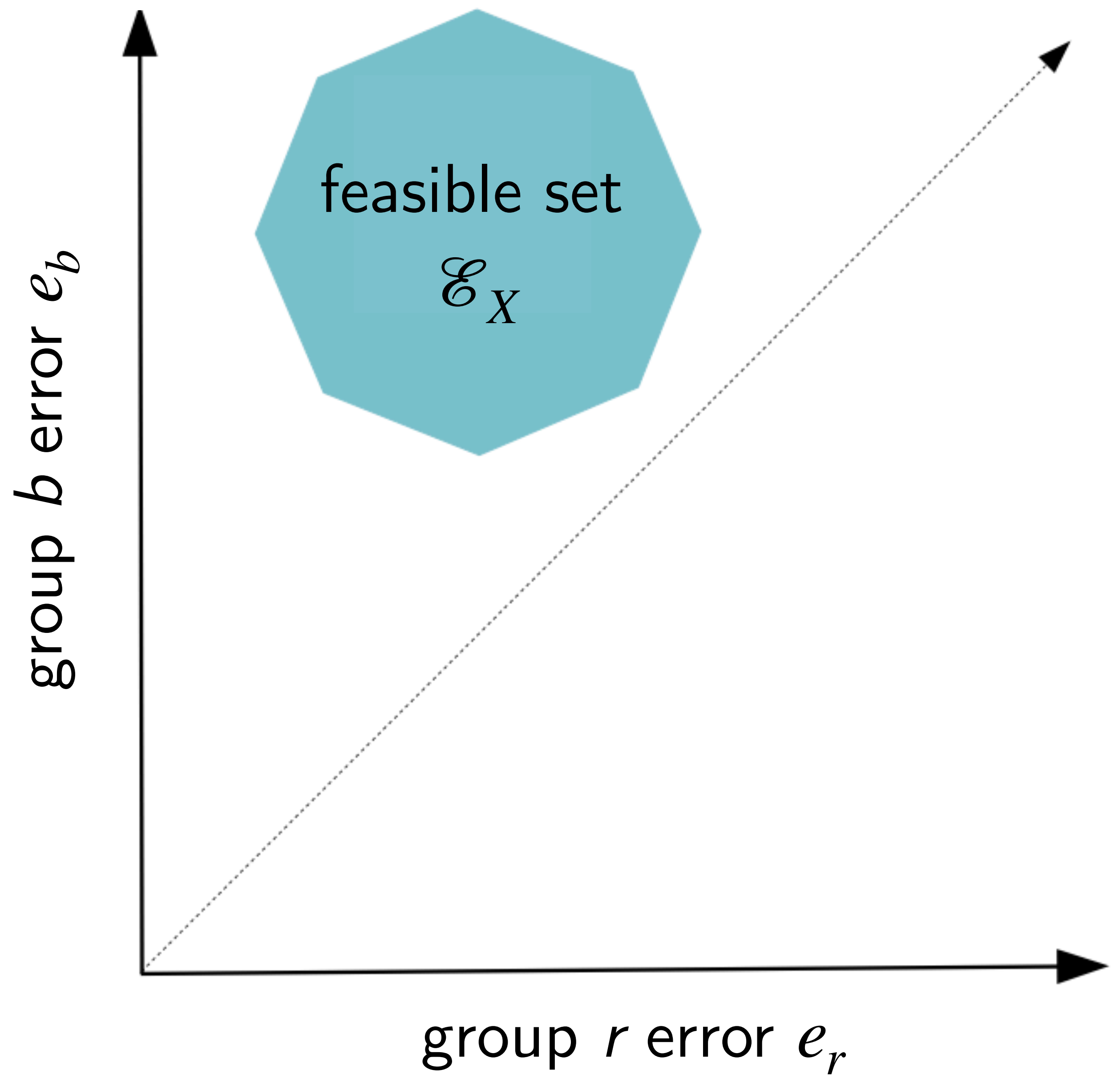
with at least one of these inequalities strict

- includes all of the previous example preferences

feasible set

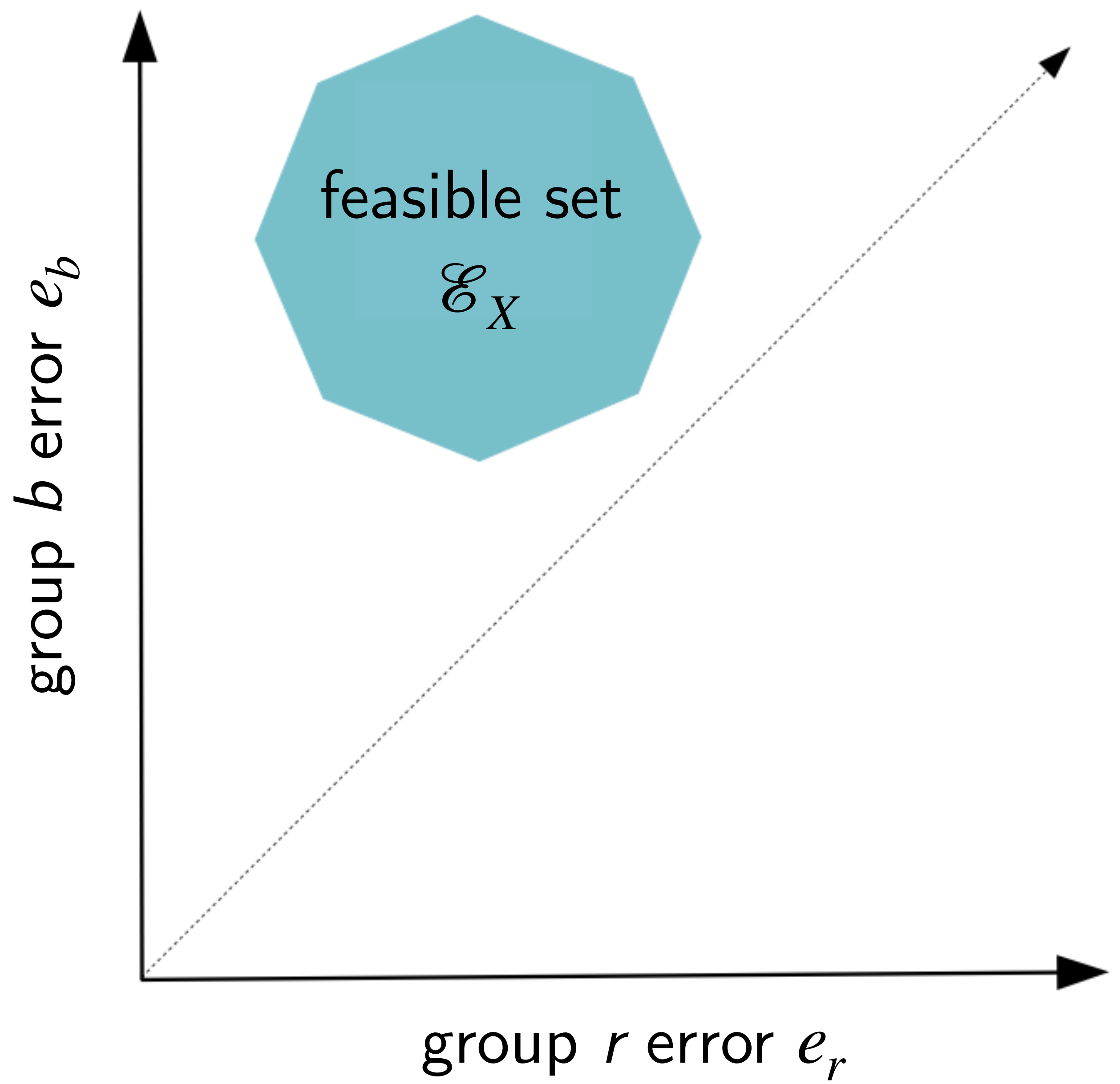


feasible set



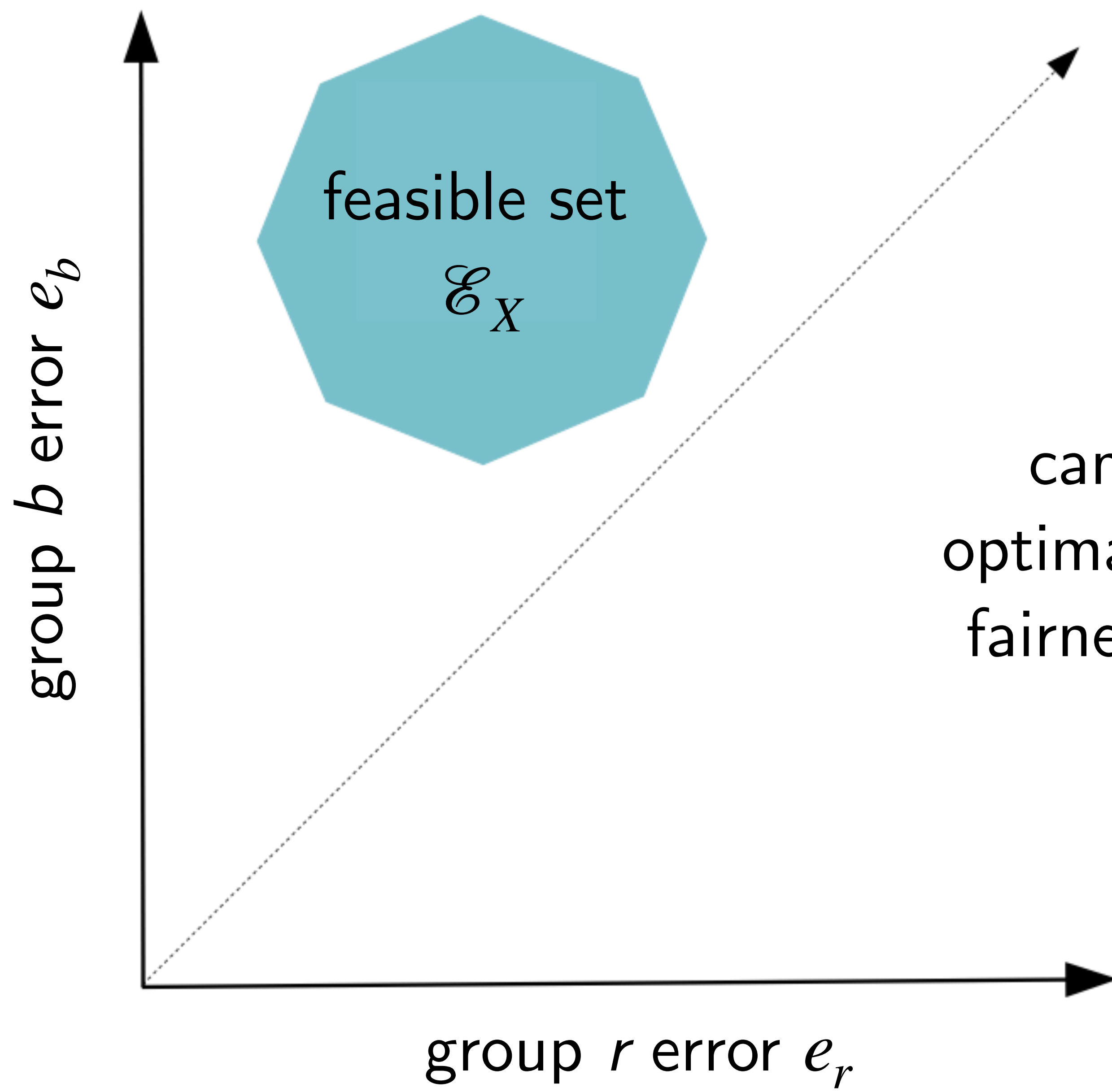
the **feasible set given** X (denoted $\mathscr{E}_X$) consists of all pairs (e_r, e_b) that can be implemented using some algorithm in $\Delta(\mathscr{A})$

fairness-accuracy frontier



the **fairness-accuracy frontier given** X (denoted $\mathscr{F}_X$) consists of all feasible (e_r, e_b) that are undominated in the $>_{FA}$-order (i.e., not possible to improve both accuracy and fairness)

fairness-accuracy frontier

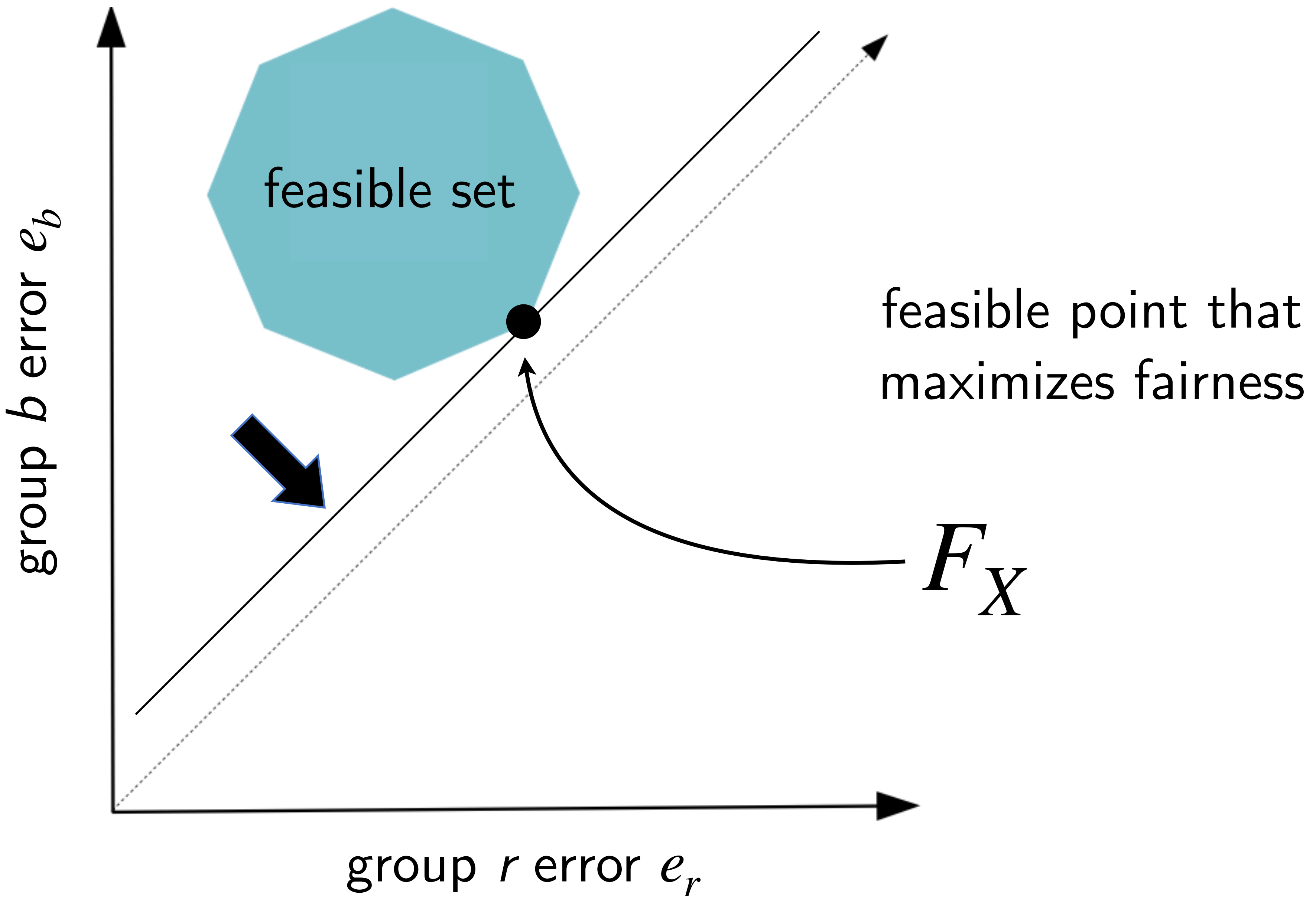


can also interpret as the optimal points across different fairness-accuracy preferences

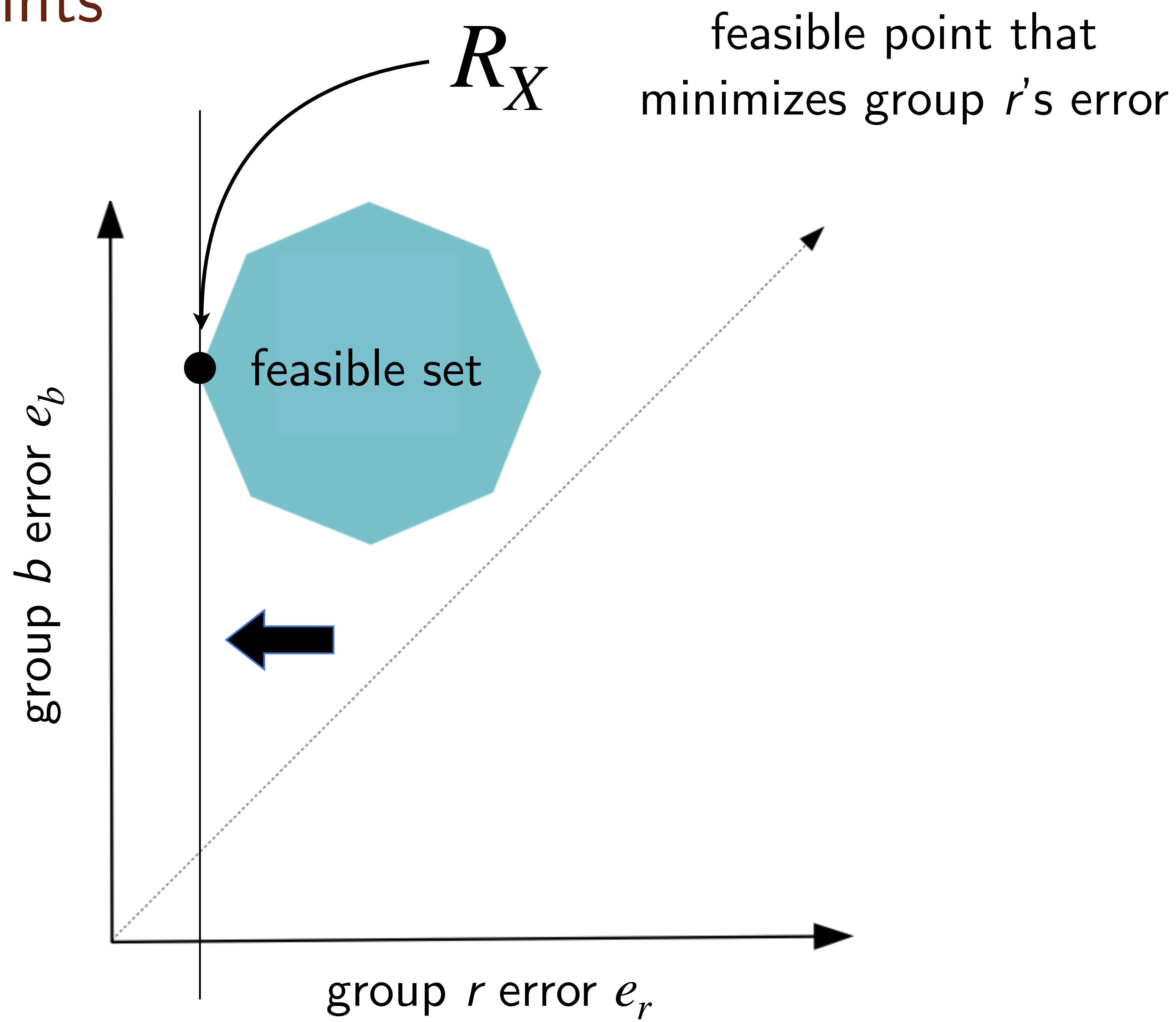
the **fairness-accuracy frontier given** X (denoted $\mathcal{F}_X$) consists of all feasible (e_r, e_b) that are undominated in the $>_{FA}$-order (i.e., not possible to improve both accuracy and fairness)

characterization of the fairness-accuracy frontier

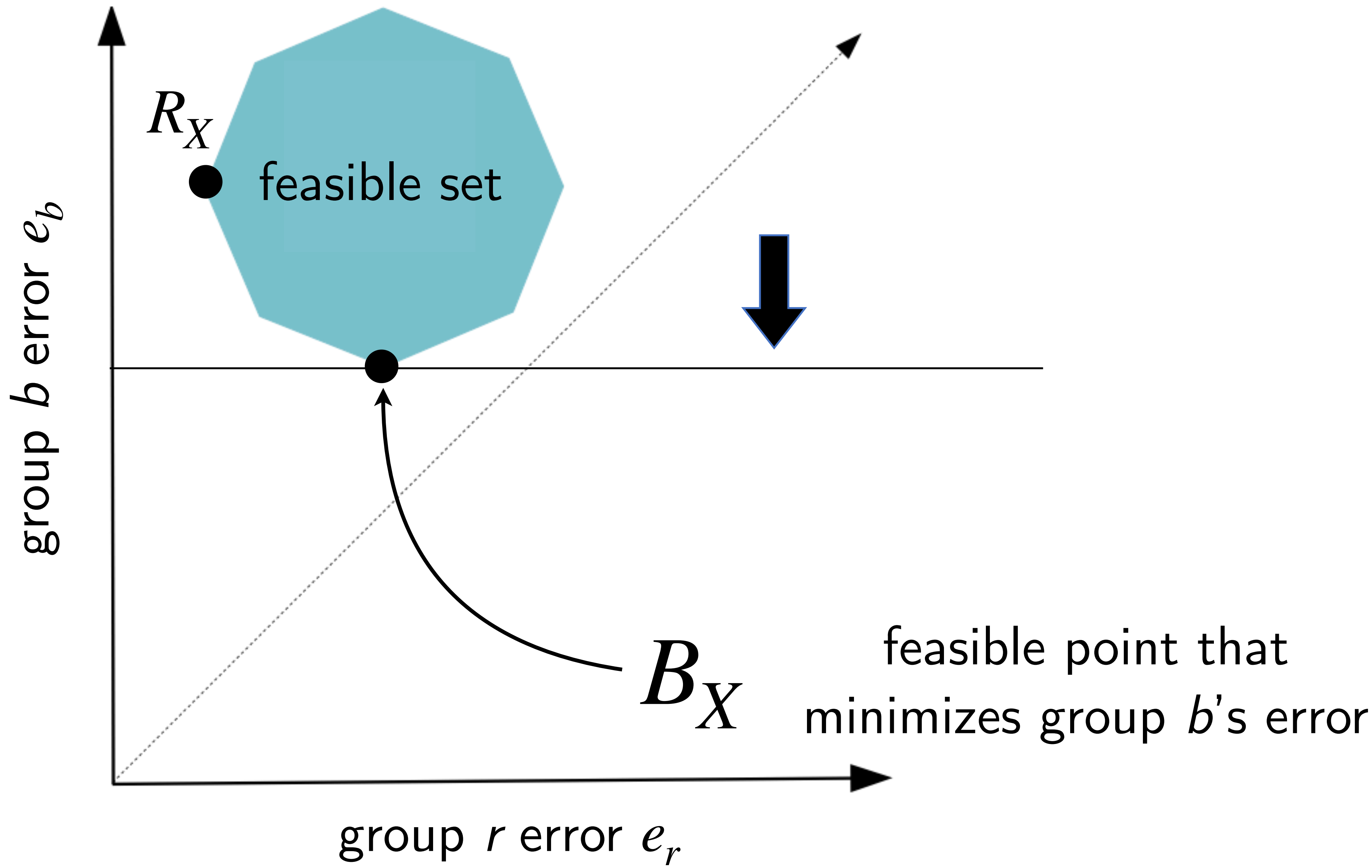
important points



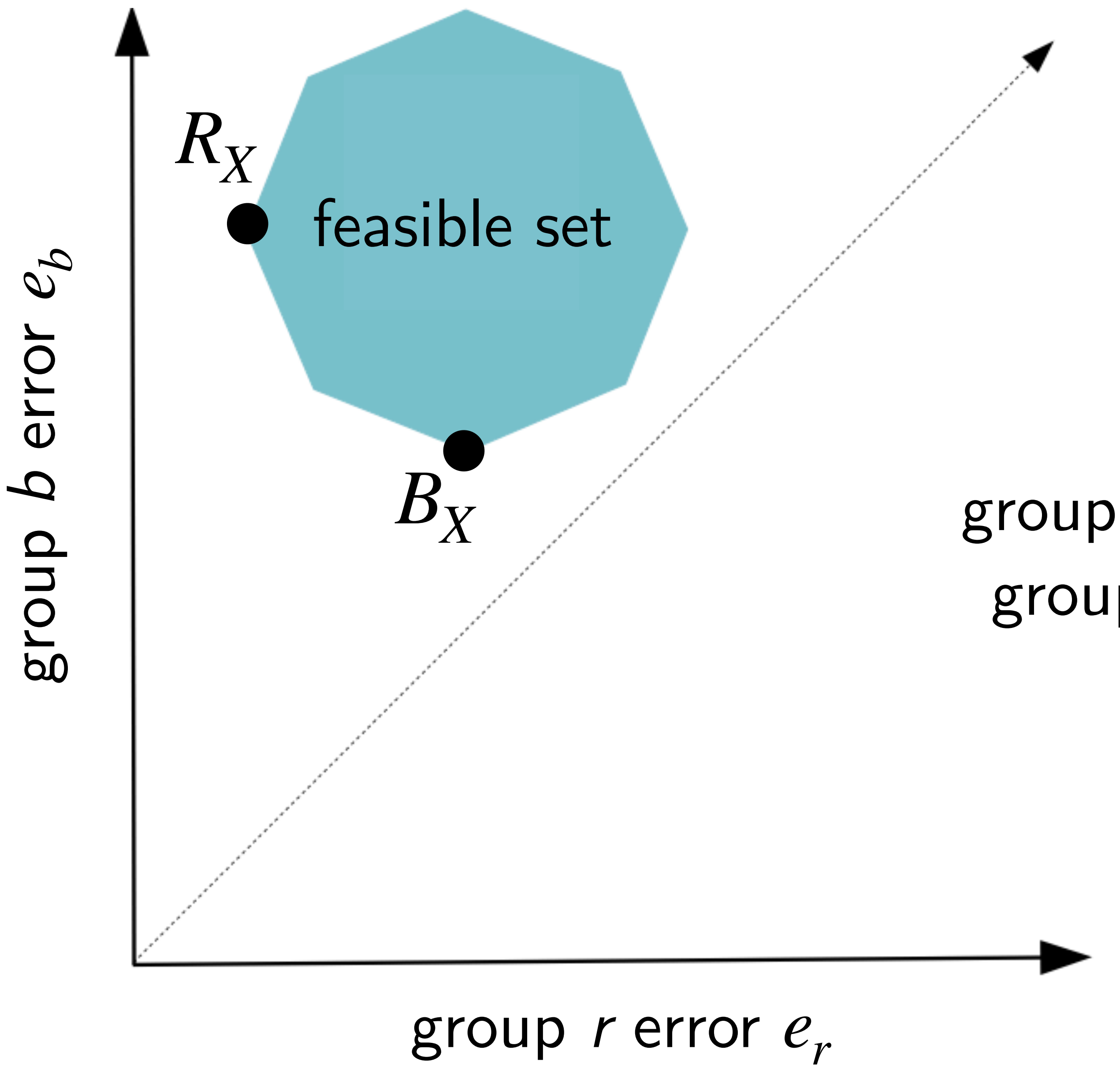
important points



important points



important points

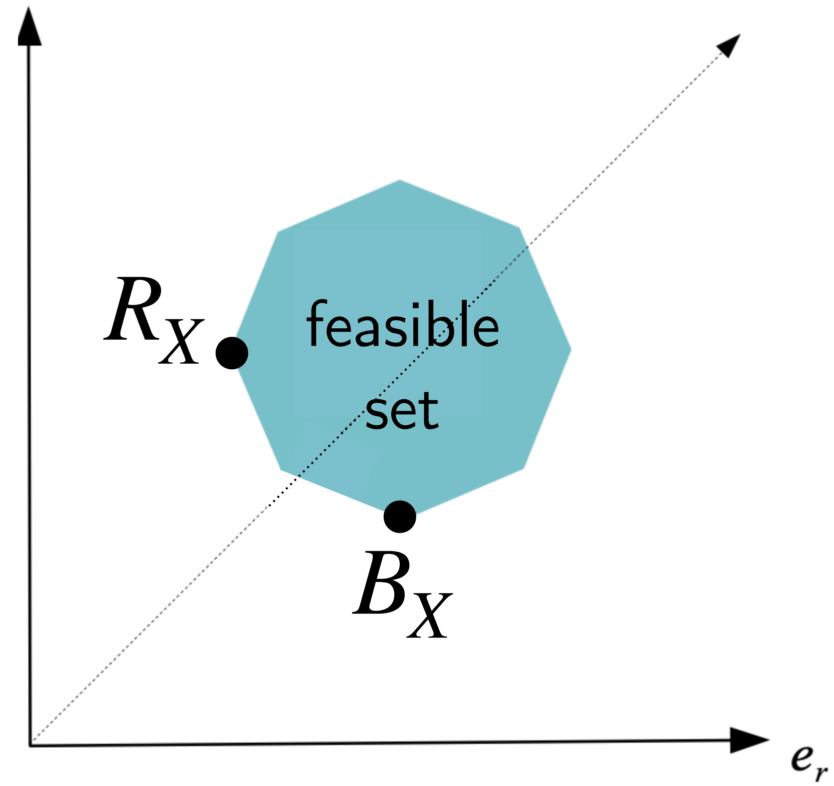


observe:

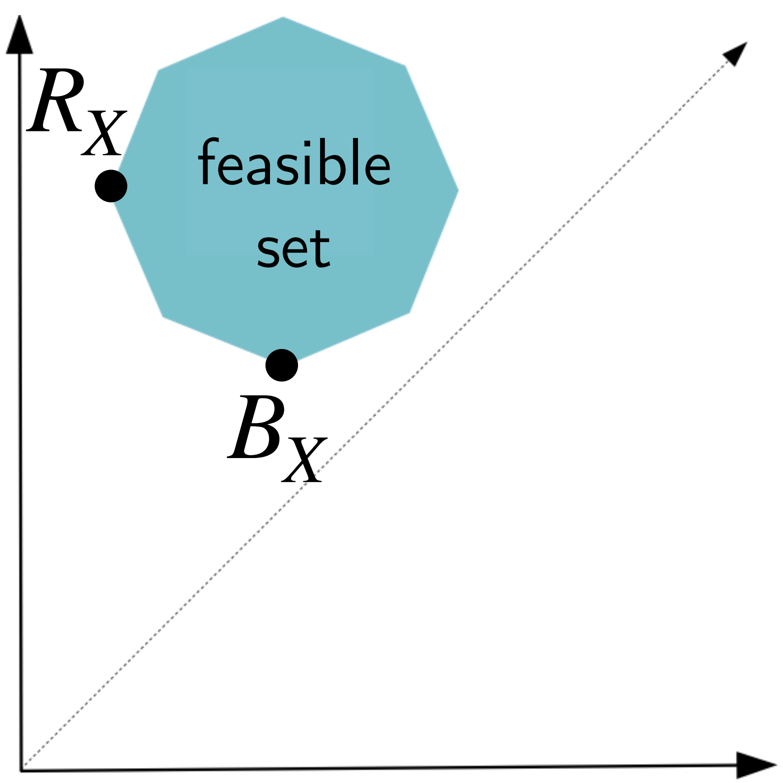
group b's error is higher than group r's even at group b's favorite point, B_X

group balance and group skew

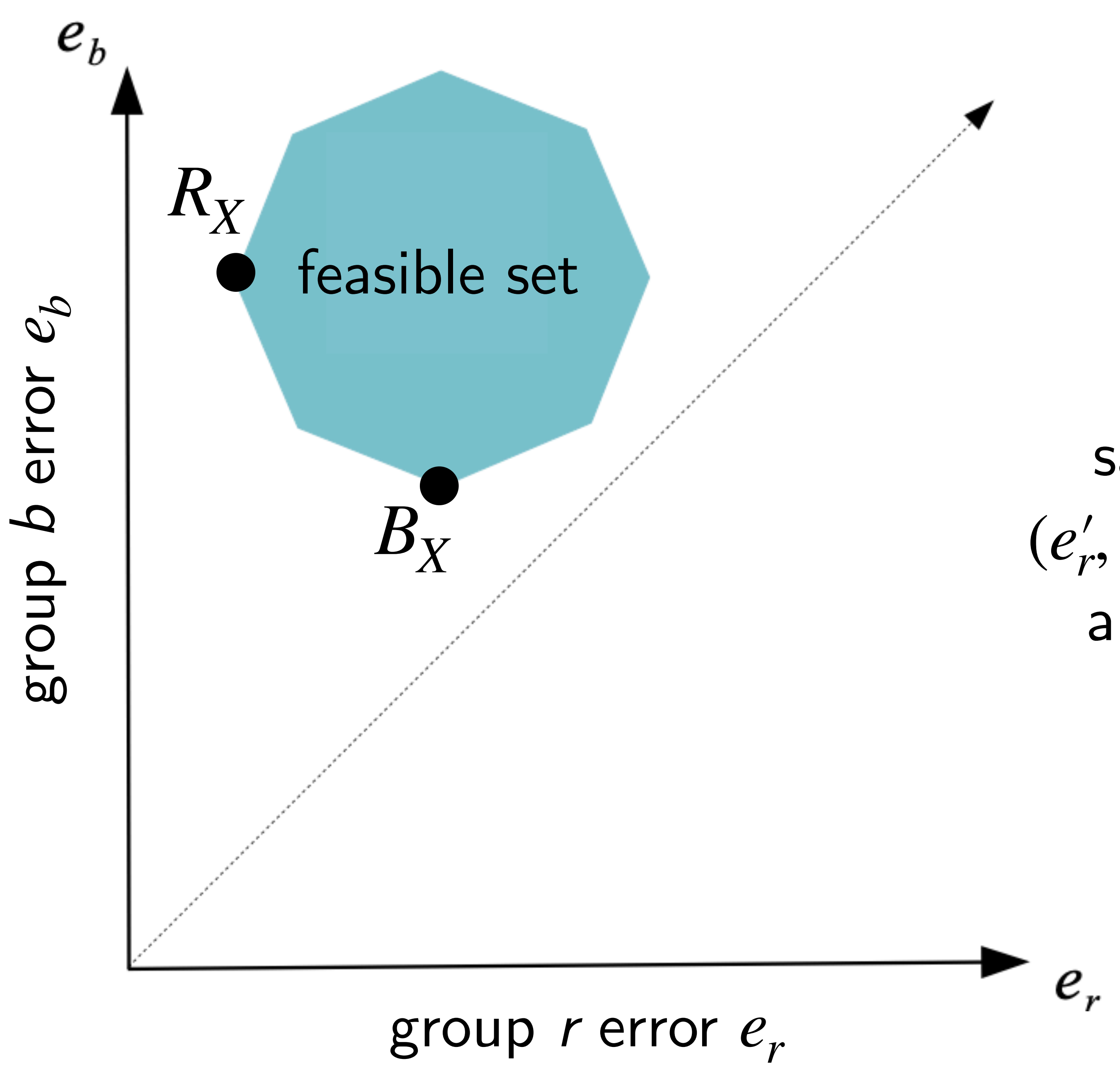
X is **group-balanced**



X is **group-skewed**

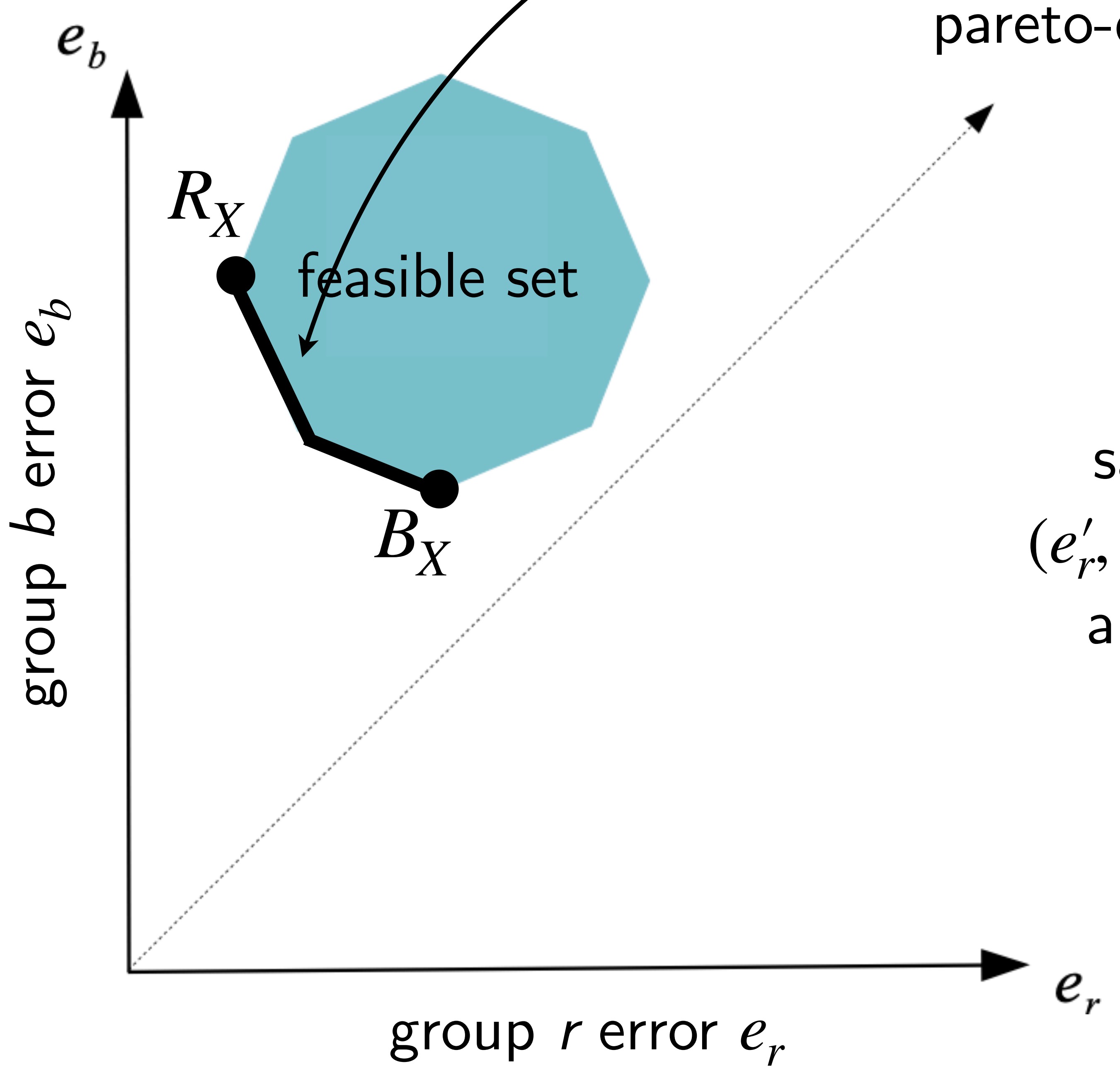


pareto frontier



say that (e_r, e_b) pareto-dominates (e_r', e_b') if both group errors are smaller and one is at least strictly smaller

pareto frontier



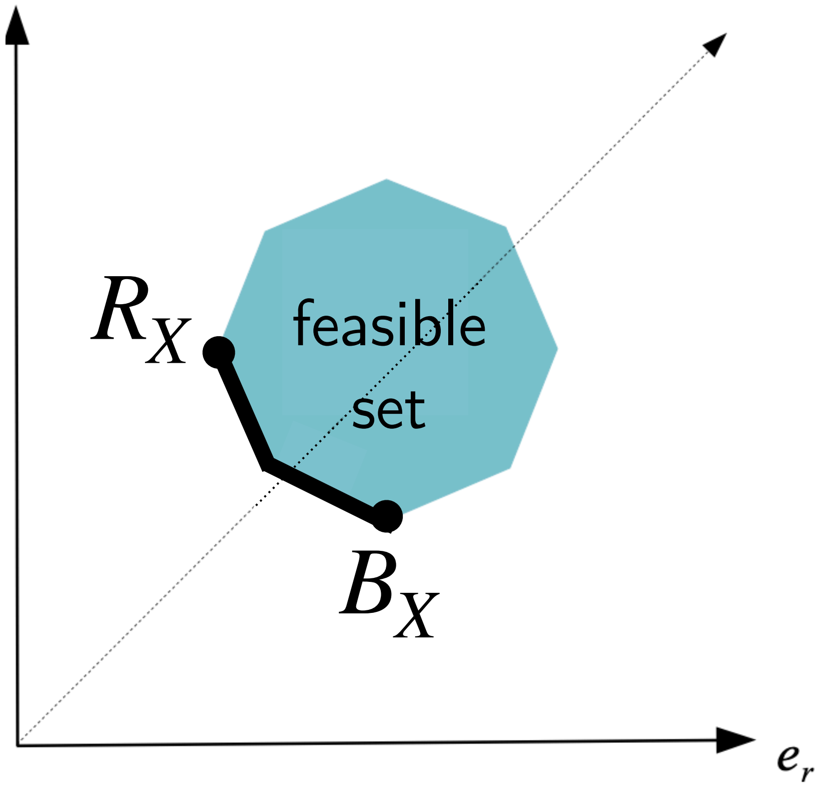
usual **pareto frontier**:

all feasible points which are not pareto-dominated

say that (e_r, e_b) pareto-dominates (e_r', e_b') if both group errors are smaller and one is at least strictly smaller

characterization of the frontier

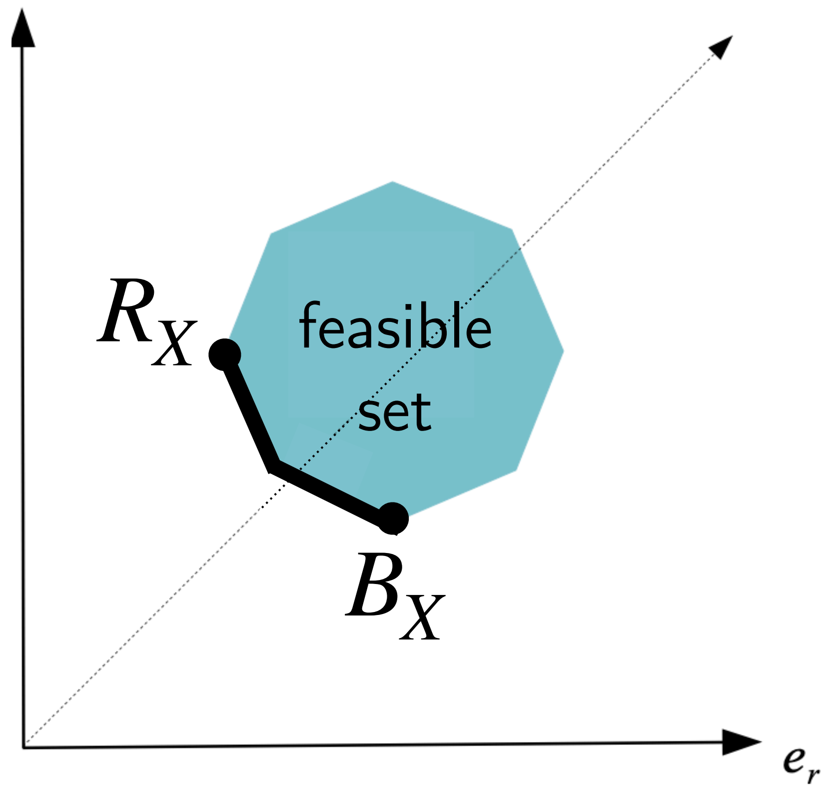
X is **group-balanced**



the fairness-accuracy frontier is precisely the usual pareto frontier

characterization of the frontier

X is **group-balanced**

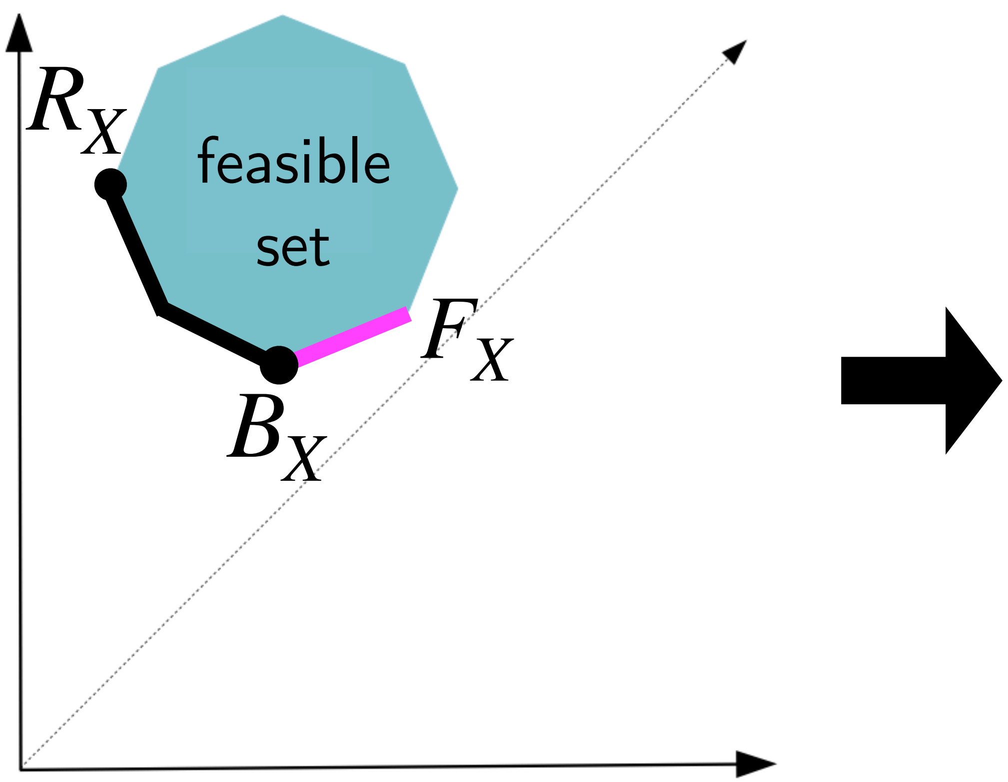


the fairness-accuracy frontier is precisely the usual pareto frontier

fairness considerations cannot justify the implementation of pareto-dominated outcomes

characterization of the frontier

X is **group-skewed**



the fairness-accuracy frontier is strictly larger than the pareto frontier

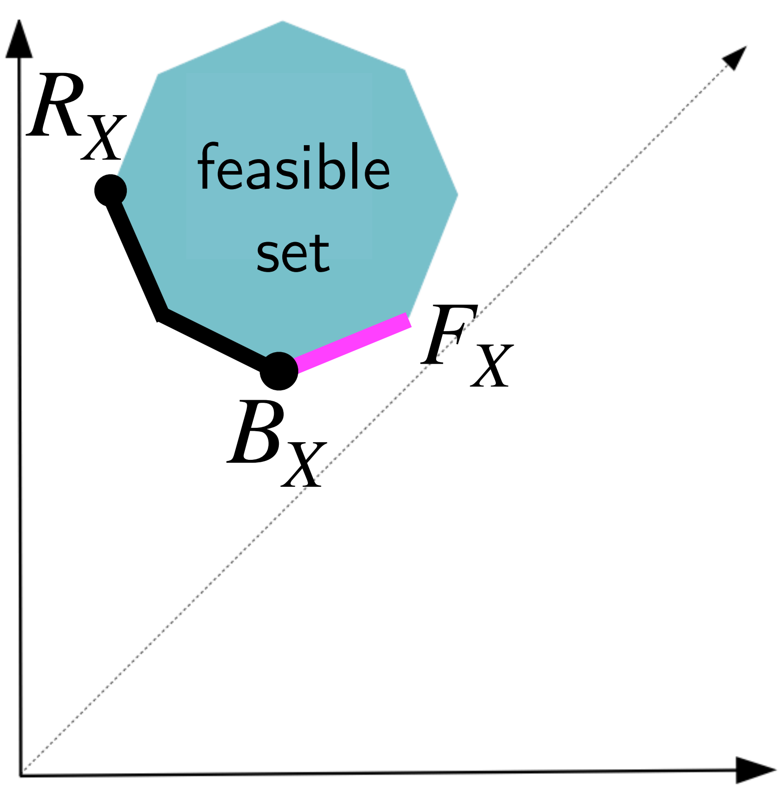
(includes pareto-dominated points)

characterization of the frontier

pareto-dominated outcomes may be optimal for the policymaker given sufficient weight on fairness concerns

(in practice, may look like choosing to ignore predictive information)

X is **group-skewed**



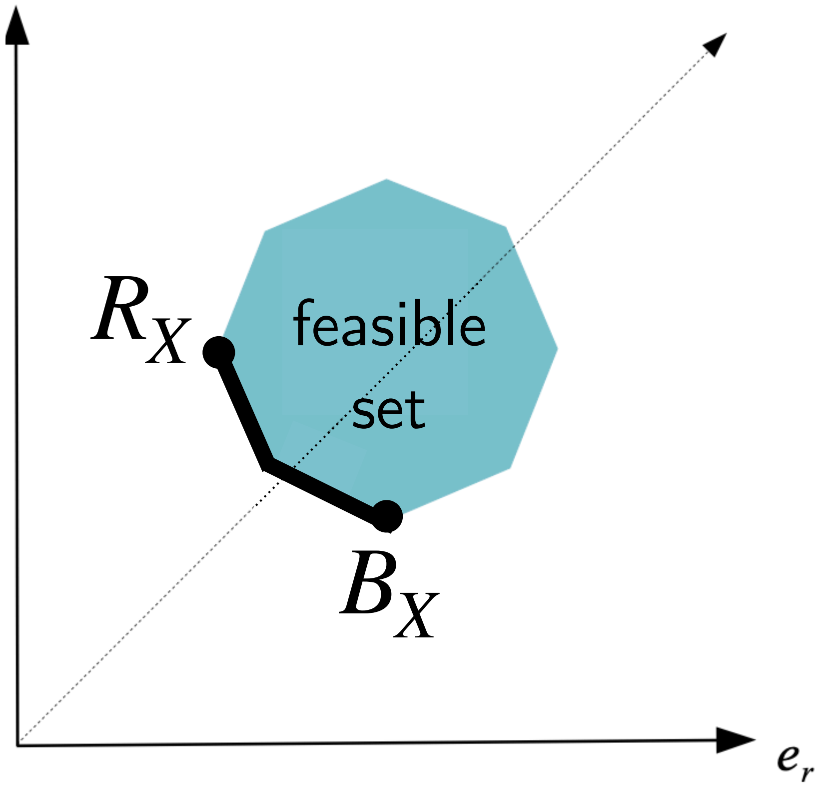
the fairness-accuracy frontier is strictly larger than the pareto frontier

(includes pareto-dominated points)

characterization of the frontier

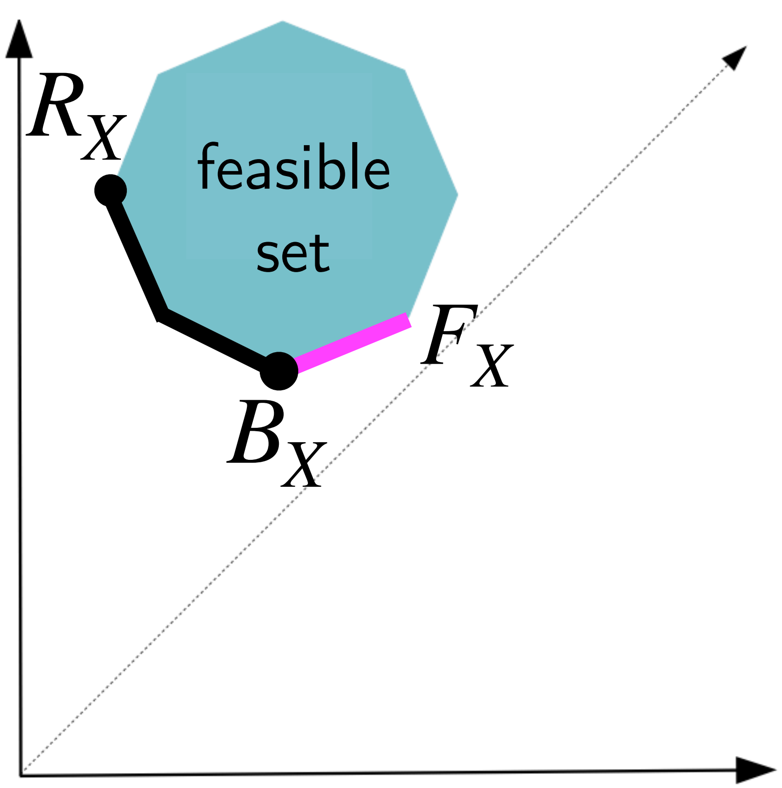
theorem:

X is **group-balanced**



the fairness-accuracy frontier is precisely the usual pareto frontier

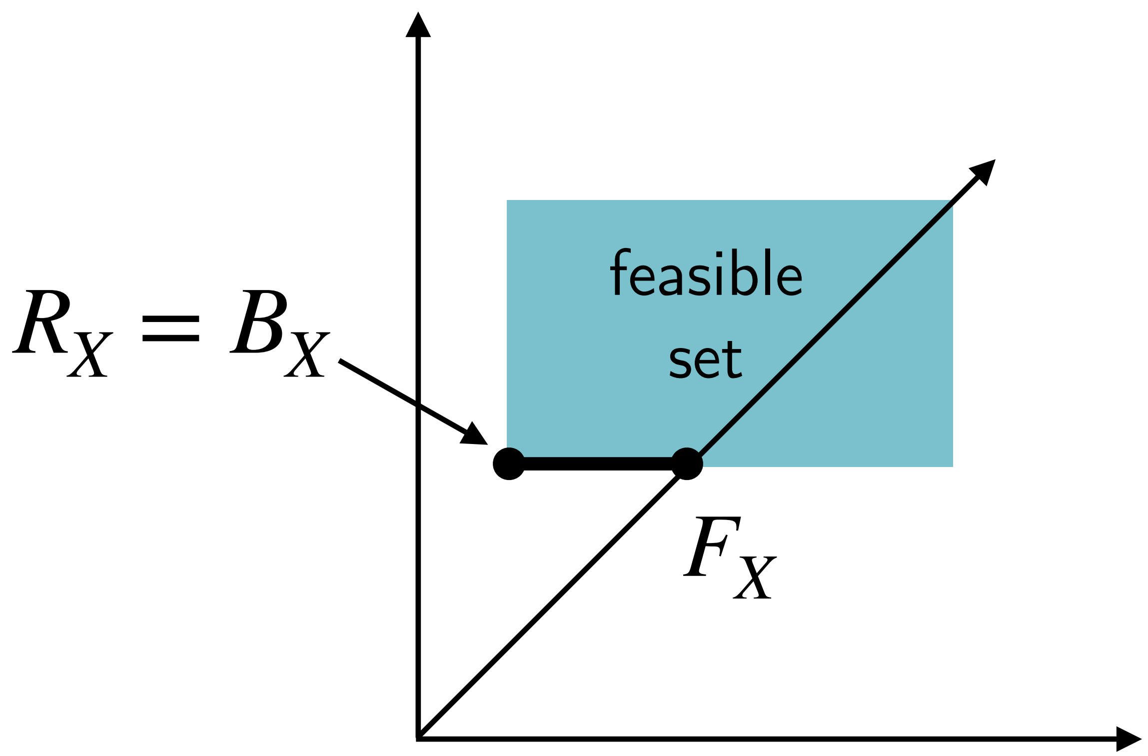
X is **group-skewed**



the fairness-accuracy frontier is strictly larger than the pareto frontier

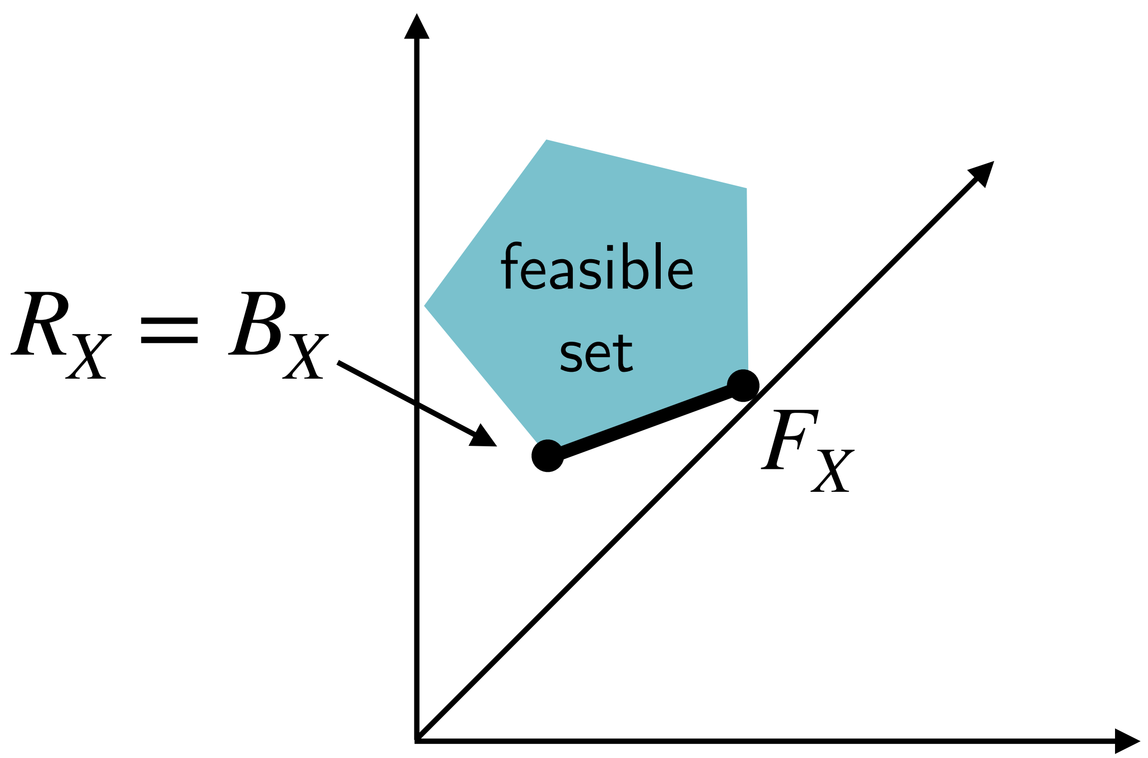
(includes pareto-dominated points)

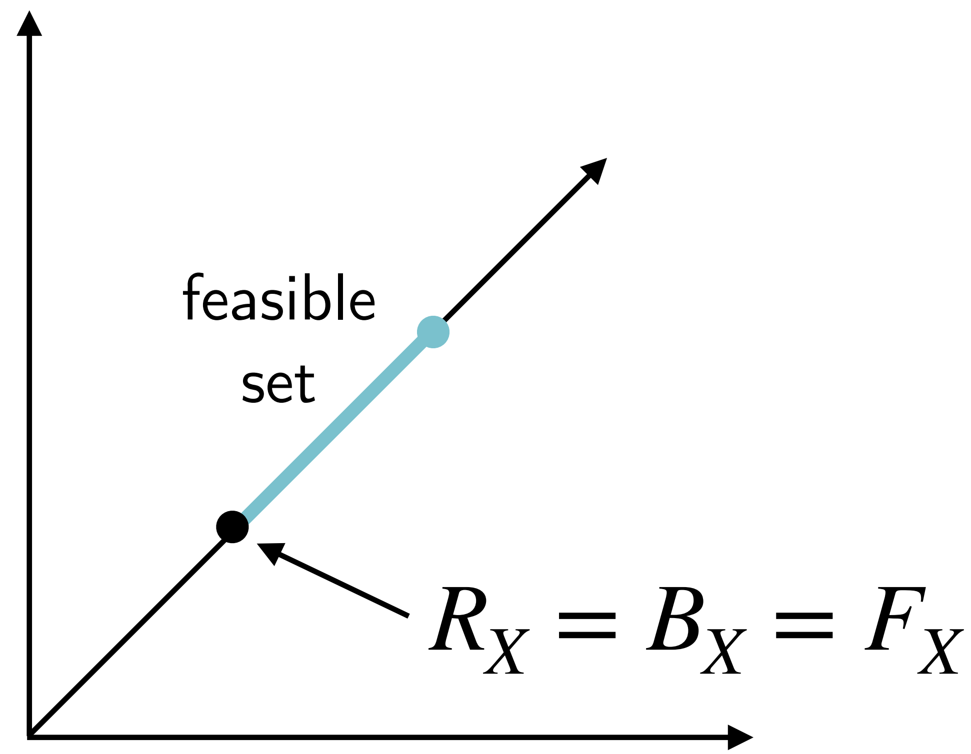
special case where G is a covariate



when G is a covariate, the feasible set and fairness-accuracy frontier simplify further

more special cases





conditional independence:

$$G \perp\!\!\!\perp Y \mid X$$

"once you know X, there is no additional predictive value to knowing G"

strong independence:

$$G \perp\!\!\!\perp (X, Y)$$

"the joint distribution of (X, Y) is the same for both groups"

interpreting group balance and group skew

why might X be **group-balanced**?

- X has a group-dependent meanings
  - high X implies high Y for group r, but low Y for group b

- different inputs in X are informative for either group
  - $X = (X_1, X_2)$ where X_1 is uninformative about Y for group r and X_2 is uninformative about Y for group b

interpreting group balance and group skew

why might X be **group-balanced**?

- X has a group-dependent meanings
  - high X implies high Y for group r, but low Y for group b
- different inputs in X are informative for either group
  - $X = (X_1, X_2)$ where X_1 is uninformative about Y for group r and X_2 is uninformative about Y for group b

why might X be **group-skewed**?

- X is asymmetrically informative
  - $Y \mid X, G = r$ more dispersed than $Y \mid X, G = b$
- e.g., medical data is recorded more accurately for high-income patients than low-income patients

two additional characterizations

see paper for two additional characterizations of the fairness-accuracy frontier

- a small set of preferences (which linearly trade off fairness and accuracy) is sufficient for recovering the full fairness-accuracy frontier
- a class of "threshold" algorithms implements the fairness-accuracy frontier

model 2:
input design

model 1: algorithm design

covariate vector

$(x_1, \ldots, x_n) \in \mathcal{X}$

a policymaker chooses an algorithm

$a : \mathcal{X} \to \Delta(\{0,1\})$

decision

$d \in \{0,1\}$

model 2: input design

covariate vector

$$(x_1, \ldots, x_n) \in \mathscr{X}$$

a policymaker
garbles the original
covariate vector

an agent chooses an algorithm

$$a : \widehat{\mathscr{X}} \to \Delta\left(\{0,1\}\right)$$

$$(\hat{x}_1, \ldots, \hat{x}_m) \in \widehat{\mathscr{X}}$$

garbled covariate vector

decision

$$d \in \{0,1\}$$

model 2: input design

LSAT score

law school administrators

coarsened LSAT score

law school admissions committee

decision

$d \in \{0,1\}$

e.g., in 1997, Berkeley law school administrators reported to their admissions committee only a coarsened LSAT score (Chan and Eyster, 2003)

garblings

original covariate vector

$$(x_1, \ldots, x_n)$$

a policymaker
chooses a garbling

$$(\hat{x}_1, \ldots, \hat{x}_m)$$

garbled covariate vector

a garbling of X is any stochastic map

$$T : \mathscr{X} \to \Delta(\widehat{\mathscr{X}})$$

garblings

original covariate vector

$(x_1, \ldots, x_n)$

a policymaker
chooses a garbling

$(\hat{x}_1, \ldots, \hat{x}_m)$

garbled covariate vector

a garbling of X is any stochastic map

$$T : \mathscr{X} \to \Delta(\widehat{\mathscr{X}})$$

original space of
covariate vectors

garblings

original covariate vector

$(x_1, \ldots, x_n)$

a policymaker chooses a garbling

$(\hat{x}_1, \ldots, \hat{x}_m)$

garbled covariate vector

a garbling of X is any stochastic map

$$T : \mathscr{X} \to \Delta(\widehat{\mathscr{X}})$$

new space of covariate vectors

garblings

$$(x_1, x_2, x_3)$$

$$\Big\downarrow \text{a policymaker chooses a garbling}$$

$$(x_1, x_2)$$

examples

- ban a specific covariate (e.g., a group identity or test score)

garblings

$$x \in \{1,2,3,4\}$$

a policymaker chooses a garbling

$$\downarrow$$

$$\hat{x} = \begin{cases} L & \text{if } x \in \{1,2\} \\ H & \text{if } x \in \{3,4\} \end{cases}$$

examples

- ban a specific covariate (e.g., a group identity or test score)
- coarsen a covariate

garblings

$$(x_1, x_2)$$

a policymaker
chooses a garbling

$$(x_1 + \varepsilon, x_2)$$

$$\varepsilon \sim \mathcal{N}(0,1)$$

examples

- ban a specific covariate (e.g., a group identity or test score)
- coarsen a covariate
- add noise to a covariate

preferences

covariate vector

$(x_1, \ldots, x_n) \in \mathscr{X}$

a policymaker chooses a garbling

can have any fairness-accuracy preference

$(\hat{x}_1, \ldots, \hat{x}_m) \in \widehat{\mathscr{X}}$

garbled covariate vector

an agent chooses an algorithm

utilitarian

$d \in \{0,1\}$

decision (e.g., whether to treat)

input design

covariate vector

$(x_1, \ldots, x_n) \in \mathscr{X}$

a policymaker
chooses a garbling

can have any
fairness-accuracy
preference

INPUT DESIGN

an agent chooses an
algorithm

$(\hat{x}_1, \ldots, \hat{x}_m) \in \widehat{\mathscr{X}} \longrightarrow d \in \{0,1\}$

garbled covariate vector

utilitarian

decision
(e.g., whether
to treat)

model 1: algorithm design

covariate vector

$(x_1, \ldots, x_n) \in \mathscr{X}$

a policymaker chooses an algorithm

$a : \mathscr{X} \to \Delta(\{0,1\})$

decision

$d \in \{0,1\}$

input design

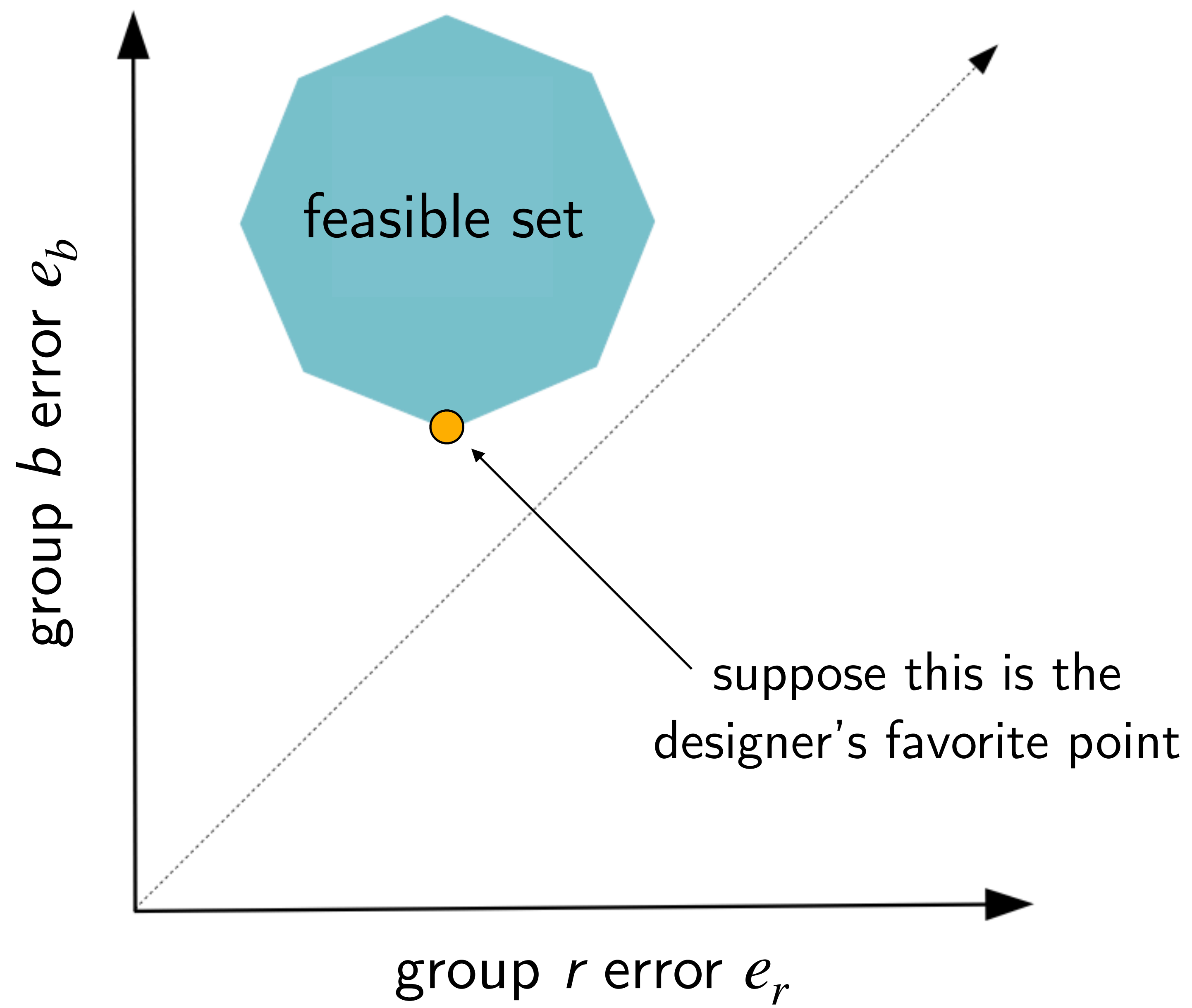
for any garbling T, let $a_T : \widehat{X} \to \{0,1\}$ denote the algorithm that a utilitarian agent optimally chooses given this garbling

definition: the input design feasible set given X is

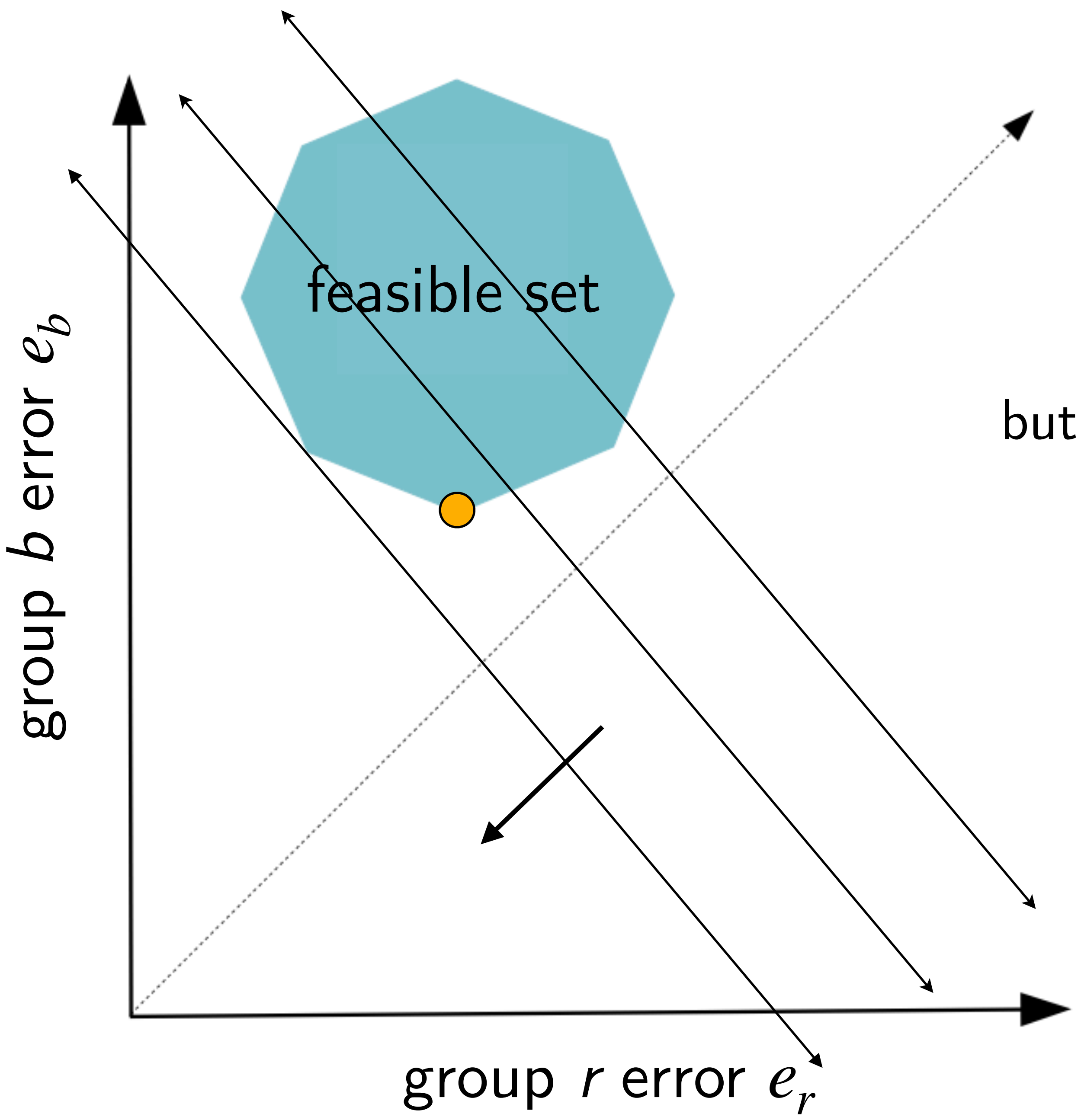
$$\mathcal{E}^*_X = \{e(a_T) \mid T \text{ is a garbling of } X\}$$

definition: the input design fairness-accuracy frontier given X (denoted $\mathcal{F}^*_X$) is the set of $>_{FA}$-undominated points in $\mathcal{E}^*_X$

how input design can help



how input design can help



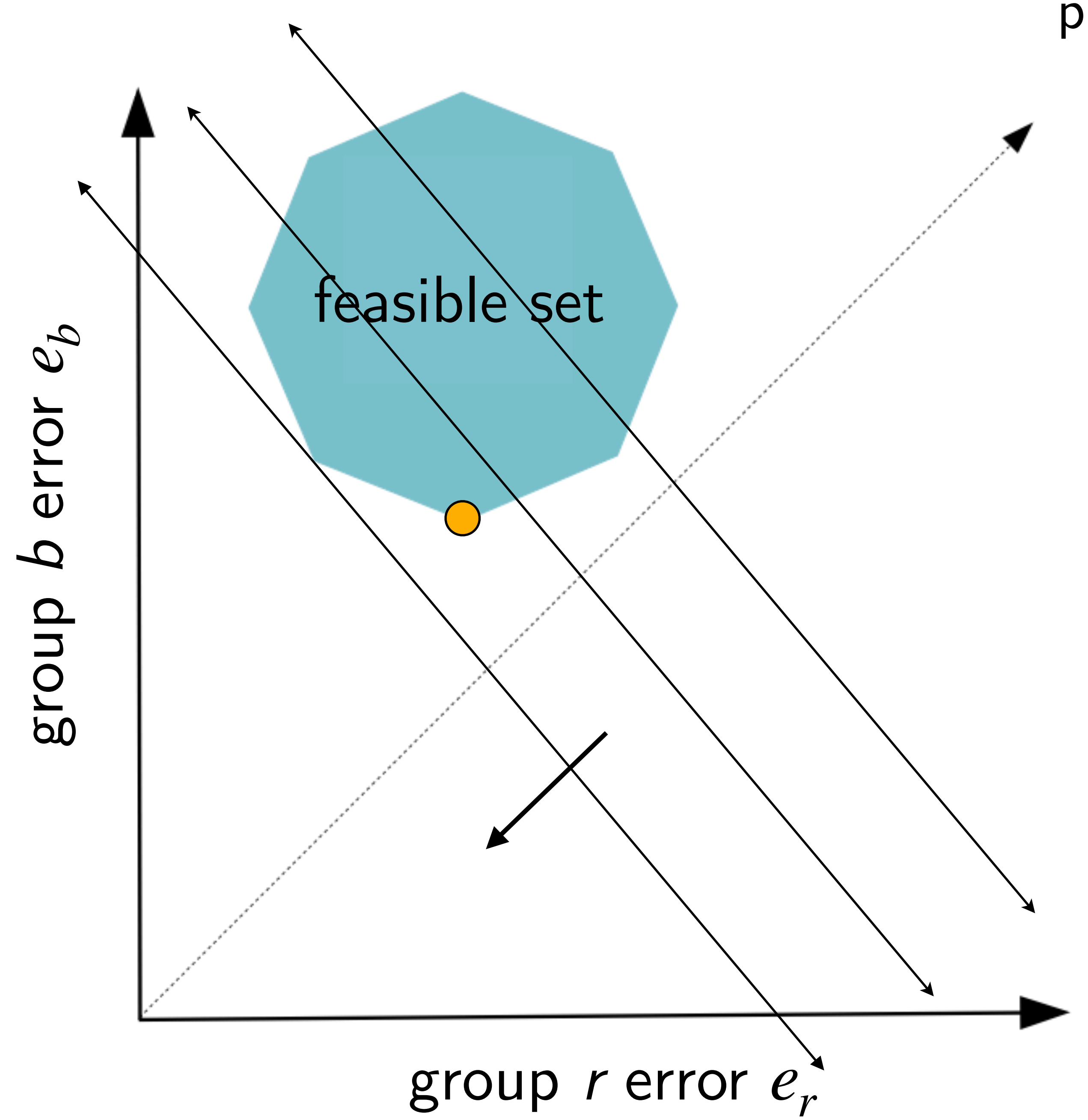
but these are the utilitarian indifference curves

how input design can help

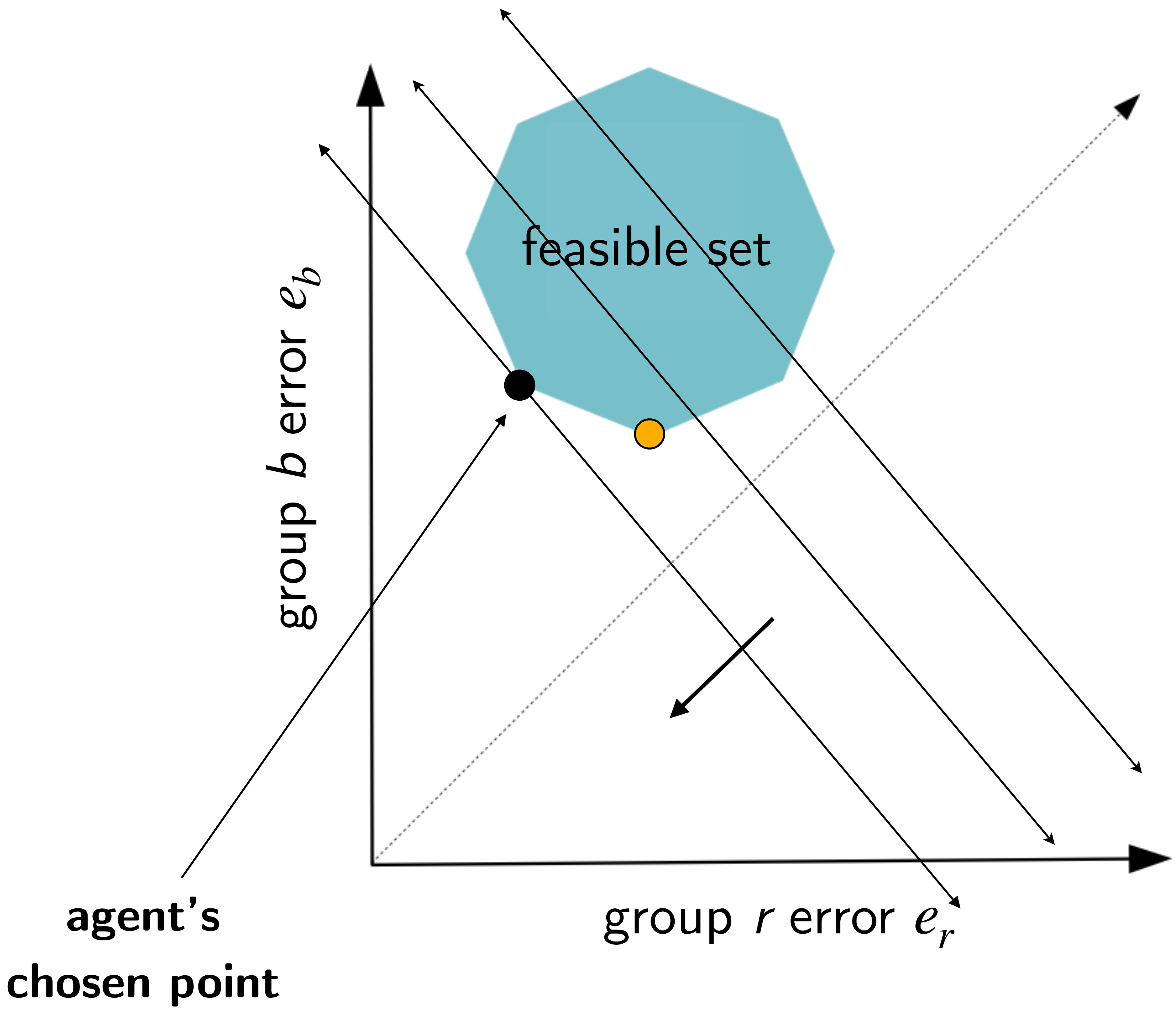
send X ungarbled

↓

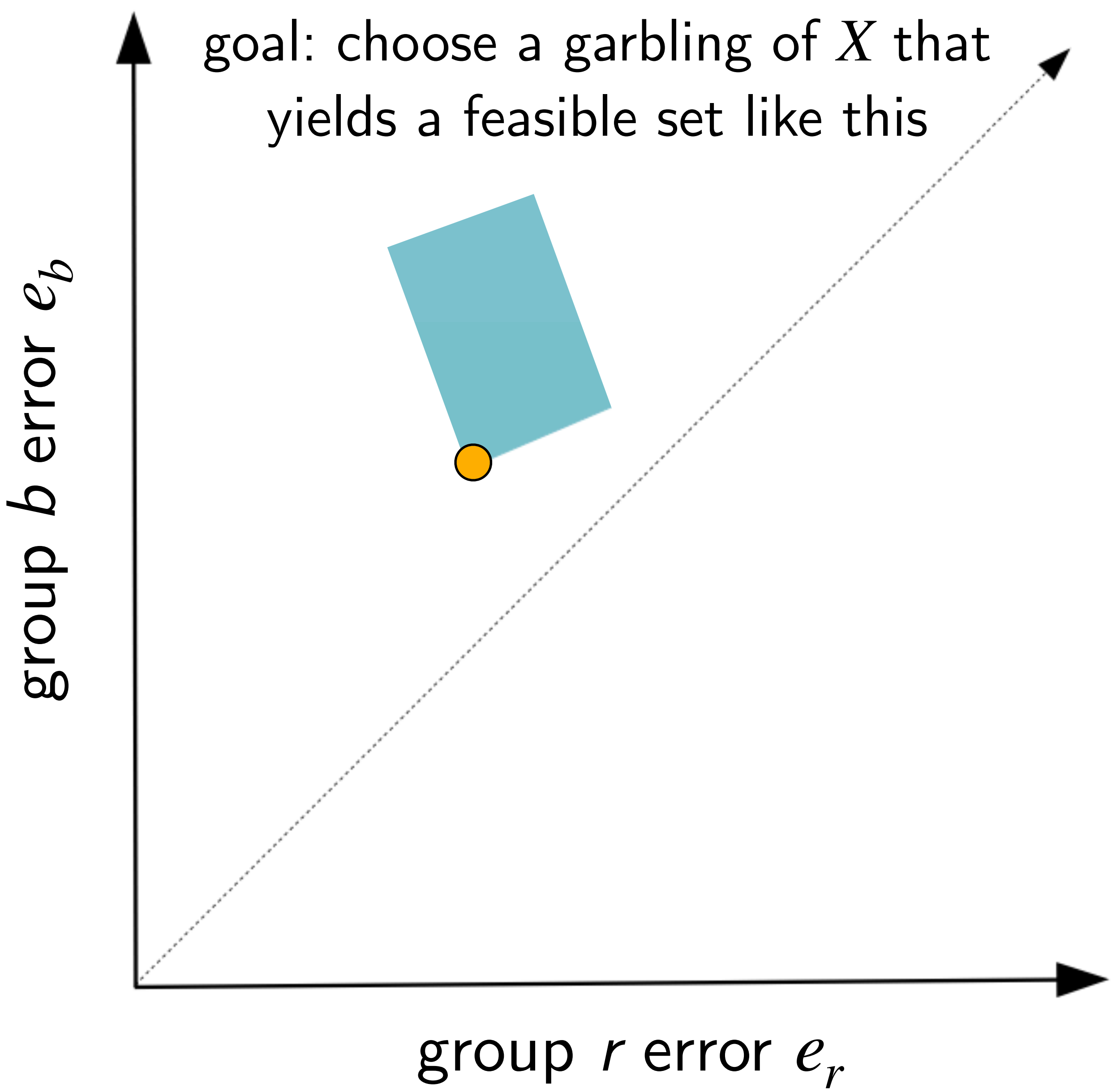
agent can implement any point in the feasible set $\mathscr{E}_X$



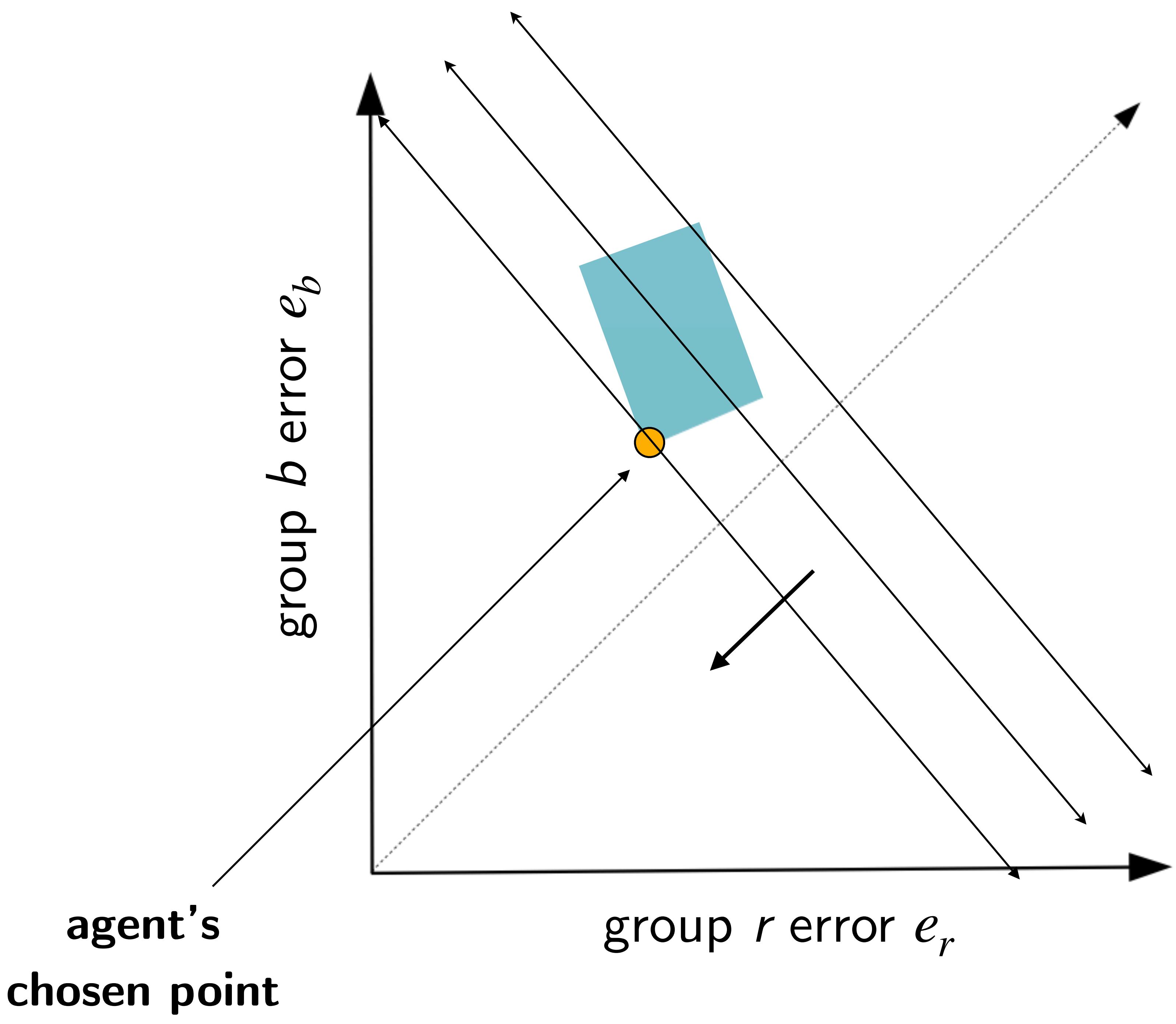
how input design can help



how input design can help

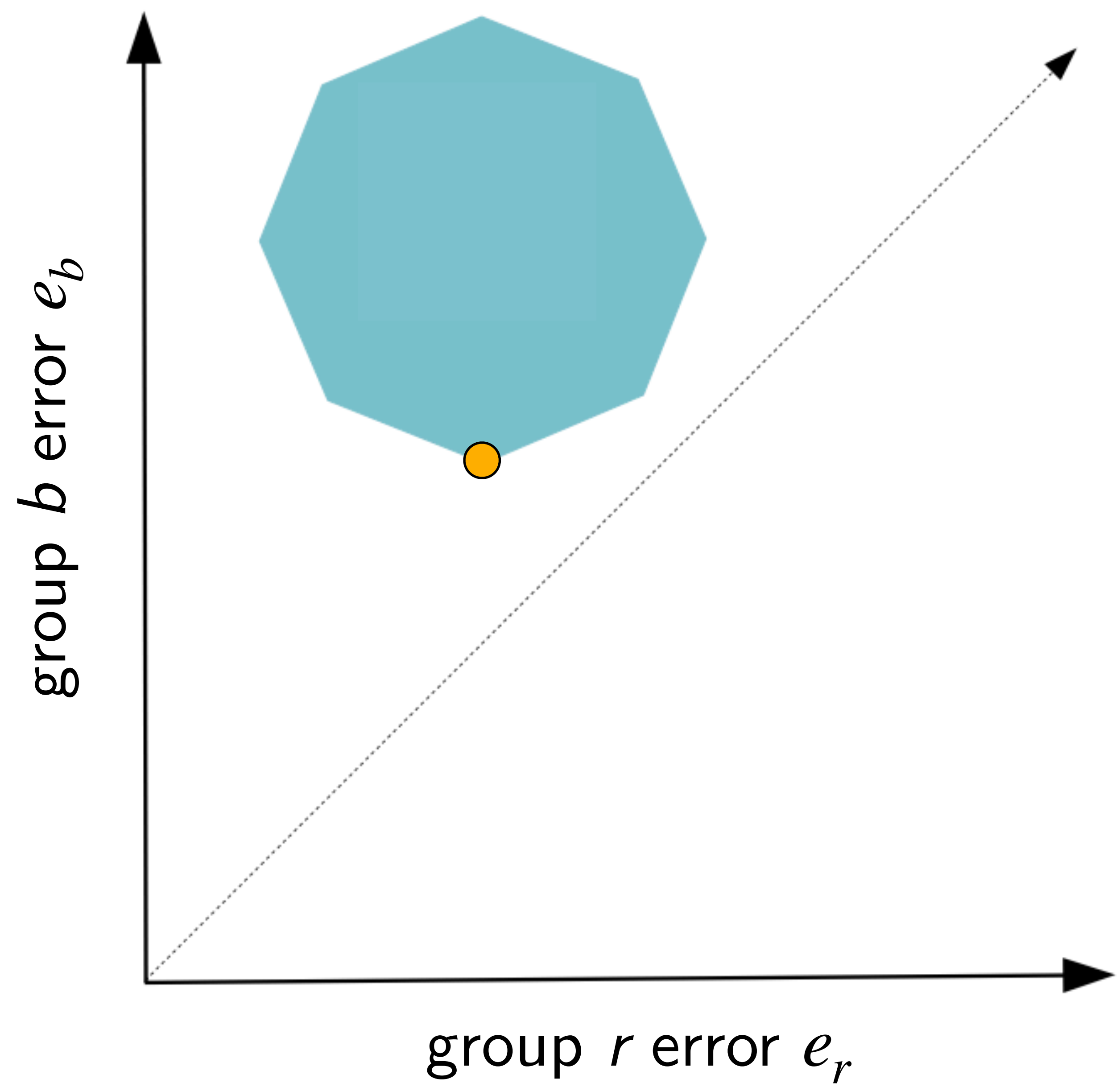


how input design can help



how input design can help

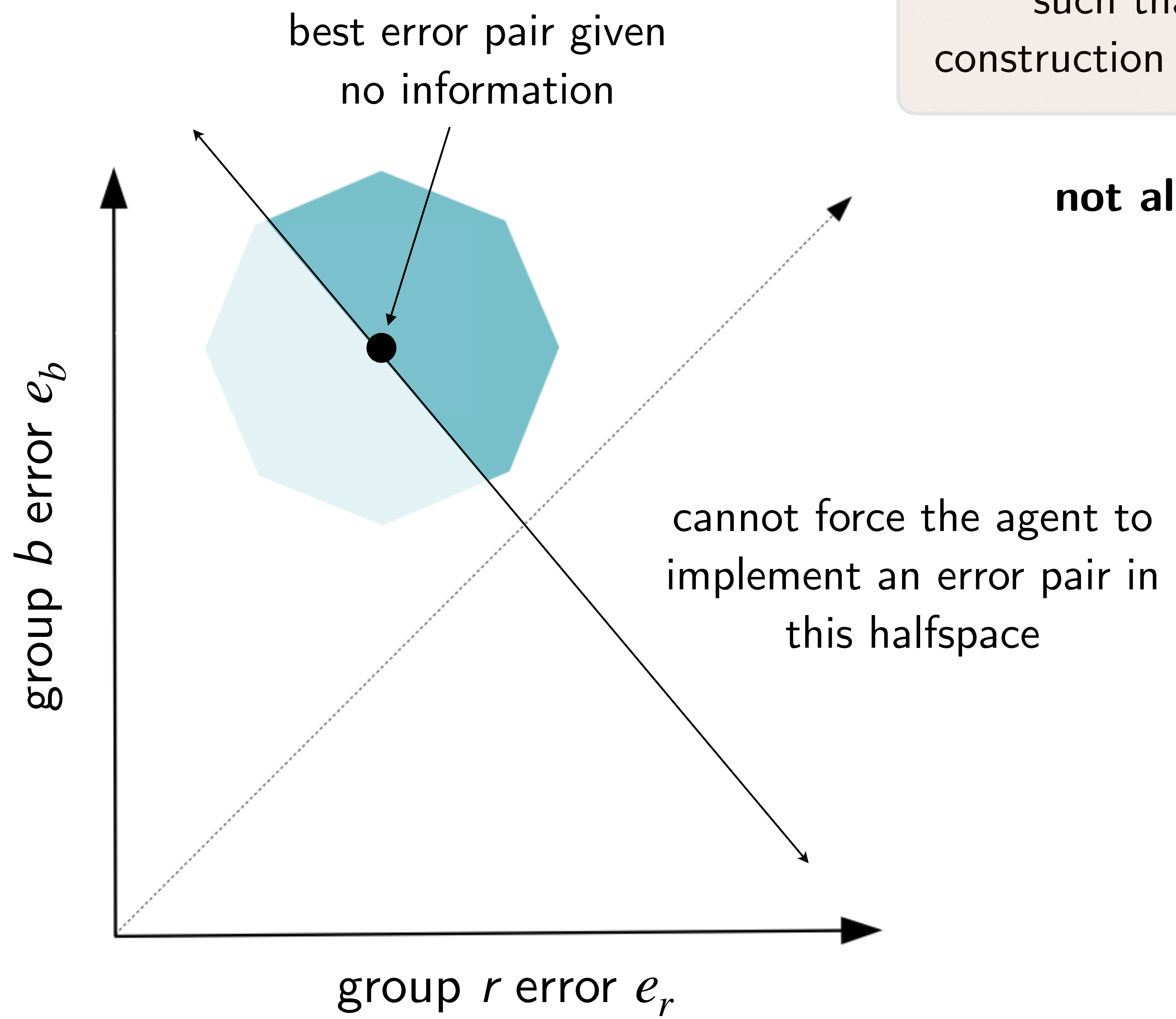
when do garbling exist such that this construction is possible?



how input design can help

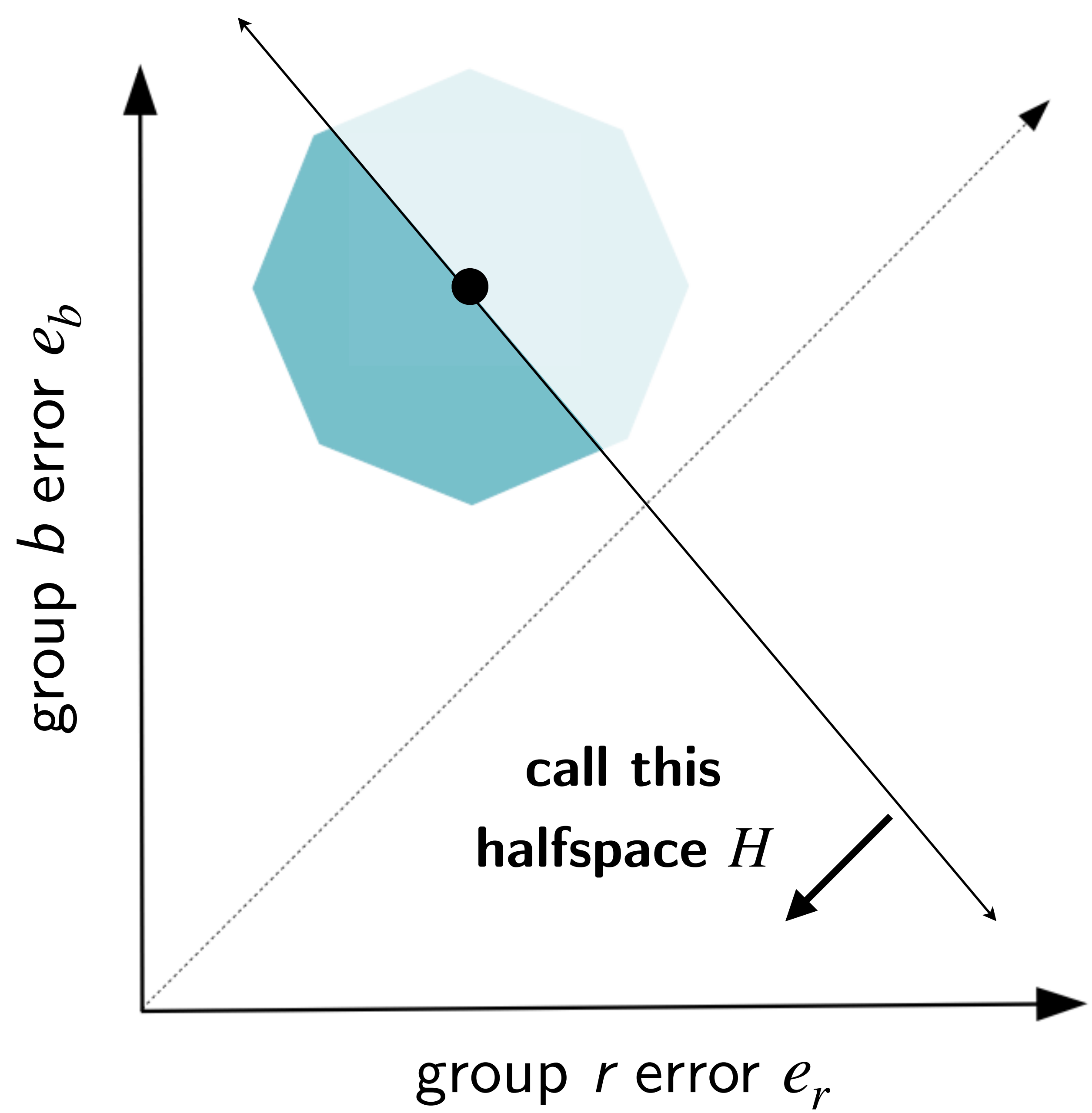
when do garbling exist such that this construction is possible?

not always



how input design can help

when do garbling exist such that this construction is possible?

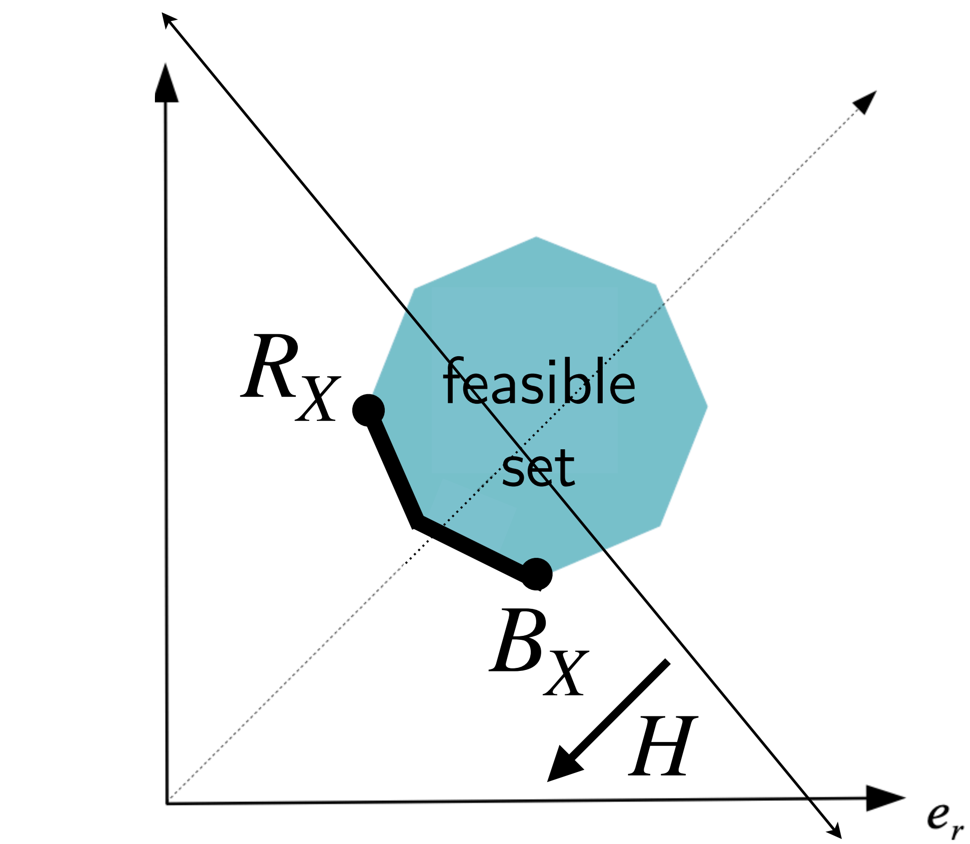


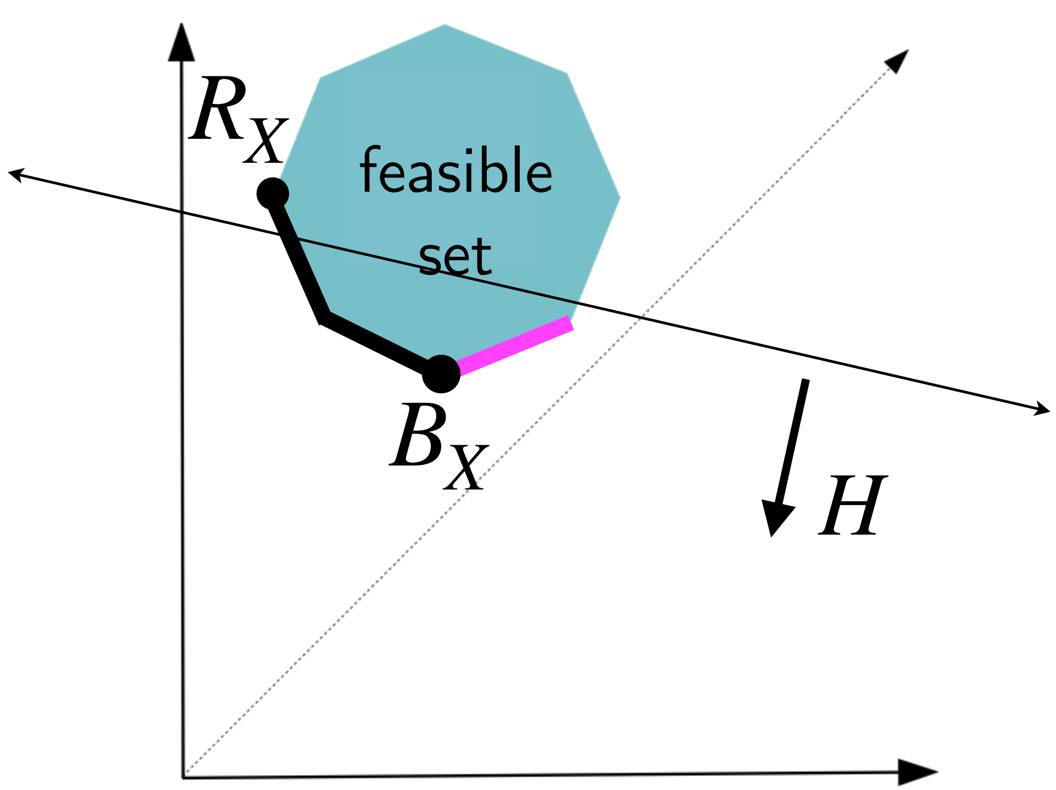
lemma:

$$\mathscr{E}_X^* = \mathscr{E}_X \cap H$$

(see also Alonso and Camara, 2016)

how powerful is input design?





proposition:

(a) if X is group-balanced, then

$$\mathscr{F}_X = \mathscr{F}_X^* \quad \Longleftrightarrow \quad R_X, B_X \in H$$

(b) if X is r-skewed, then

$$\mathscr{F}_X = \mathscr{F}_X^* \quad \Longleftrightarrow \quad R_X, F_X \in H$$

could banning a covariate ever be strictly optimal?

when the policymaker has control of the algorithm (model 1), it is **never** strictly optimal to ban a covariate

- Blackwell (1951)

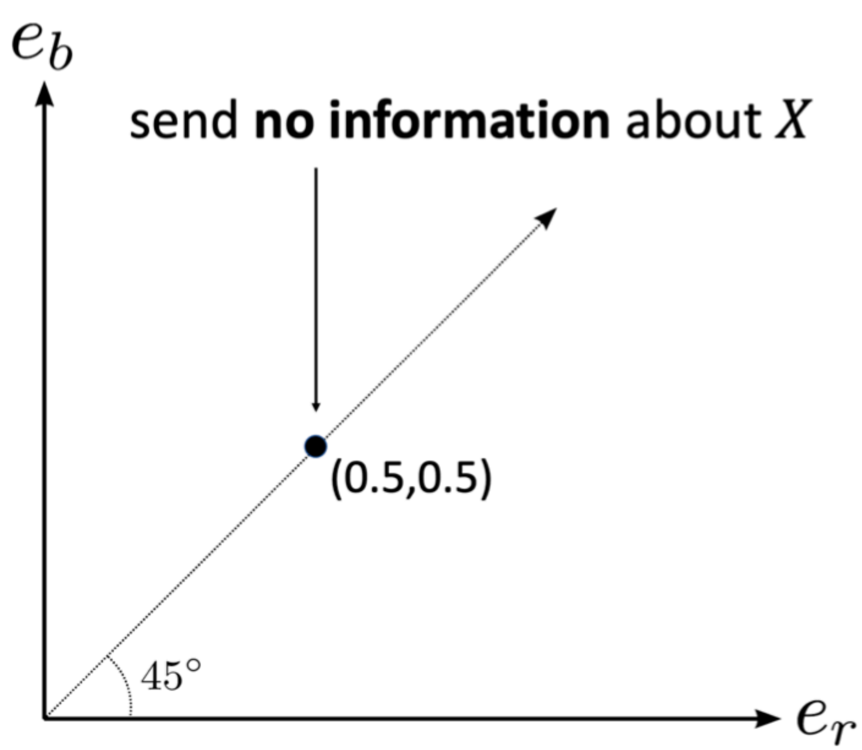
because of misaligned preferences between the policymaker and agent, banning a covariate **can be strictly optimal** in our framework

simple example where banning a covariate is optimal

- $Y \in \{0,1\}$ with $P(Y = 1 \mid G = g) = 1/2$ for both groups g
- $X \in \{0,1\}$ is a binary covariate
  - $X = Y$ with probability 1 if $G = r$
  - $X = Y$ with probability 0.6 if $G = b$
- the policymaker is Egalitarian (payoff is $-|e_r - e_b|$)

simple example where banning a covariate is optimal

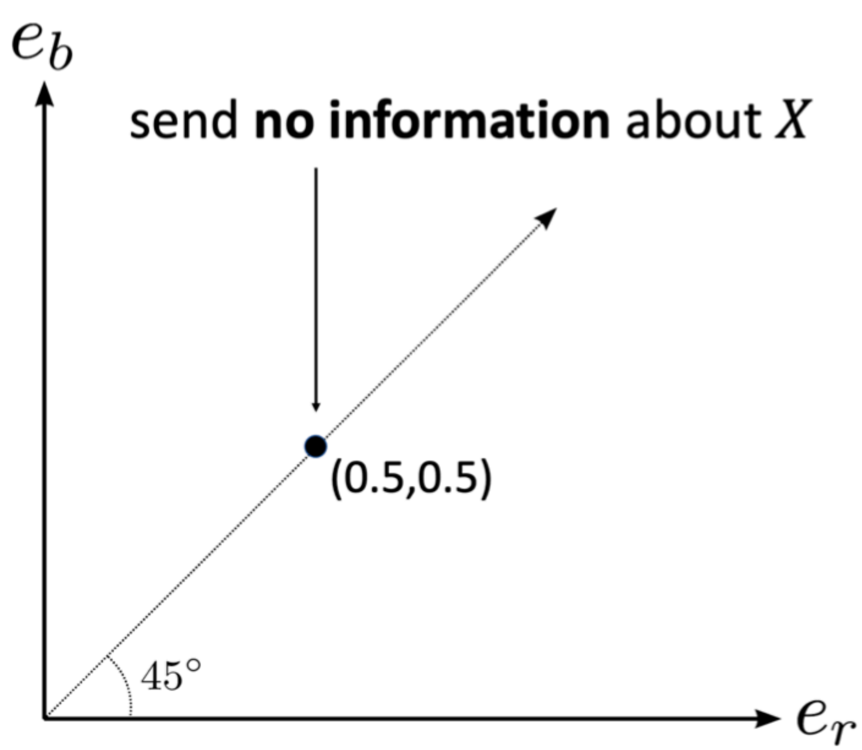
- $Y \in \{0,1\}$ with $P(Y = 1 \mid G = g) = 1/2$ for both groups g
- $X \in \{0,1\}$ is a binary covariate
  - $X = Y$ with probability 1 if $G = r$
  - $X = Y$ with probability 0.6 if $G = b$
- the policymaker is Egalitarian (payoff is $-|e_r - e_b|$)



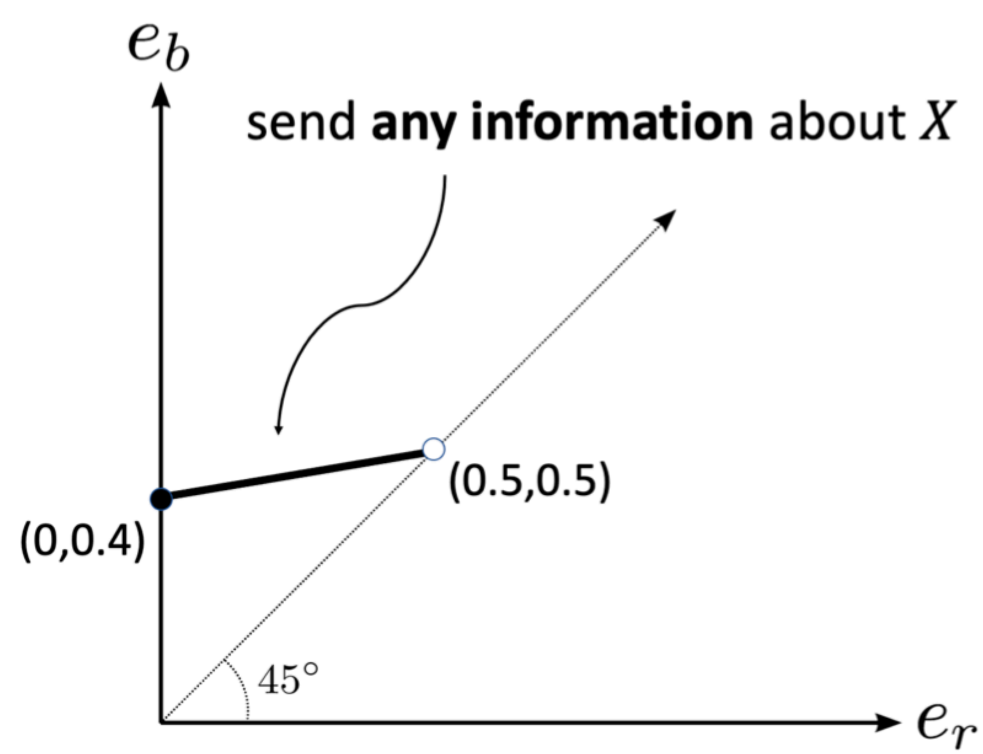
the policymaker's payoff is zero

simple example where banning a covariate is optimal

- $Y \in \{0,1\}$ with $P(Y = 1 \mid G = g) = 1/2$ for both groups g
- $X \in \{0,1\}$ is a binary covariate
  - $X = Y$ with probability 1 if $G = r$
  - $X = Y$ with probability 0.6 if $G = b$
- the policymaker is Egalitarian (payoff is $-|e_r - e_b|$)



the policymaker's payoff is zero



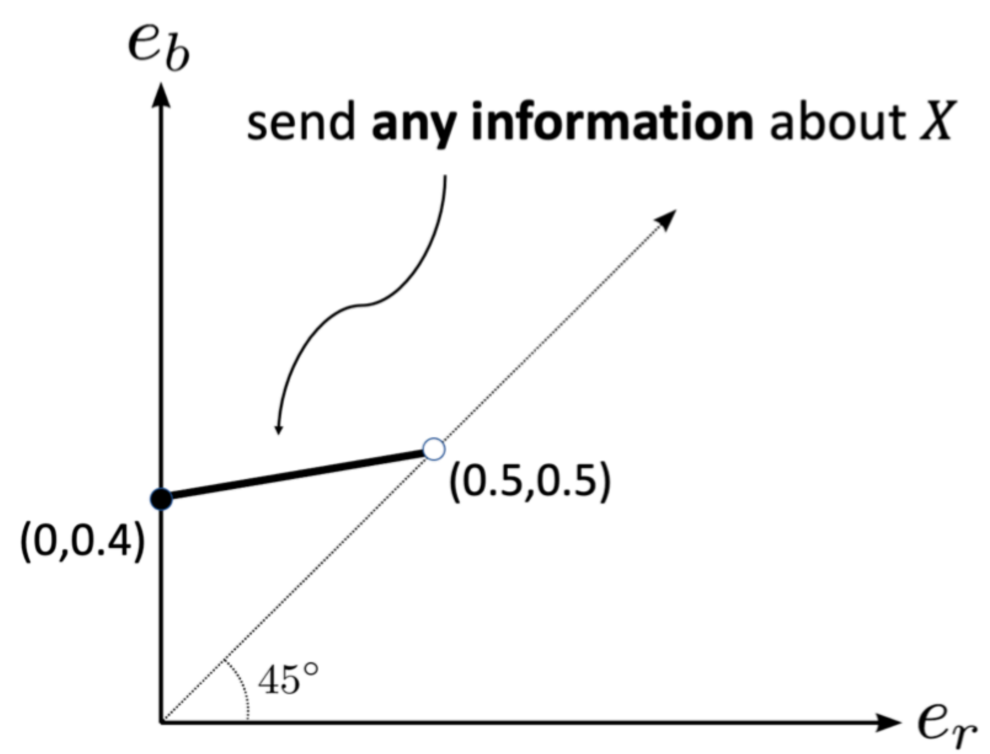
the policymaker's payoff is strictly negative

simple example where banning a covariate is optimal

- $Y \in \{0,1\}$ with $P(Y = 1 \mid G = g) = 1/2$ for both groups g
- $X \in \{0,1\}$ is a binary covariate
  - $X = Y$ with probability 1 if $G = r$
  - $X = Y$ with probability 0.6 if $G = b$
- the policymaker is Egalitarian (payoff is $-|e_r - e_b|$)

intuition:

the utilitarian agent will use all permitted information to make more accurate decisions, but accuracy increases faster for group r than group b



the policymaker's payoff is strictly negative

could banning a covariate ever be strictly optimal?

when the policymaker has control of the algorithm (model 1), it is **never** strictly optimal to ban a covariate

- Blackwell (1951)

because of misaligned preferences between the policymaker and agent, banning a covariate **can be strictly optimal** in our framework

could banning a covariate ever be strictly optimal?

when the policymaker has control of the algorithm (model 1), it is **never** strictly optimal to ban a covariate

- Blackwell (1951)

because of misaligned preferences between the policymaker and agent, banning a covariate **can be strictly optimal** in our framework

(**result)** ...but this only possible when group identity is not available

could banning a covariate ever be strictly optimal?

when the policymaker has control of the algorithm (model 1), it is **never** strictly optimal to ban a covariate

- Blackwell (1951)

because of misaligned preferences between the policymaker and agent, banning a covariate **can be strictly optimal** in our framework

(**result)** ...but this only possible when group identity is not available

definition: write $\mathscr{F}_{X,X'} >_{FA} \mathscr{F}_X$ if every $e \in \mathscr{F}_X$ is FA-dominated by some $e \in \mathscr{F}_{X,X'}$

- every designer with a FA preference is made strictly better off by garbling (X, X') rather than by garbling X alone

preferences

original covariate vector

$$(\underbrace{x_1, \ldots, x_{n-1}}_{\text{GPA, essays, etc.}}, x_n)$$

test score

a policymaker chooses a garbling

?

a **utilitarian** agent chooses an algorithm

$d \in \{0,1\}$

decision

could be optimal to drop x_n entirely
(for some policymaker preference)

preferences

group identity

original covariate vector

$$(\underbrace{x_1, \ldots, x_{n-1},}_{\text{GPA, essays, etc.}} x_n, g)$$

test score

a policymaker
chooses a garbling

?

never optimal to drop x_n entirely
(for **any** policymaker preference)

a **utilitarian** agent
chooses an algorithm

$d \in \{0,1\}$
decision

could banning a covariate ever be strictly optimal?

result: $\mathscr{F}_{X,X',G} >_{FA} \mathscr{F}_{X,G}$ for all "minimally informative" X'

i.e., every policymaker (with any fairness-accuracy preference) is made strictly worse off by banning any (minimally informative) covariate **when group identity g is available**

could banning a covariate ever be strictly optimal?

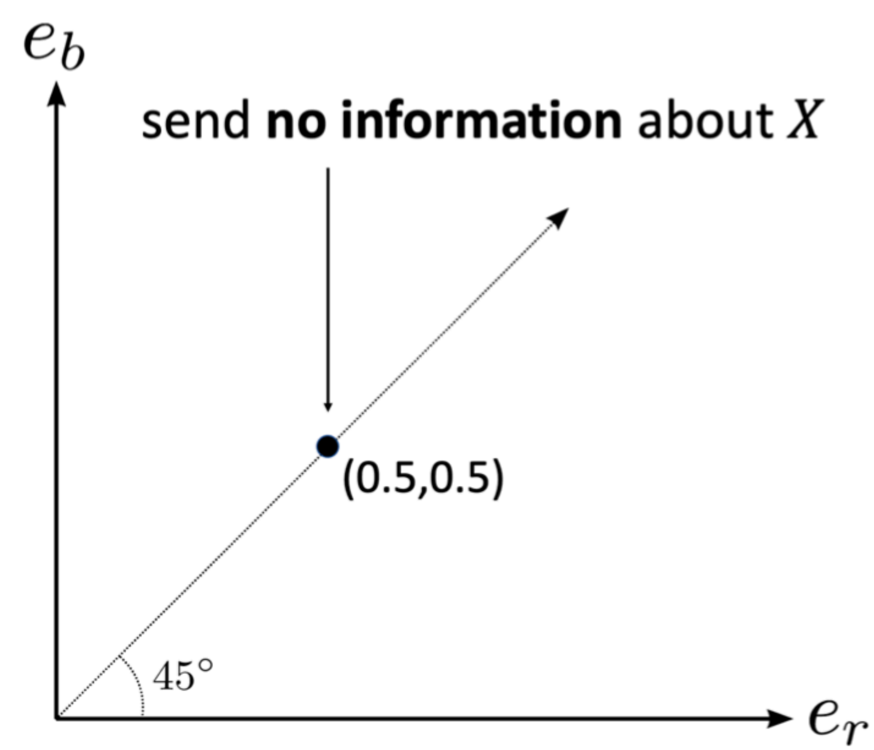
result: $\mathscr{F}_{X,X',G} >_{FA} \mathscr{F}_{X,G}$ for all "minimally informative" X'

i.e., every policymaker (with any fairness-accuracy preference) is made strictly worse off by banning any (minimally informative) covariate **when group identity g is available**

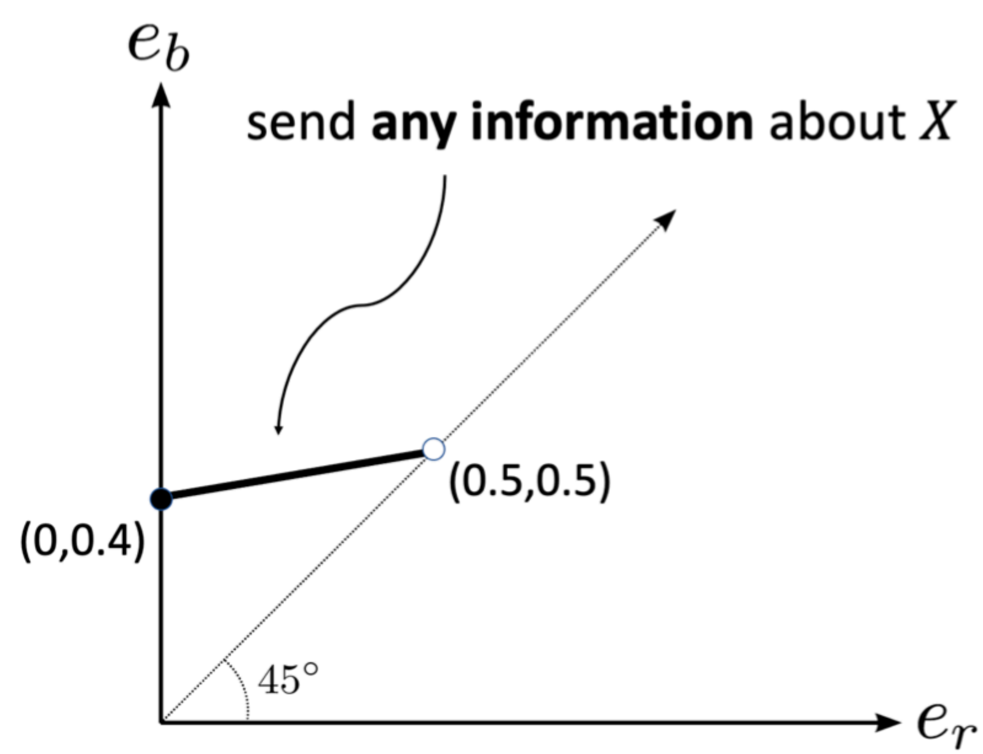
intuition: when g is available, can choose a group-dependent garbling of the covariate, e.g., add noise if $g = r$ but not if $g = b$

back to the example

- $Y \in \{0,1\}$ with $P(Y = 1 \mid G = g) = 1/2$ for both groups g
- $X \in \{0,1\}$ is a binary covariate
  - $X = Y$ with probability 1 if $G = r$
  - $X = Y$ with probability 0.6 if $G = b$
- the policymaker is Egalitarian (payoff is $-|e_r - e_b|$)



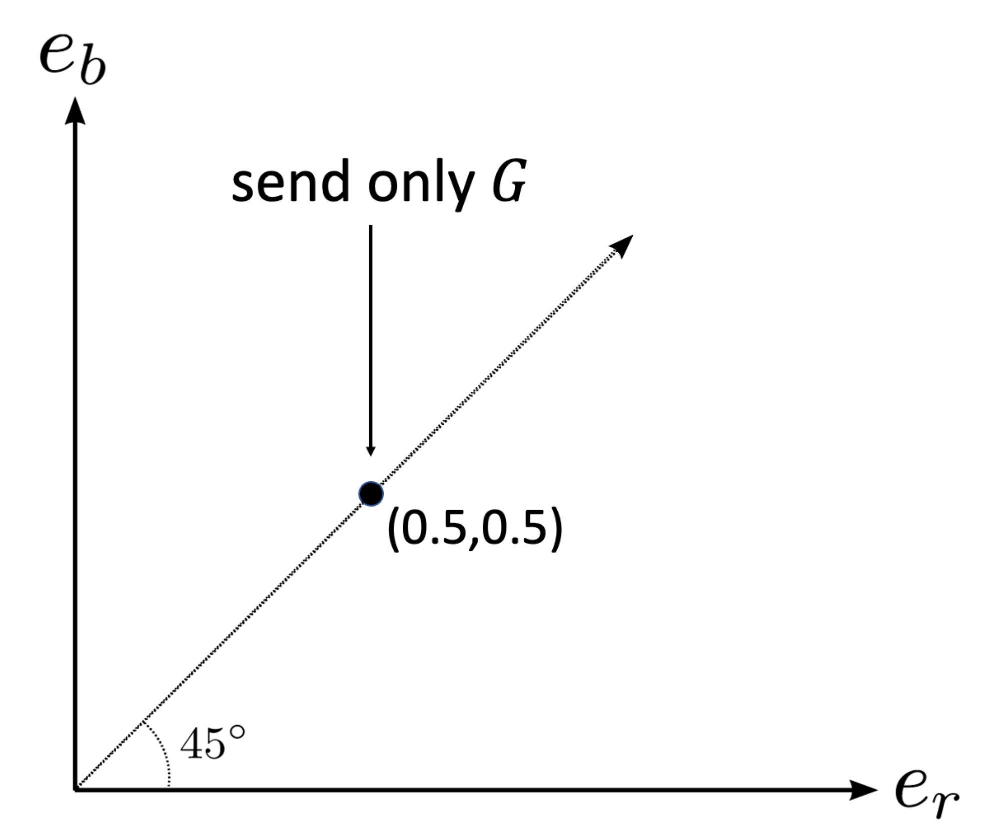
the policymaker's payoff is zero



the policymaker's payoff is strictly negative

back to the example

- $Y \in \{0,1\}$ with $P(Y = 1 \mid G = g) = 1/2$ for both groups g
- $X \in \{0,1\}$ is a binary covariate
  - $X = Y$ with probability 1 if $G = r$
  - $X = Y$ with probability 0.6 if $G = b$
- the policymaker is Egalitarian (payoff is $-|e_r - e_b|$)

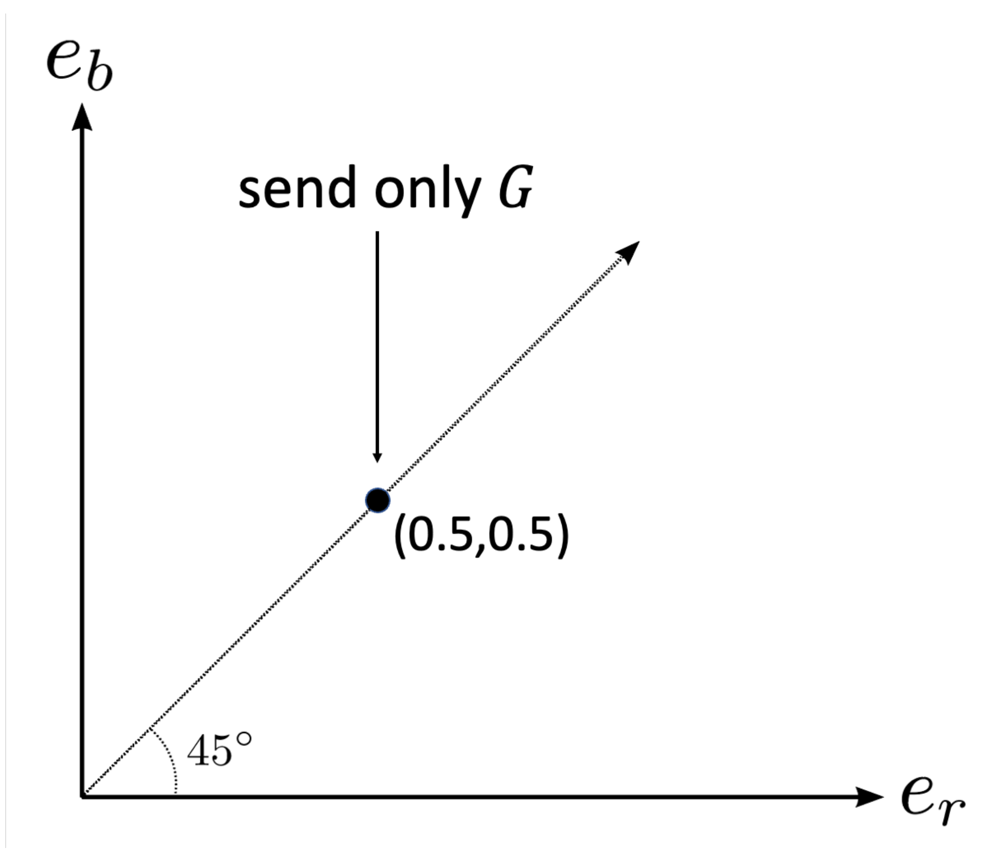


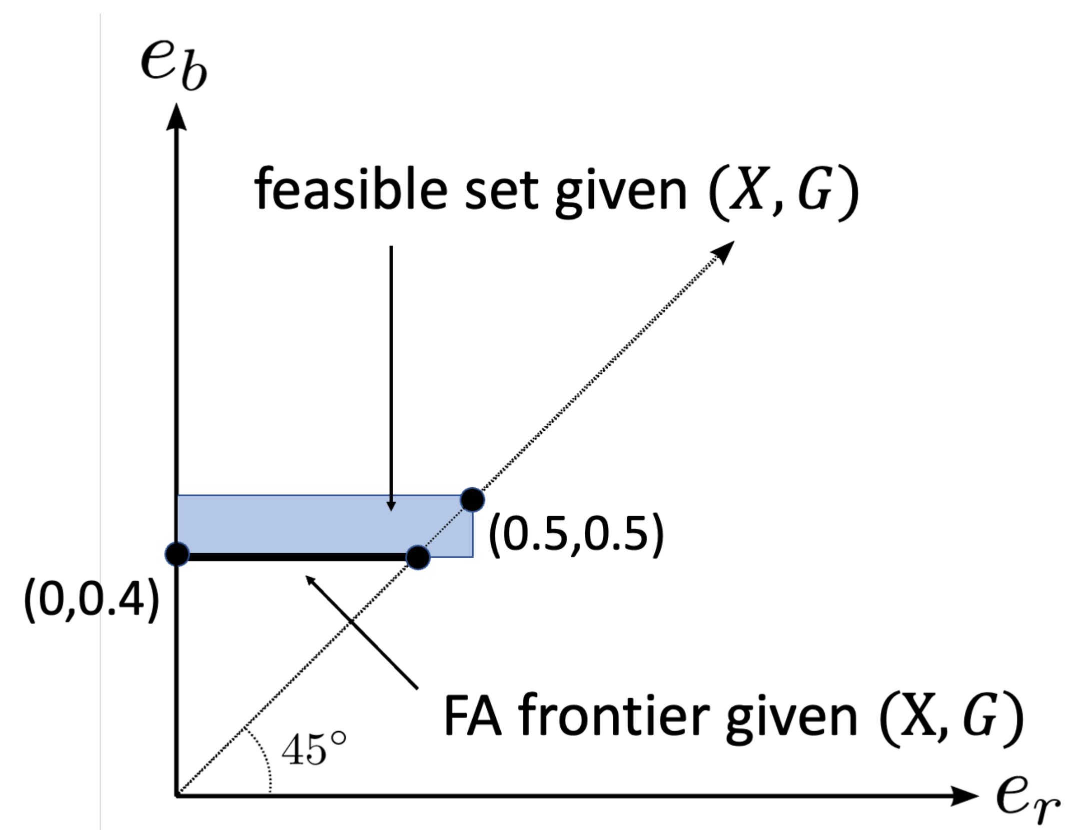
now suppose the policymaker additionally has access to G

so the full covariate vector is (X, G)

back to the example

- $Y \in \{0,1\}$ with $P(Y = 1 \mid G = g) = 1/2$ for both groups g
- $X \in \{0,1\}$ is a binary covariate
  - $X = Y$ with probability 1 if $G = r$
  - $X = Y$ with probability 0.6 if $G = b$
- the policymaker is Egalitarian (payoff is $-|e_r - e_b|$)





comment on test scores

considering test scores, our result says that...

- if g is available, then excluding test scores is welfare-reducing for all policymakers with the ability to garble available covariates
- if g is not available, then it may be better for a sufficiently fairness-minded policymaker to completely exclude test scores

banning affirmative action may lead universities with certain preferences to ban use of test scores

empirical application

taking the framework to data

- have so far focused on general conceptual findings that hold across settings
- our framework can also be used to better understand the fairness-accuracy tradeoffs in specific datasets
- illustrate this on a healthcare dataset (see paper for second illustration)

healthcare application

the data is from Obermeyer et al. (2019)

- 48,784 patient observations
- the covariate vector X includes 139 demographic + medical covariates
- group identities: Black or White, denoted $G \in \{b, w\}$
- true health needs are measured in the data as each patient's total number of active chronic illnesses in the subsequent year

healthcare application

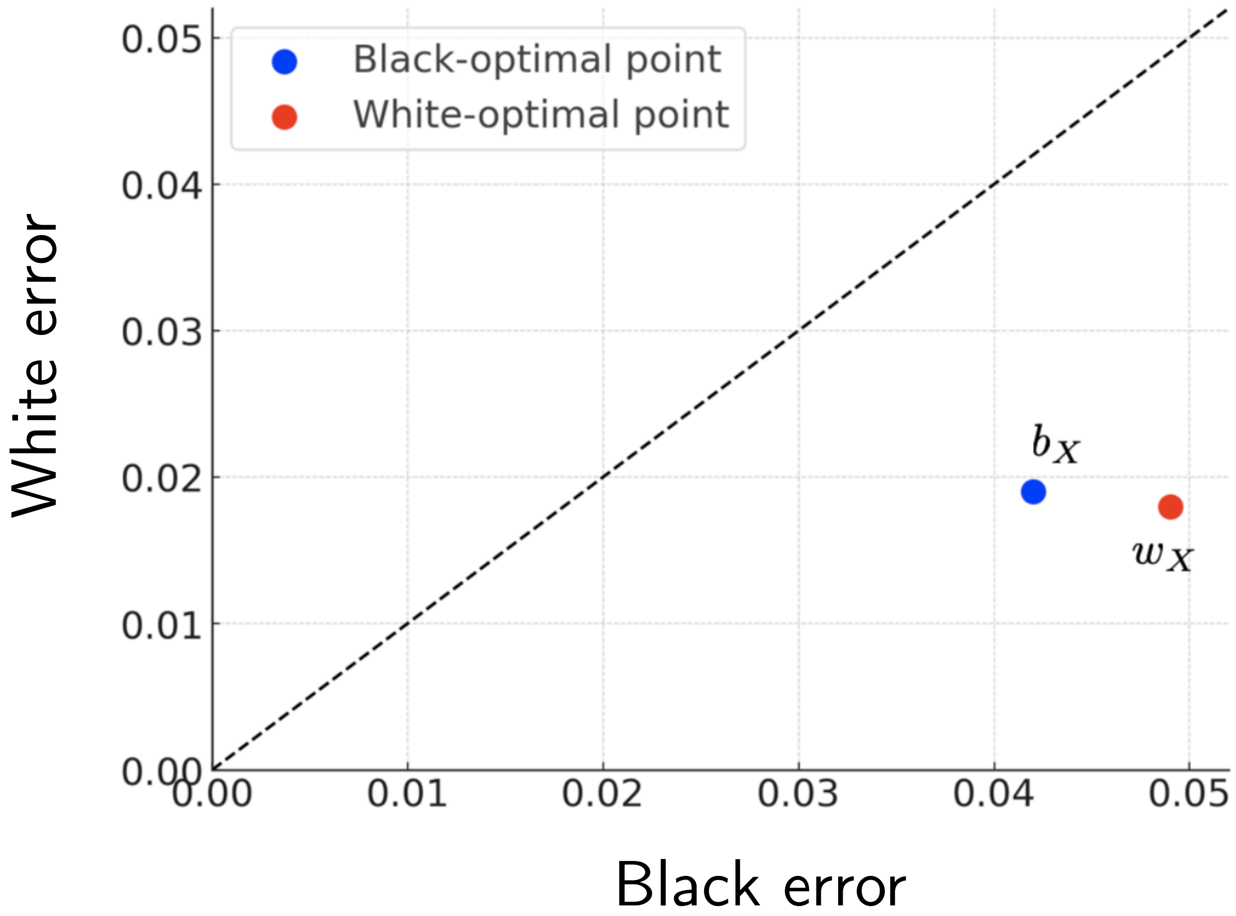
the data is from Obermeyer et al. (2019)

- 48,784 patient observations
- the covariate vector X includes 139 demographic + medical covariates
- group identities: Black or White, denoted $G \in \{b, w\}$
- true health needs are measured in the data as each patient's total number of active chronic illnesses in the subsequent year

the prediction problem:

- the hospital used these covariates to identify 3% of patients to automatically enroll in an intensive healthcare program
- $Y =$ indicator for whether the patient's health needs are in the top 3%
- consider algorithms $a : \mathscr{X} \to \{0,1\}$ and loss function $\ell(d, y) = 1(d \neq y)$
  - algorithms are more accurate if they have a lower misclassification rate for each group
  - more fair if the disparity between the misclassification rates for the two groups is smaller

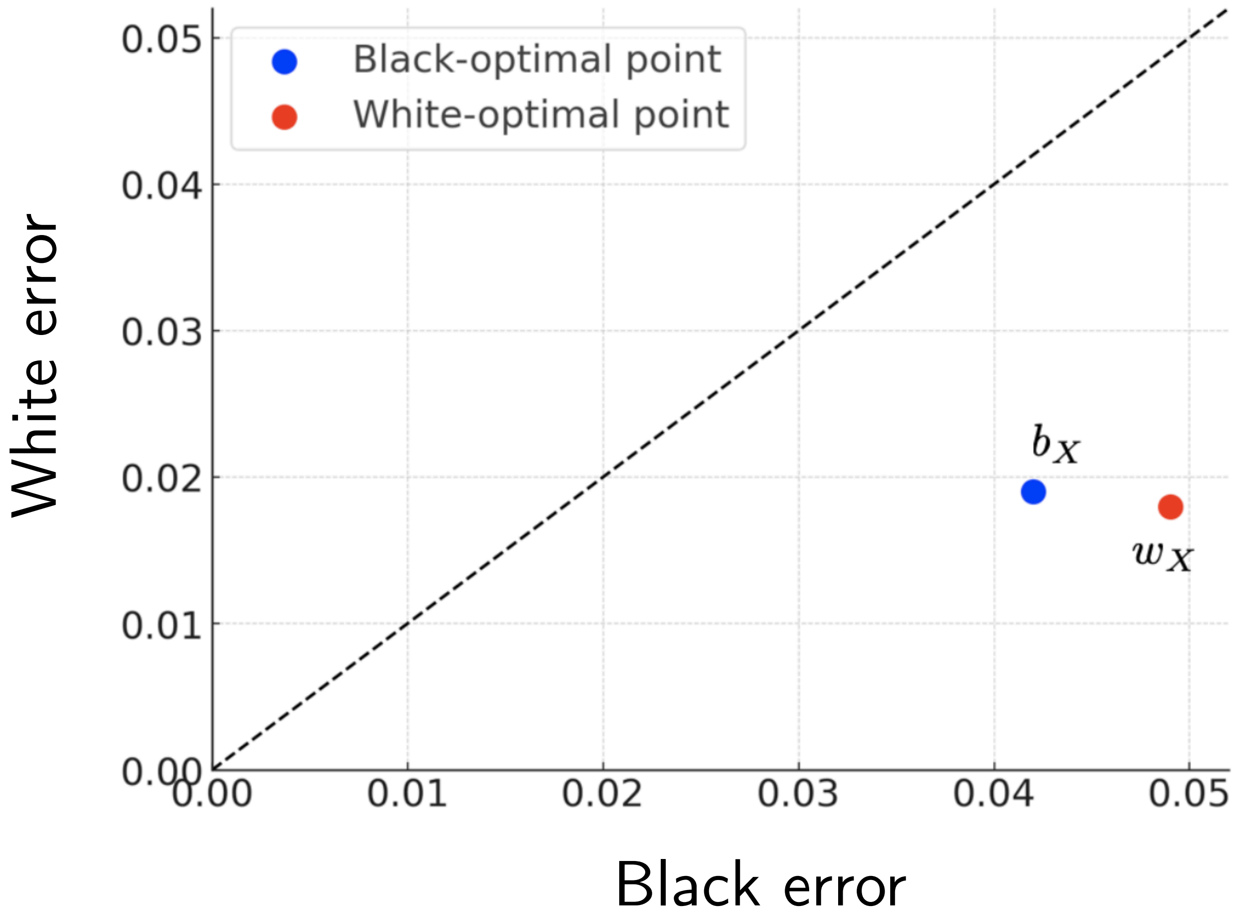
group-balance versus group-skew



how we estimate these group optimal points:

- split the sample into "training" and "test"
- use the training sample to identify the algorithm that minimizes group g's error
- assess error of this algorithm on the test sample for each group

group-balance versus group-skew



how we estimate these group optimal points:

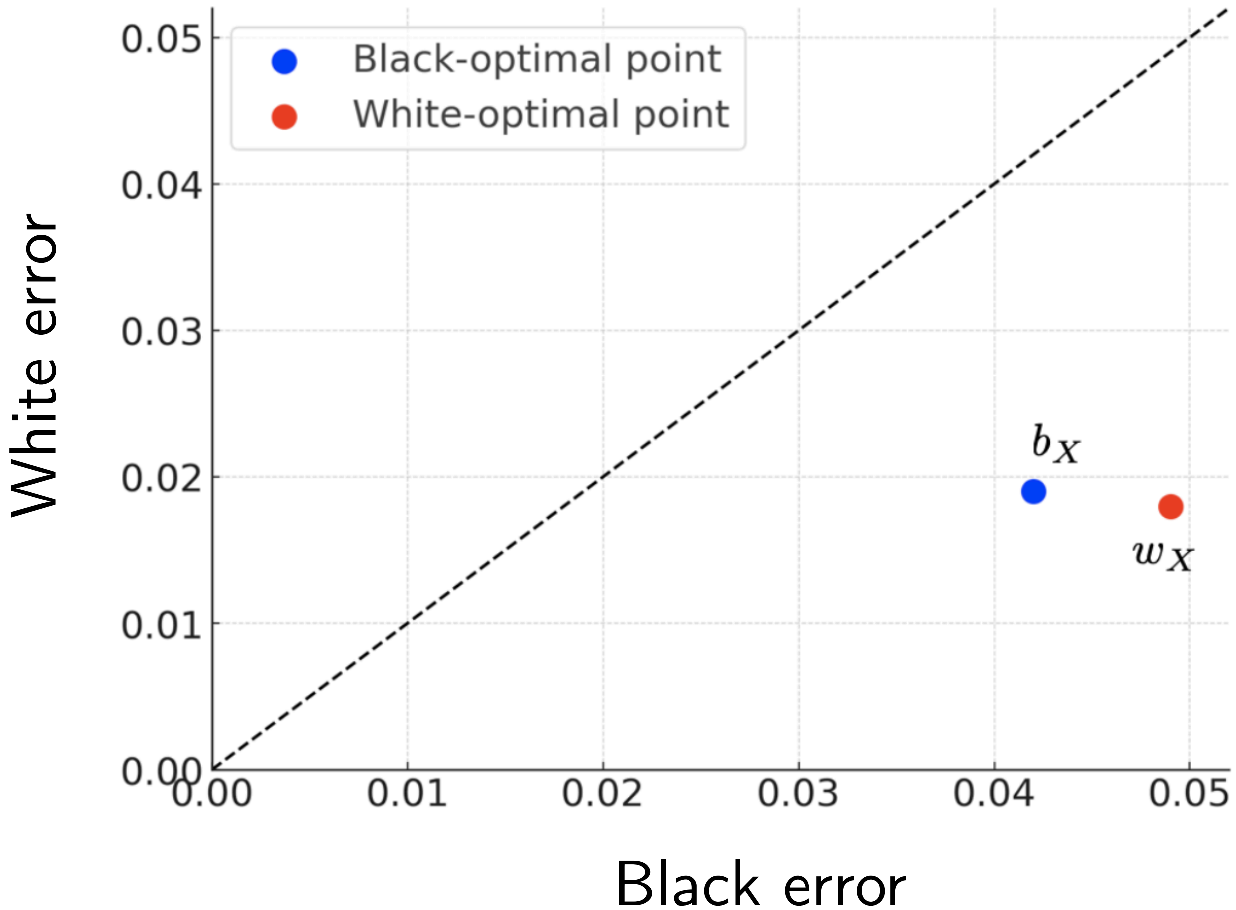
- split the sample into "training" and "test"
- use the training sample to identify the algorithm that minimizes group g's error
- assess error of this algorithm on the test sample for each group

group-skewed:

Black error is higher even at the Black-optimal point

(statistically significant, see paper for details)

group-balance versus group-skew



how we estimate these group optimal points:

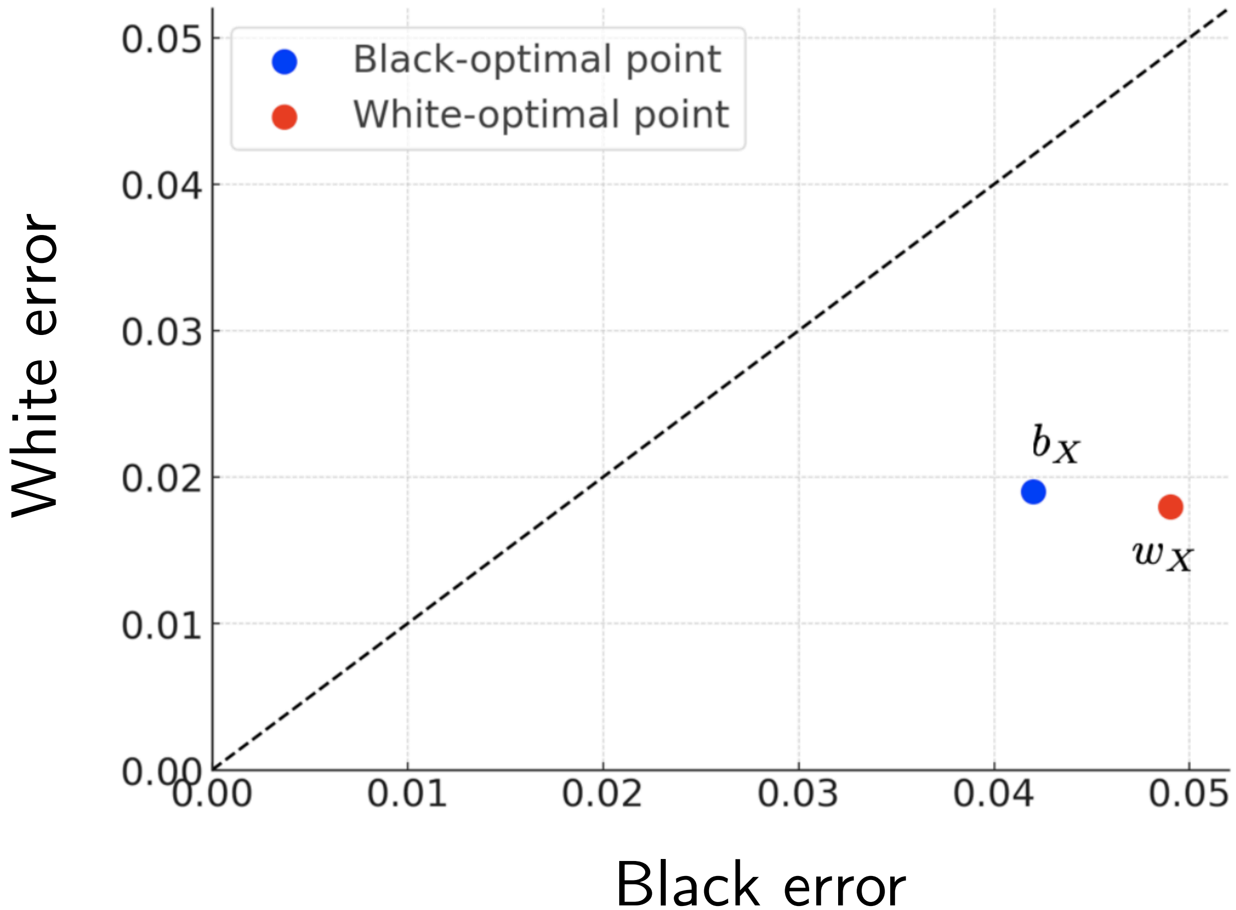
- split the sample into "training" and "test"
- train a **random forest algorithm** to minimize group g's error on the training sample
- assess error of this algorithm on the test sample for each group

group-skewed:

Black error is higher even at the Black-optimal point

(statistically significant, see paper for details)

group-balance versus group-skew



how we estimate these group optimal points:

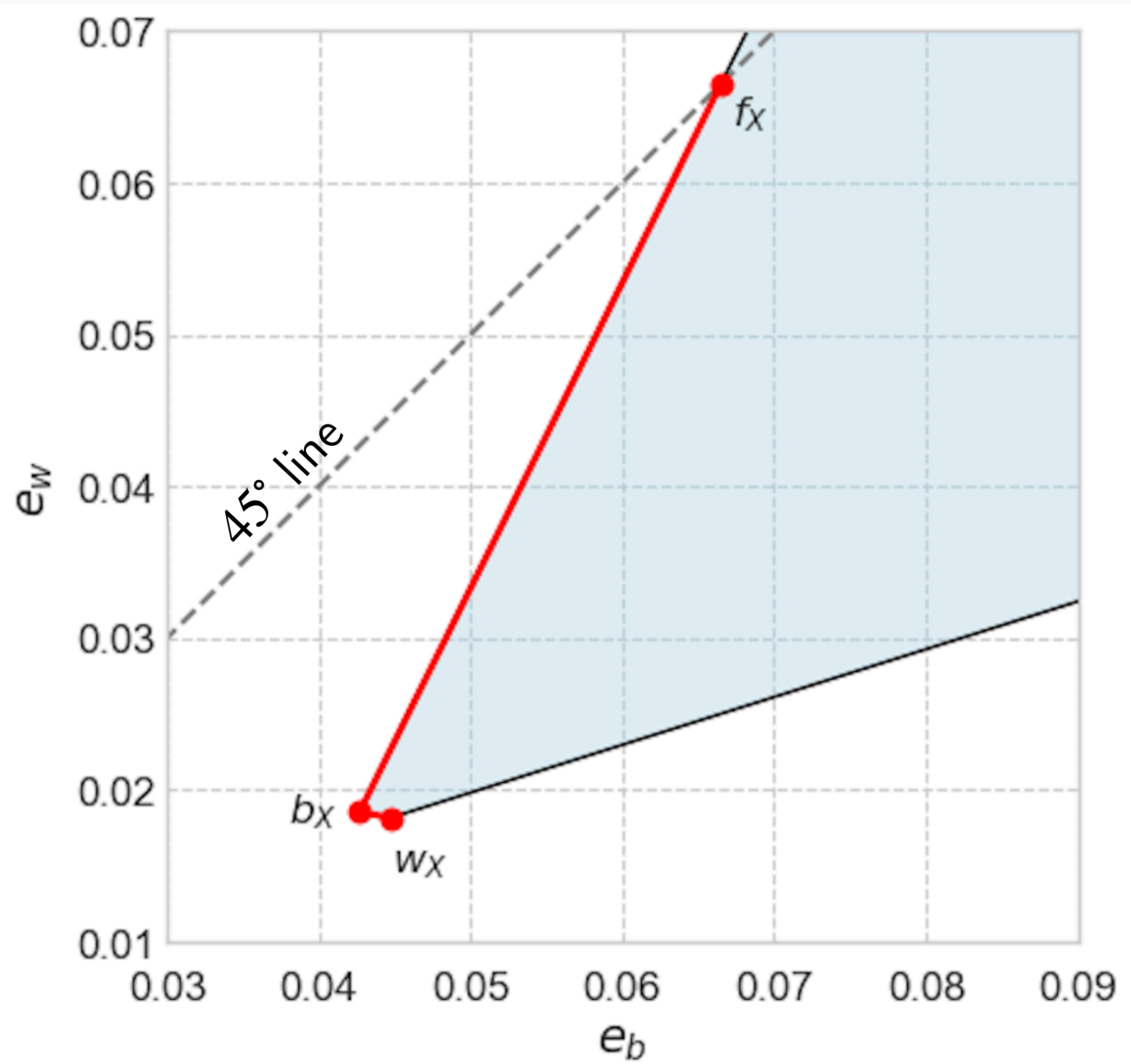
- split the sample into "training" and "test"
- train a **linear classifier** to minimize group g's error on the training sample
- assess error of this algorithm on the test sample for each group

group-skewed:

Black error is higher even at the Black-optimal point

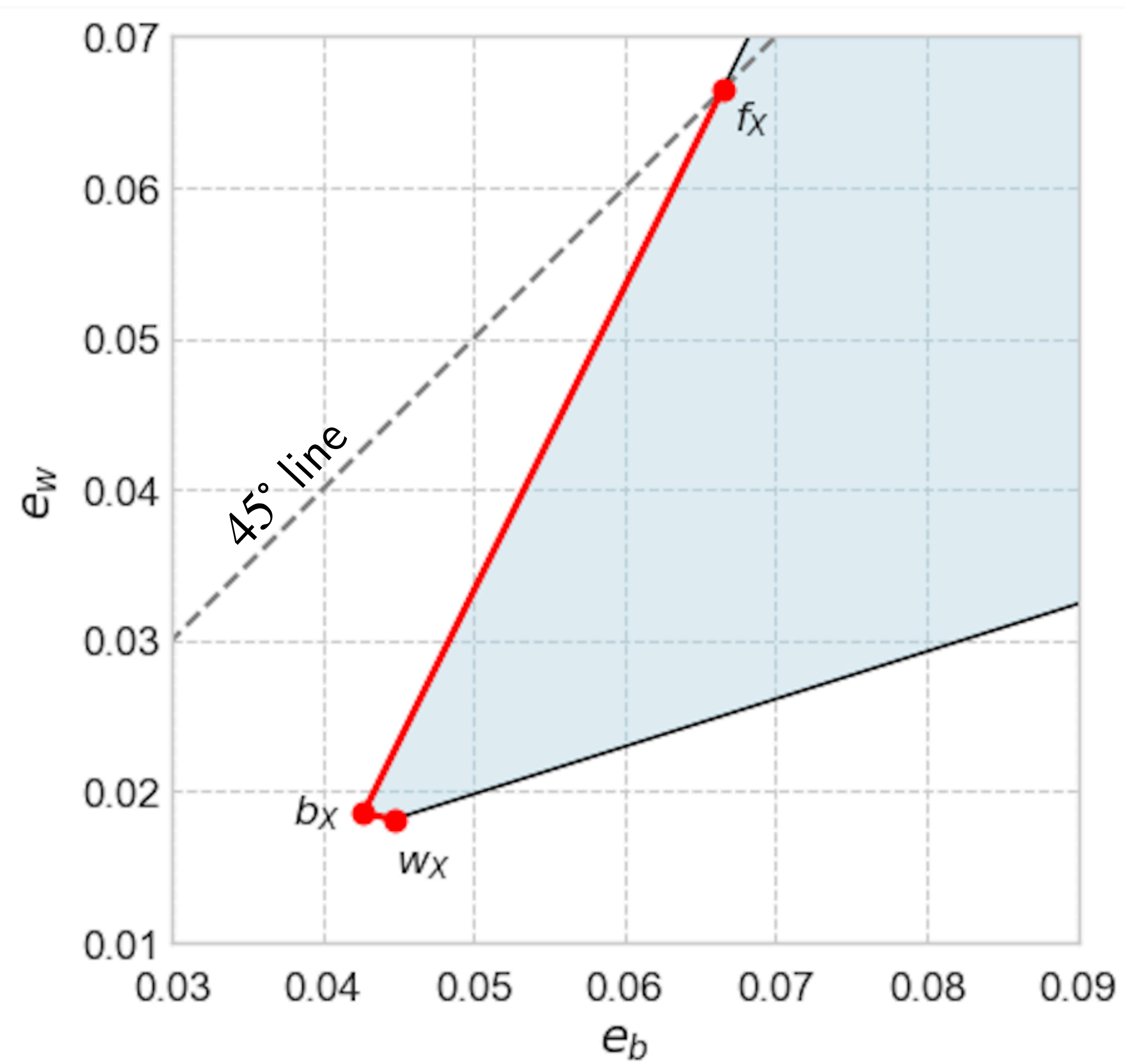
(statistically significant, see paper for details)

fairness-accuracy frontier



(depicted for the set of linear classifiers)

fairness-accuracy frontier

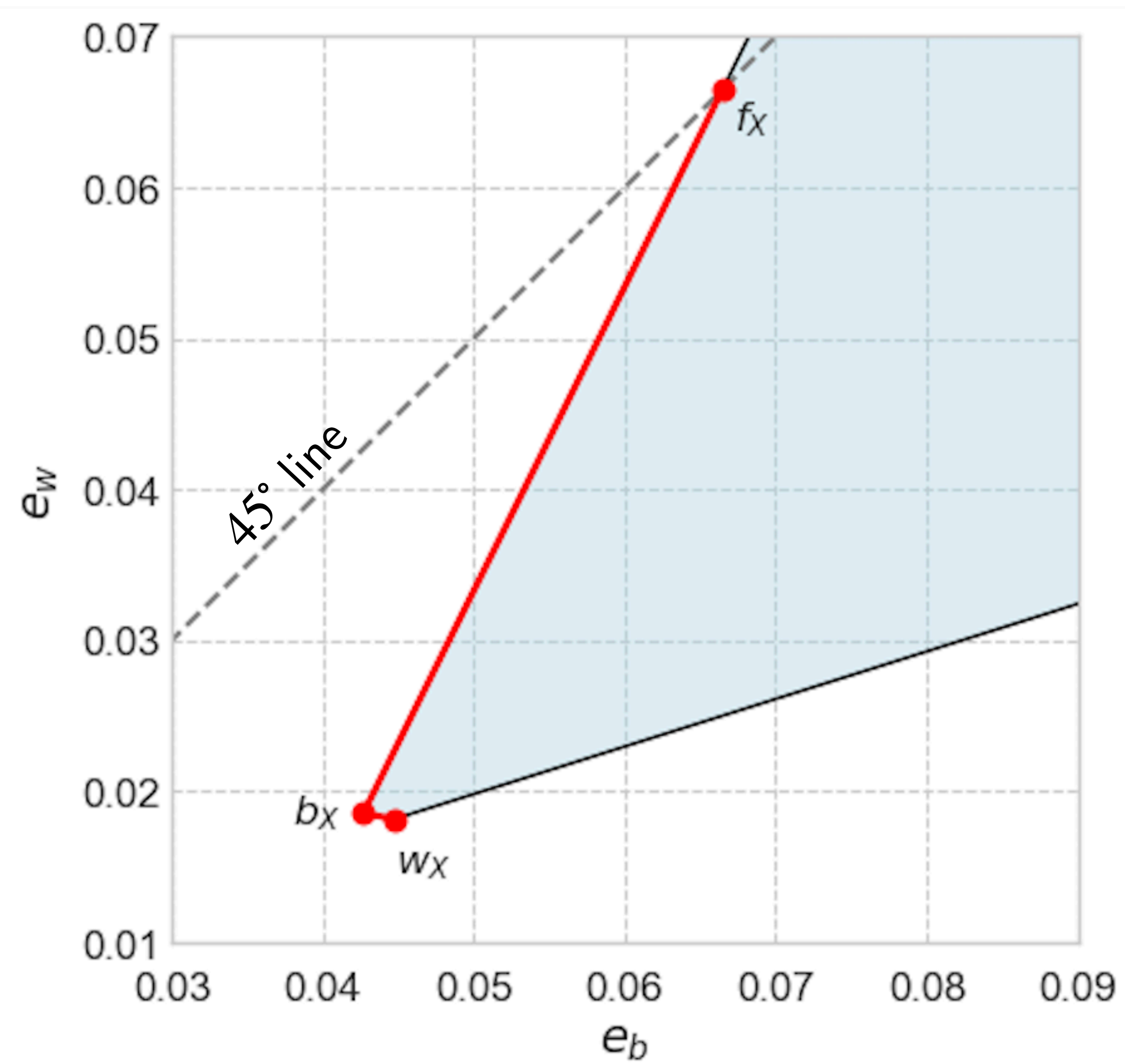


(depicted for the set of linear classifiers)

- strong fairness-accuracy conflict:

  main tradeoff is whether the designer is willing to increase errors for **both groups** to improve fairness

fairness-accuracy frontier



(depicted for the set of linear classifiers)

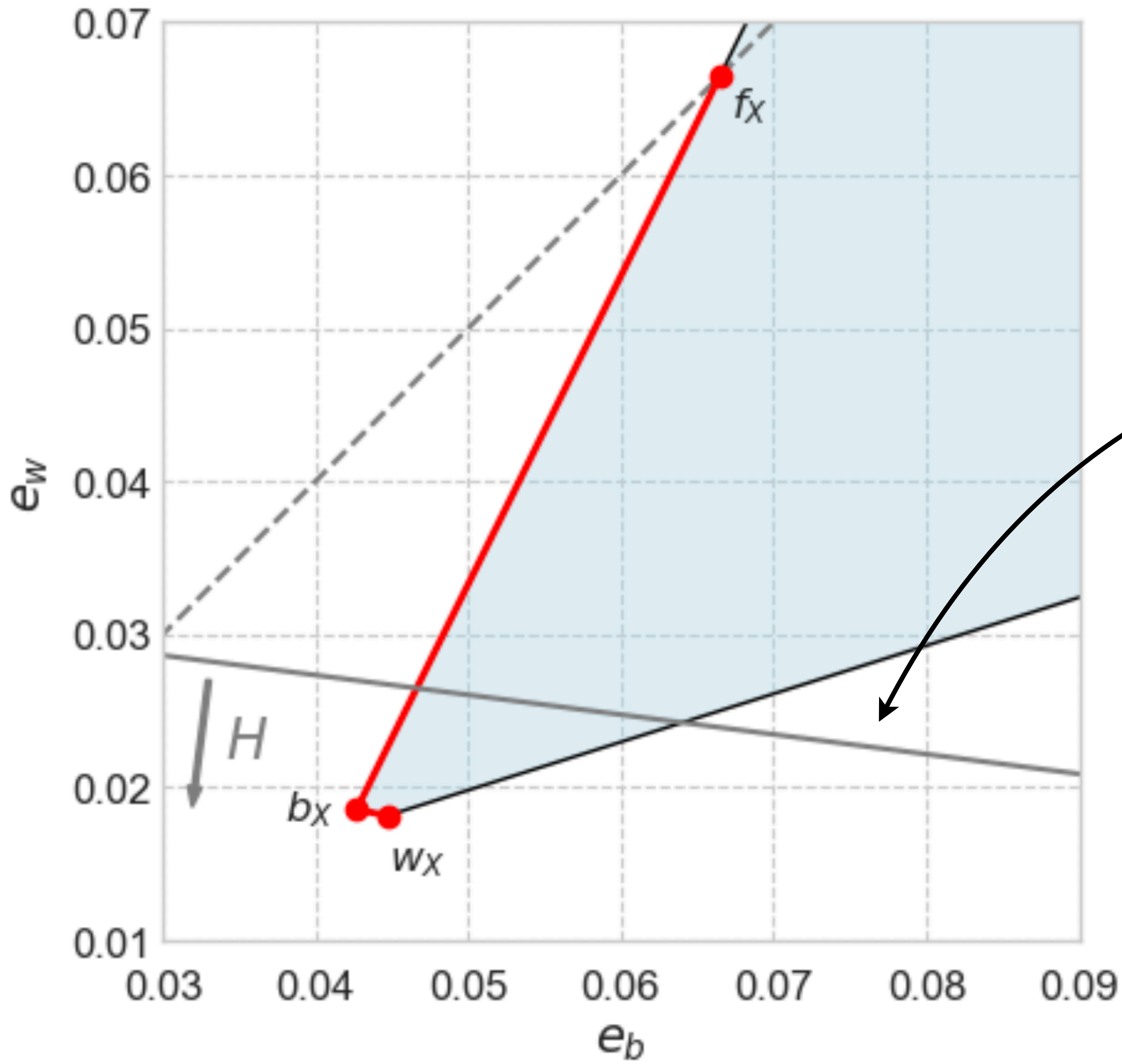
- strong fairness-accuracy conflict:

  main tradeoff is whether the designer is willing to increase errors for **both groups** to improve fairness

- qualitatively resembles Conditional Independence case ($G \perp Y \mid X$)

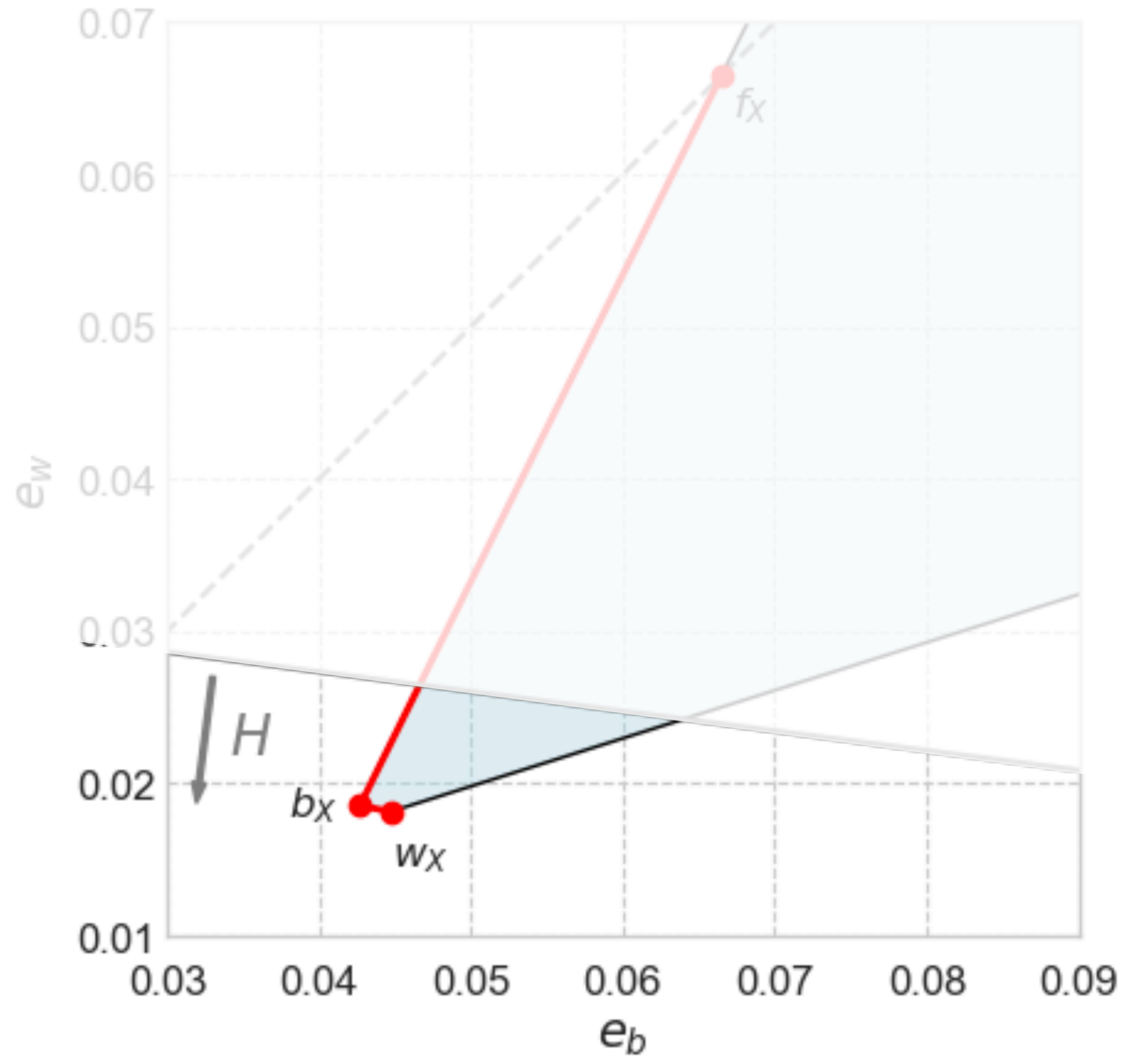
- consistent with a setting in which:
  - the optimal algorithm is the same for both groups
  - measured covariates are more predictive for White patients than Black patients

input design is with loss

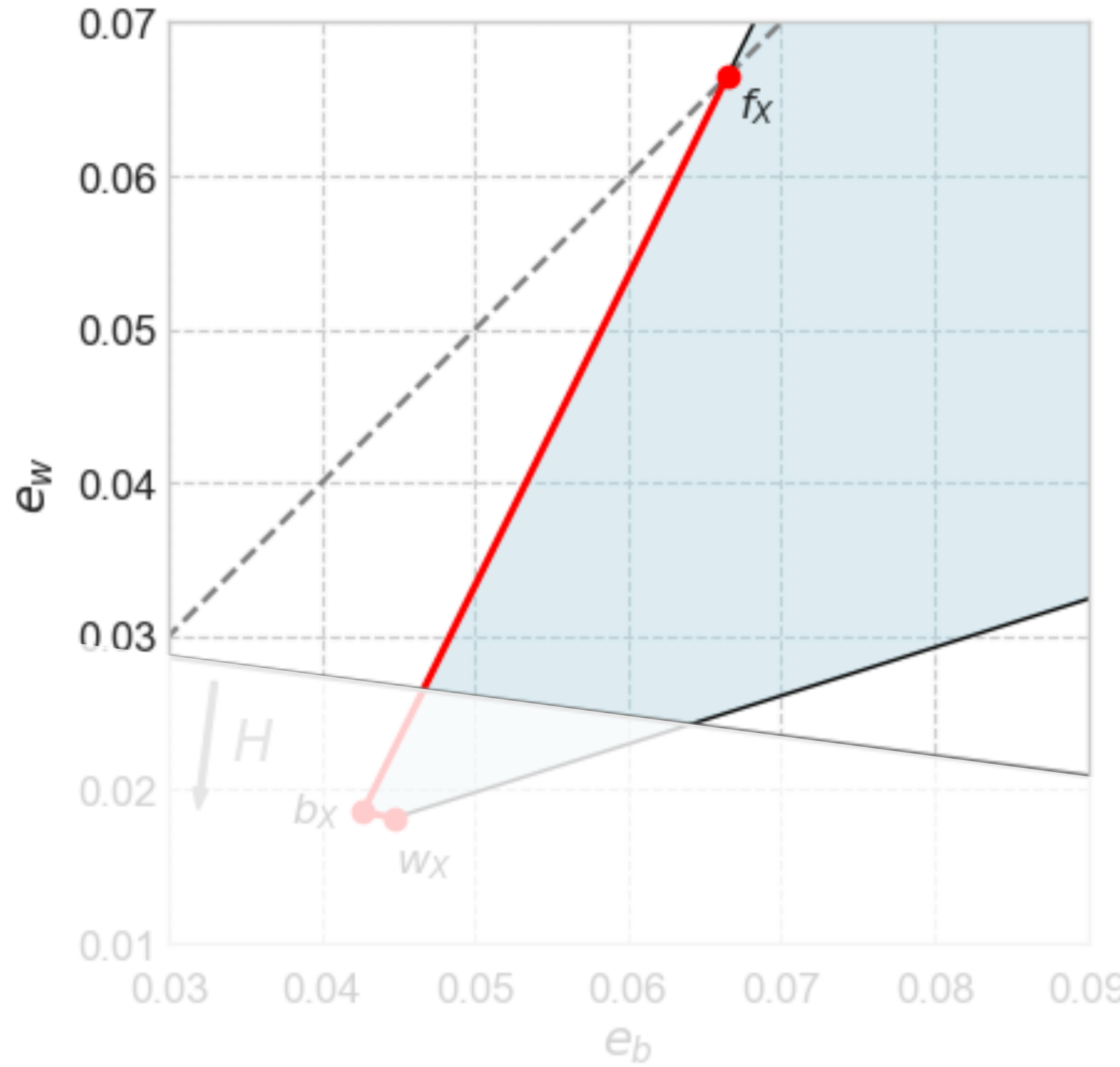


"no information"
indifference curve for
a utilitarian agent

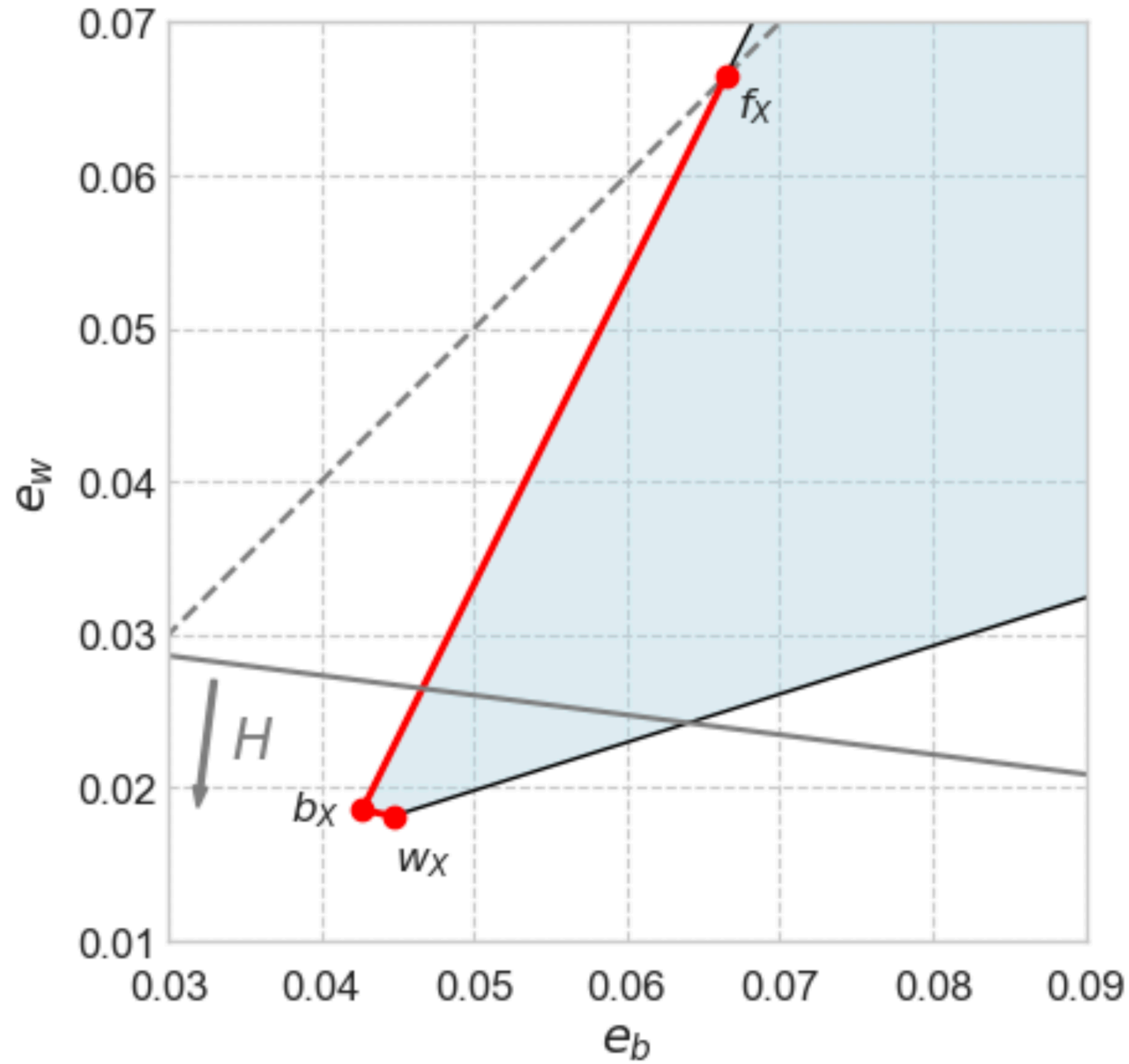
input design is with loss



input design is with loss

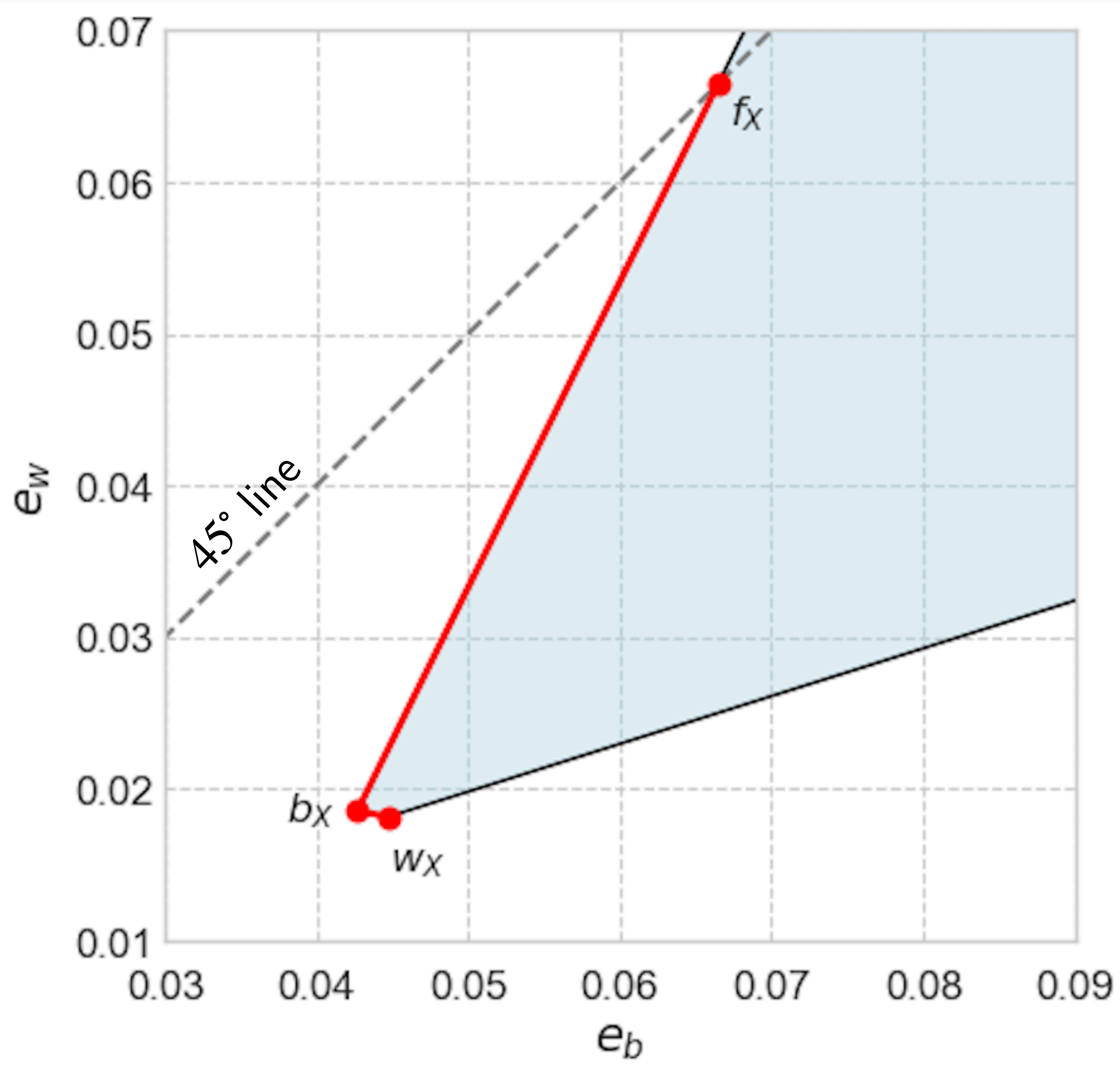


input design is with loss



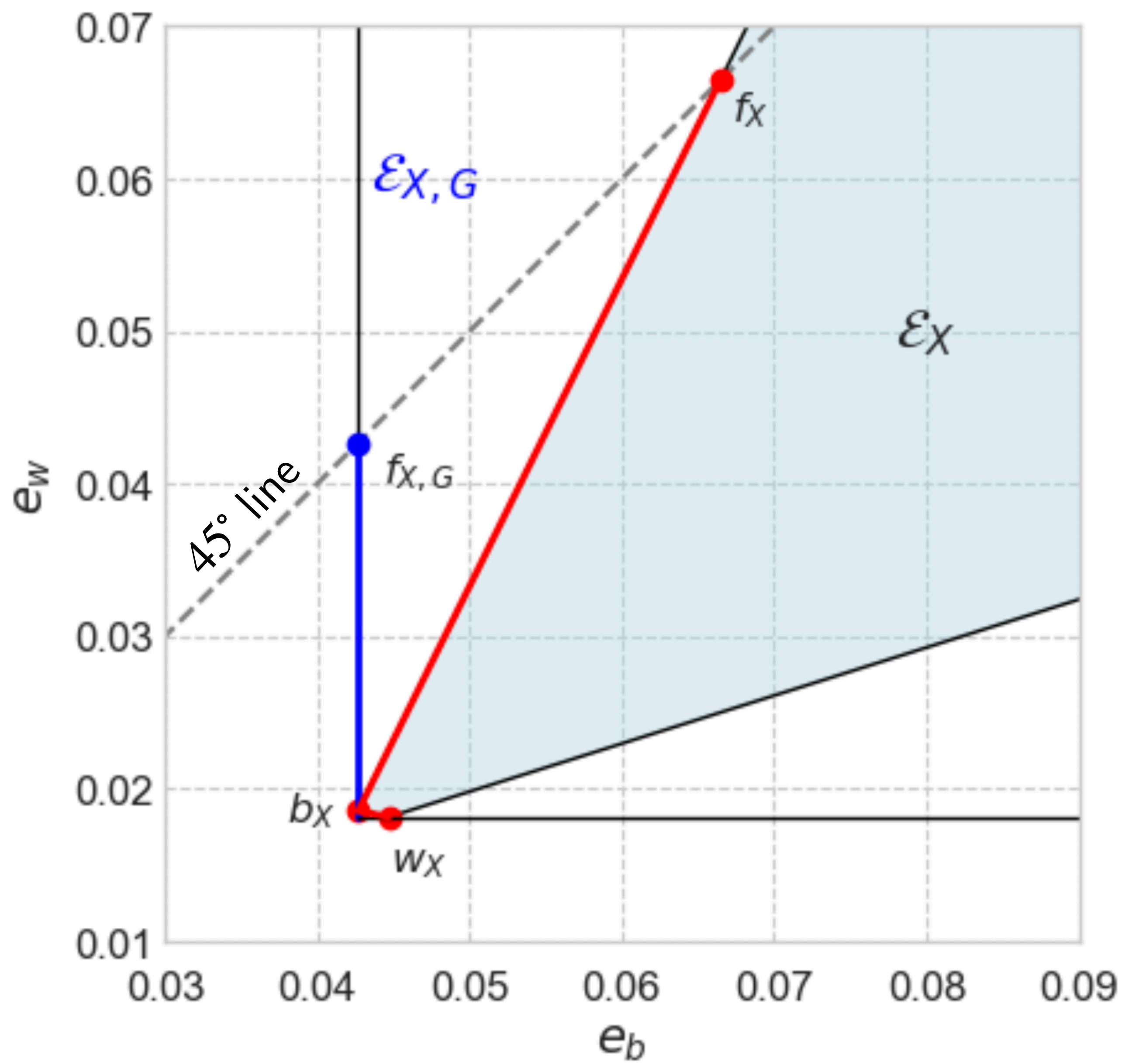
not all designers with FA preferences can implement their favorite outcome using input design

adding group identity as a covariate



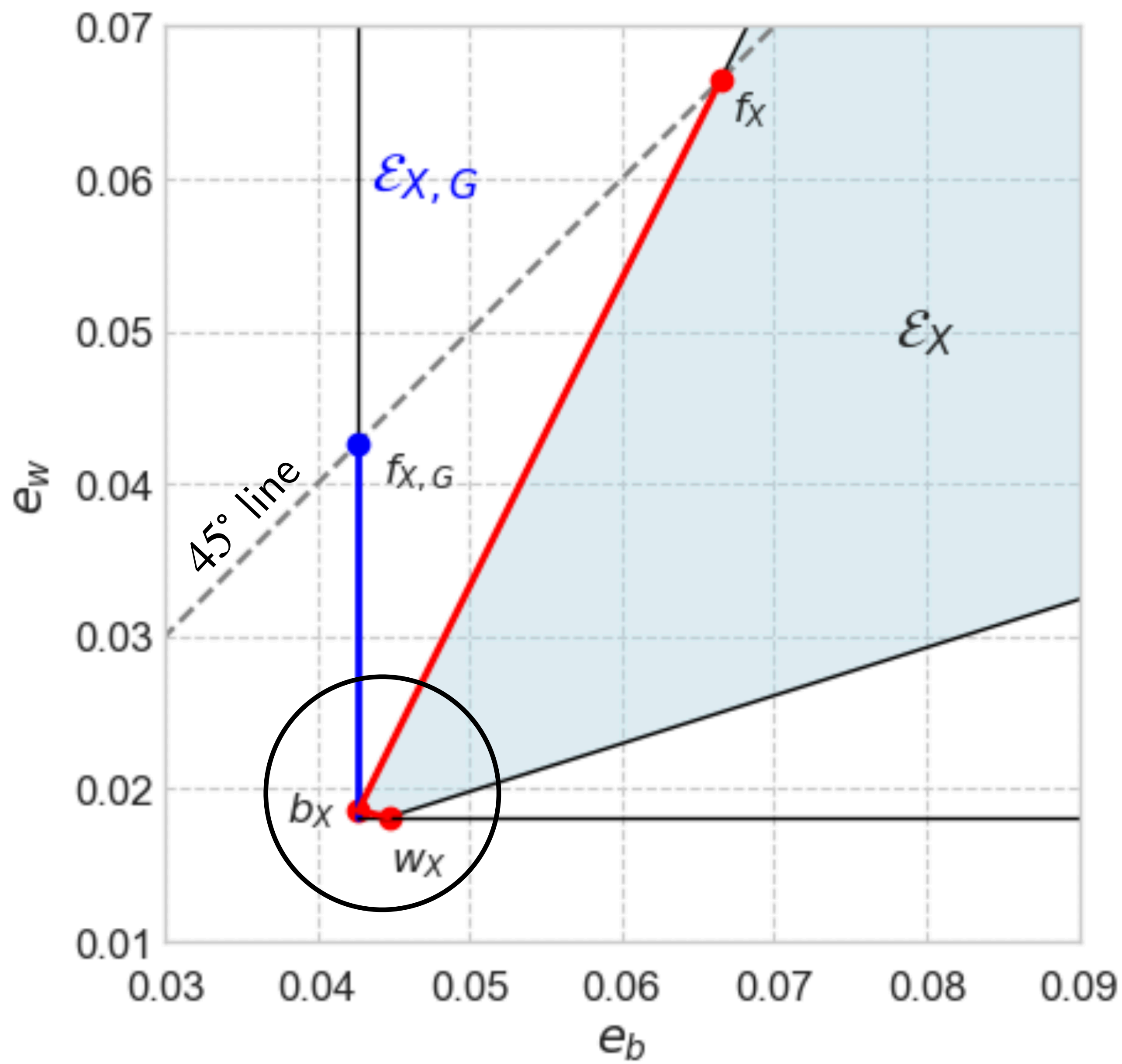
there are currently active debates regarding whether to include race as a group variable in healthcare prediction algorithms

adding group identity as a covariate



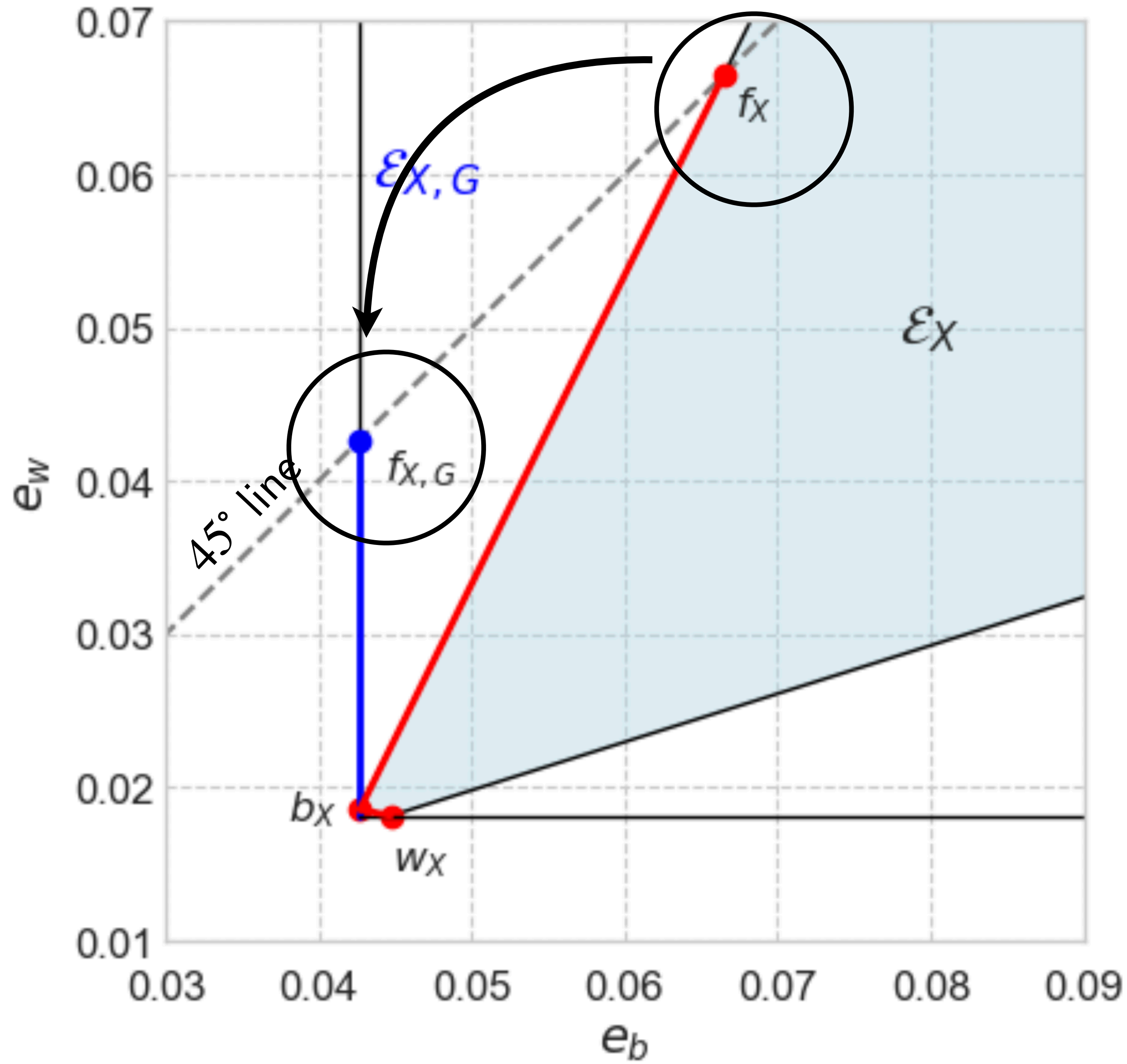
here's how the FA frontier changes when a separate algorithm is permitted for each group

adding group identity as a covariate



adding group identity has little effect on
the utilitarian-optimal point

adding group identity as a covariate



the largest effect is on the fairness-optimal point

conclusion

we formalize a fairness-accuracy frontier for the evaluation of algorithms

demonstrate that qualitative conclusions can be made which hold uniformly over a large class of designer preferences

- e.g., when inputs are group-balanced, Pareto-dominated outcomes are not optimal even with strong fairness preferences
- when it is possible to choose group-dependent garblings of covariates, then banning covariates is never optimal

the framework is useful not just conceptually, but also to the empirical evaluation of algorithms that are used in practice

Testing the Fairness-Accuracy Improvability of Algorithms

Eric Auerbach
(Northwestern)

Annie Liang
(Northwestern)

Max Tabord-Meehan
(UChicago)

Kyohei Okumura
(Northwestern)

introduction

- disparate impact has been empirically documented in a range of applications
- but the organizations that deploy these algorithms also value other objectives such as accuracy and profit
- when an algorithm has a disparate impact, is it possible to reduce that disparity without compromising the organization's other objectives?
- the answer to this question is legally relevant:

  disparate impact that would otherwise be prohibited under US federal law is often permissible if necessary to achieve a business interest

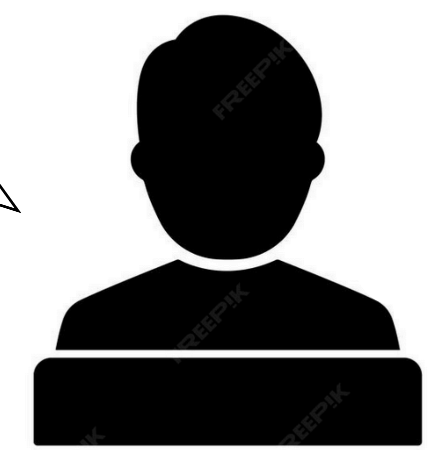
PART 1:

ESTABLISHING DISPARATE IMPACT

FIRM

(employs an algorithm, e.g., to make hiring decisions)

this algorithm has disproportionate harms for blue people



CHALLENGER

(e.g., a commission or private individual)

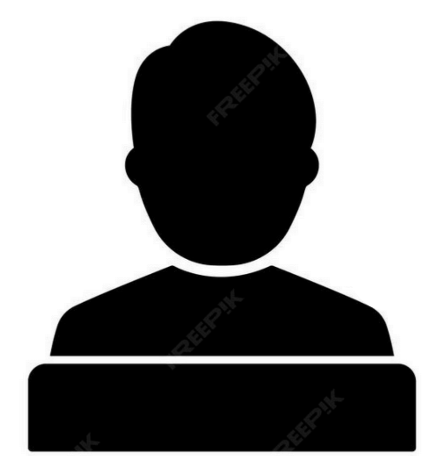
FIRM

(employs an algorithm, e.g., to make hiring decisions)

this algorithm is a business necessity, i.e., it is necessary to achieve a legitimate nondiscriminatory interest

PART 2:

ESTABLISHING BUSINESS NECESSITY



CHALLENGER

(e.g., a commission or private individual)

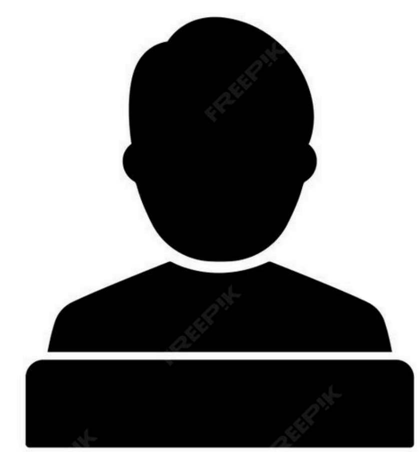
PART 3:

IS THERE A VALID LESS-DISCRIMINATORY ALTERNATIVE?

FIRM

(employs an algorithm, e.g., to make hiring decisions)

this alternative algorithm would achieve those same business objectives, and has less disparate impact



CHALLENGER

(e.g., a commission or private individual)

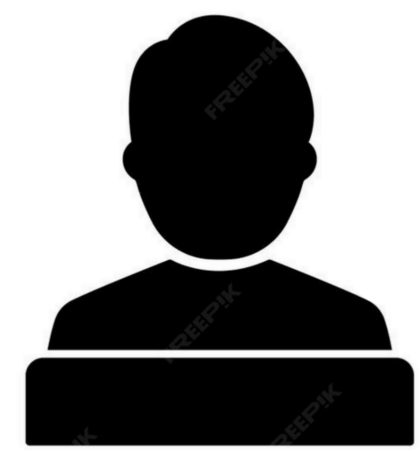
PART 3:

IS THERE A VALID LESS-DISCRIMINATORY ALTERNATIVE?

FIRM

(employs an algorithm, e.g., to make hiring decisions)

this alternative algorithm would achieve those same business objectives, and has less disparate impact



WINS

CHALLENGER

(e.g., a commission or private individual)

there exist established statistical procedures for this part →

PART 1:

ESTABLISHING DISPARATE IMPACT

there exist established statistical procedures for this part

PART 1:

ESTABLISHING DISPARATE IMPACT

our paper focuses on developing a statistical framework and tests for evaluating these latter parts

PART 2:

ESTABLISHING BUSINESS NECESSITY

PART 3:

IS THERE A VALID LESS-DISCRIMINATORY ALTERNATIVE?

setup

- each subject is described by three variables:
  - **type** Y taking values in $\mathscr{Y}$
  - **group** $G \in \mathscr{G} = \{r, b\}$
  - **covariate vector** X taking values in $\mathscr{X}$
- an algorithm is a map $a : \mathscr{X} \to \mathscr{D}$ from covariate vectors into a decision in $\mathscr{D}$
- there is a primitive set of permissible algorithms $\mathscr{A}$
- in the population, $(X, Y, G) \sim P$ (with no restrictions on P)
- analyst does not know P, but observes a sample $\{(X_i, Y_i, G_i)\}$ consisting of n i.i.d. observations from P

accuracy and fairness

- there is an accuracy utility function $u_A : \mathscr{X} \times \mathscr{Y} \times \mathscr{D} \to \mathbb{R}$ and a (possibly identical) fairness utility function $u_F : \mathscr{X} \times \mathscr{Y} \times \mathscr{D} \to \mathbb{R}$

accuracy and fairness

- there is an accuracy utility function $u_A : \mathscr{X} \times \mathscr{Y} \times \mathscr{D} \to \mathbb{R}$ and a (possibly identical) fairness utility function $u_F : \mathscr{X} \times \mathscr{Y} \times \mathscr{D} \to \mathbb{R}$

stand-in for any business objective
unrelated to fairness across groups

accuracy and fairness

- there is an accuracy utility function $u_A : \mathcal{X} \times \mathcal{Y} \times \mathcal{D} \to \mathbb{R}$ and a (possibly identical) fairness utility function $u_F : \mathcal{X} \times \mathcal{Y} \times \mathcal{D} \to \mathbb{R}$

- we consider accuracy and fairness criteria that can be formulated as

$$\underbrace{\begin{aligned} U_A^g(a) &= E_P[u_A(X, Y, a(X)) \mid G = g] \\ U_F^g(a) &= E_P[u_F(X, Y, a(X)) \mid G = g] \end{aligned}}_{\substack{\text{expected utility for either group using the} \\ \text{respective utility function}}}$$

accuracy and fairness

- there is an accuracy utility function $u_A : \mathscr{X} \times \mathscr{Y} \times \mathscr{D} \to \mathbb{R}$ and a (possibly identical) fairness utility function $u_F : \mathscr{X} \times \mathscr{Y} \times \mathscr{D} \to \mathbb{R}$
- we consider accuracy and fairness criteria that can be formulated as

$$U_A^g(a) = E_P[u_A(X, Y, a(X)) \mid G = g]$$
$$U_F^g(a) = E_P[u_F(X, Y, a(X)) \mid G = g]$$

definition: algorithm a_1 is **more accurate** than algorithm a_0 if

$$U_A^r(a_1) > U_A^r(a_0) \text{ and } U_A^b(a_1) > U_A^b(a_0)$$

and **more fair** than algorithm a_0 if

$$|\, U_F^r(a_1) - U_F^b(a_1) \,| < |\, U_F^r(a_0) - U_F^b(a_0) \,|$$

accuracy- and fairness- improvability

definition: fix any $\Delta_r, \Delta_b, \Delta_f \in \mathbb{R}_+$. the algorithm a_1 constitutes a $(\Delta_r, \Delta_b, \Delta_f)$ -improvement on the algorithm a_0 if

$$\underbrace{\frac{U_A^r(a_1)}{U_A^r(a_0)} > 1 + \Delta_r,}_{\substack{\Delta_r\text{-percent increase in} \\ \text{accuracy for group } r}}$$

accuracy- and fairness- improvability

definition: fix any $\Delta_r, \Delta_b, \Delta_f \in \mathbb{R}_+$. the algorithm a_1 constitutes a $(\Delta_r, \Delta_b, \Delta_f)$ -improvement on the algorithm a_0 if

$$\frac{U_A^r(a_1)}{U_A^r(a_0)} > 1 + \Delta_r, \quad \underbrace{\frac{U_A^b(a_1)}{U_A^b(a_0)} > 1 + \Delta_b}_{\text{Δ_b-percent increase in accuracy for group b}},$$

accuracy- and fairness- improvability

definition: fix any $\Delta_r, \Delta_b, \Delta_f \in \mathbb{R}_+$. the algorithm a_1 constitutes a $(\Delta_r, \Delta_b, \Delta_f)$ -improvement on the algorithm a_0 if

$$\frac{U_A^r(a_1)}{U_A^r(a_0)} > 1 + \Delta_r, \quad \frac{U_A^b(a_1)}{U_A^b(a_0)} > 1 + \Delta_b \text{ , and } \underbrace{\frac{|\,U_F^r(a_1) - U_F^b(a_1)\,|}{|\,U_F^r(a_0) - U_F^b(a_0)\,|} < 1 - \Delta_f}_{\substack{\Delta_f\text{-percent reduction in} \\ \text{disparate impact}}}$$

accuracy- and fairness- improvability

definition: fix any $\Delta_r, \Delta_b, \Delta_f \in \mathbb{R}_+$. the algorithm a_1 constitutes a $(\Delta_r, \Delta_b, \Delta_f)$ -improvement on the algorithm a_0 if

$$\frac{U_A^r(a_1)}{U_A^r(a_0)} > 1 + \Delta_r, \quad \frac{U_A^b(a_1)}{U_A^b(a_0)} > 1 + \Delta_b \text{ , and } \frac{|\,U_F^r(a_1) - U_F^b(a_1)\,|}{|\,U_F^r(a_0) - U_F^b(a_0)\,|} < 1 - \Delta_f$$

definition: algorithm a_0 is **FA-dominated within class** $\mathscr{A}$ if there exists an algorithm $a_1 \in \mathscr{A}$ that $(0,0,0)$-improves on a_0

- can strictly reduce disparate impact without compromising on accuracy for either group

accuracy- and fairness- improvability

definition: fix any $\Delta_r, \Delta_b, \Delta_f \in \mathbb{R}_+$. the algorithm a_1 constitutes a $(\Delta_r, \Delta_b, \Delta_f)$ -improvement on the algorithm a_0 if

$$\frac{U_A^r(a_1)}{U_A^r(a_0)} > 1 \quad , \quad \frac{U_A^b(a_1)}{U_A^b(a_0)} > 1 \quad , \text{ and } \frac{|\,U_F^r(a_1) - U_F^b(a_1)\,|}{|\,U_F^r(a_0) - U_F^b(a_0)\,|} < 1$$

definition: algorithm a_0 is **FA-dominated within class** $\mathscr{A}$ if there exists an algorithm $a_1 \in \mathscr{A}$ that $(0,0,0)$-improves on a_0

- can strictly reduce disparate impact without compromising on accuracy for either group

accuracy- and fairness- improvability

definition: fix any $\Delta_r, \Delta_b, \Delta_f \in \mathbb{R}_+$. the algorithm a_1 constitutes a $(\Delta_r, \Delta_b, \Delta_f)$ -improvement on the algorithm a_0 if

$$\frac{U_A^r(a_1)}{U_A^r(a_0)} > 1 \quad , \quad \frac{U_A^b(a_1)}{U_A^b(a_0)} > 1 \quad , \text{ and } \frac{|\,U_F^r(a_1) - U_F^b(a_1)\,|}{|\,U_F^r(a_0) - U_F^b(a_0)\,|} < 1$$

definition: algorithm a_0 is **FA-dominated within class** $\mathscr{A}$ if there exists an algorithm $a_1 \in \mathscr{A}$ that $(0{,}0{,}0)$-improves on a_0

- can strictly reduce disparate impact without compromising on accuracy for either group

directly related to the business-necessity defense in a disparate impact case

accuracy- and fairness- improvability

definition: fix any $\Delta_r, \Delta_b, \Delta_f \in \mathbb{R}_+$. the algorithm a_1 constitutes a $(\Delta_r, \Delta_b, \Delta_f)$ -improvement on the algorithm a_0 if

$$\frac{U_A^r(a_1)}{U_A^r(a_0)} > 1 \quad , \quad \frac{U_A^b(a_1)}{U_A^b(a_0)} > 1 \quad , \text{ and } \frac{|\, U_F^r(a_1) - U_F^b(a_1)\,|}{|\, U_F^r(a_0) - U_F^b(a_0)\,|} < 1 - \delta$$

definition: algorithm a_0 is δ-**fairness improvable within class** $\mathscr{A}$ if there exists an algorithm $a_1 \in \mathscr{A}$ that $(0,0,\delta)$-improves on a_0

- can reduce disparate impact by δ percent without compromising on accuracy for either group

accuracy- and fairness- improvability

definition: fix any $\Delta_r, \Delta_b, \Delta_f \in \mathbb{R}_+$. the algorithm a_1 constitutes a $(\Delta_r, \Delta_b, \Delta_f)$ -improvement on the algorithm a_0 if

$$\frac{U_A^r(a_1)}{U_A^r(a_0)} > 1+\delta \quad , \quad \frac{U_A^b(a_1)}{U_A^b(a_0)} > 1+\delta \quad \text{, and} \quad \frac{|\, U_F^r(a_1) - U_F^b(a_1)\,|}{|\, U_F^r(a_0) - U_F^b(a_0)\,|} < 1$$

definition: algorithm a_0 is δ-**accuracy improvable within class** $\mathscr{A}$ if there exists an algorithm $a_1 \in \mathscr{A}$ that $(\delta, \delta, 0)$-improves on a_0

- can improve accuracy by δ percent for both groups without compromising on fairness

accuracy- and fairness- improvability

definition: fix any $\Delta_r, \Delta_b, \Delta_f \in \mathbb{R}_+$. the algorithm a_1 constitutes a $(\Delta_r, \Delta_b, \Delta_f)$ -improvement on the algorithm a_0 if

$$\frac{U_A^r(a_1)}{U_A^r(a_0)} > 1+\delta \quad , \quad \frac{U_A^b(a_1)}{U_A^b(a_0)} > 1+\delta \quad \text{, and} \quad \frac{|\,U_F^r(a_1) - U_F^b(a_1)\,|}{|\,U_F^r(a_0) - U_F^b(a_0)\,|} < 1$$

definition: algorithm a_0 is δ-**accuracy improvable within class** $\mathscr{A}$ if there exists an algorithm $a_1 \in \mathscr{A}$ that $(\delta, \delta, 0)$-improves on a_0

- can improve accuracy by δ percent for both groups without compromising on fairness

not legally relevant, but an interesting complement on the previous perspective

what we want to evaluate

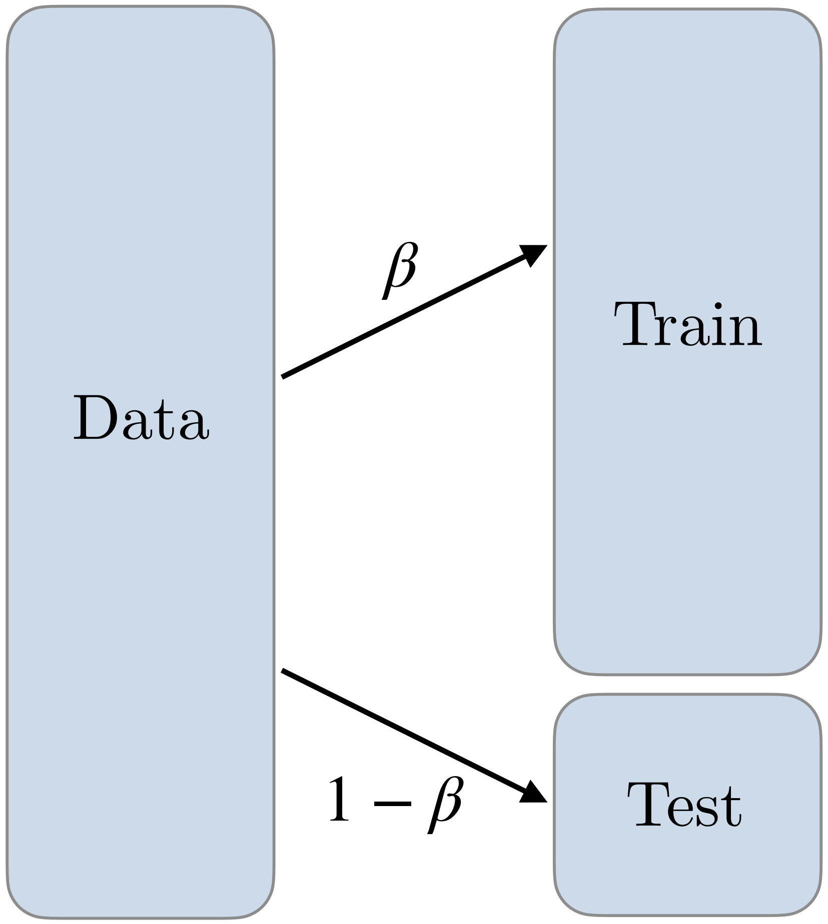
our goal is to evaluate the accuracy- and fairness-improvability of a status quo algorithm within a given class of algorithms

formally, we will test the null hypothesis

H_0: algorithm a_0 is not δ-fairness (or accuracy) improvable within class $\mathscr{A}$

proposed approach

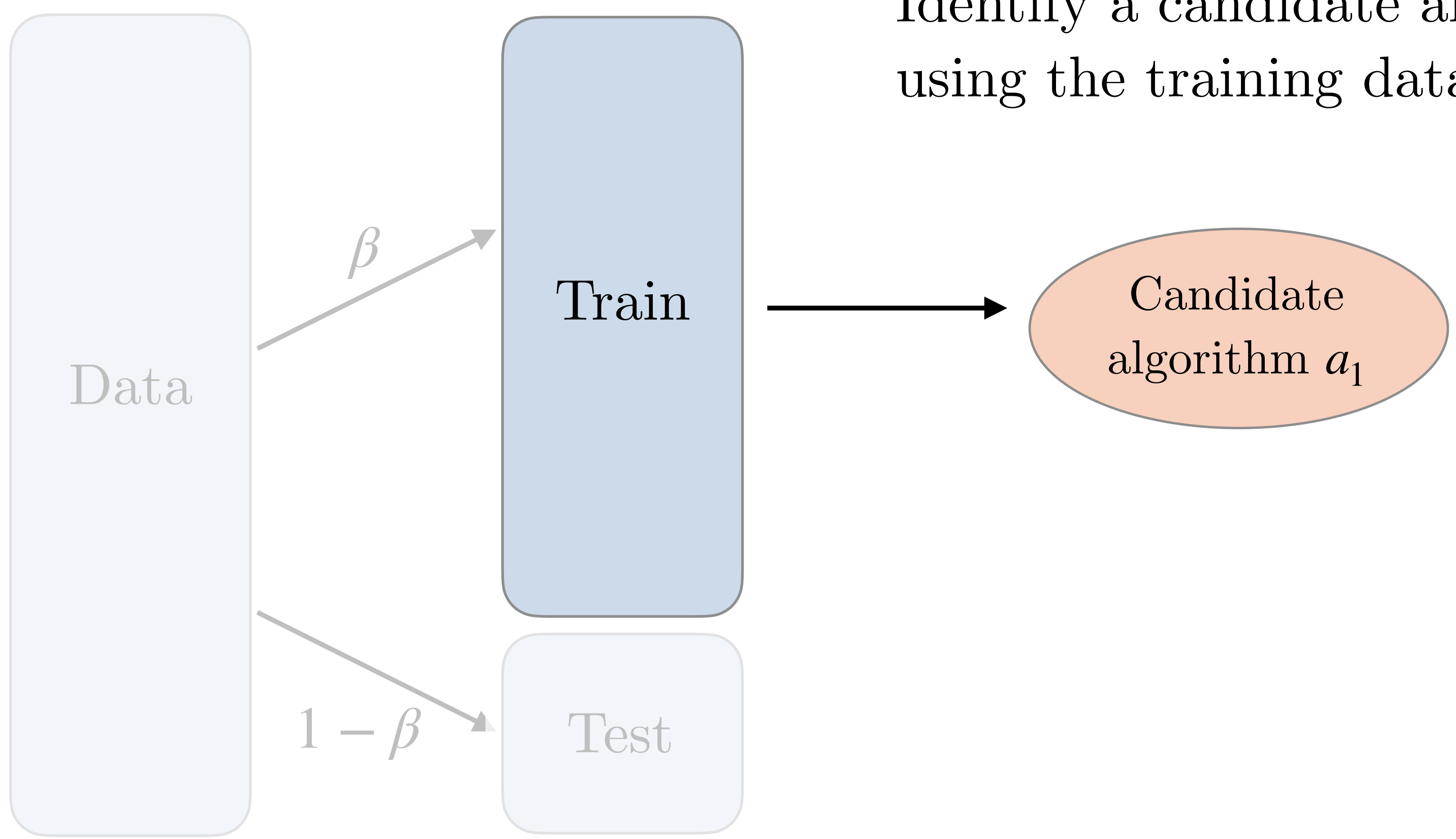
our proposed procedure



Step 1:

Randomly split the data into train and test sets.

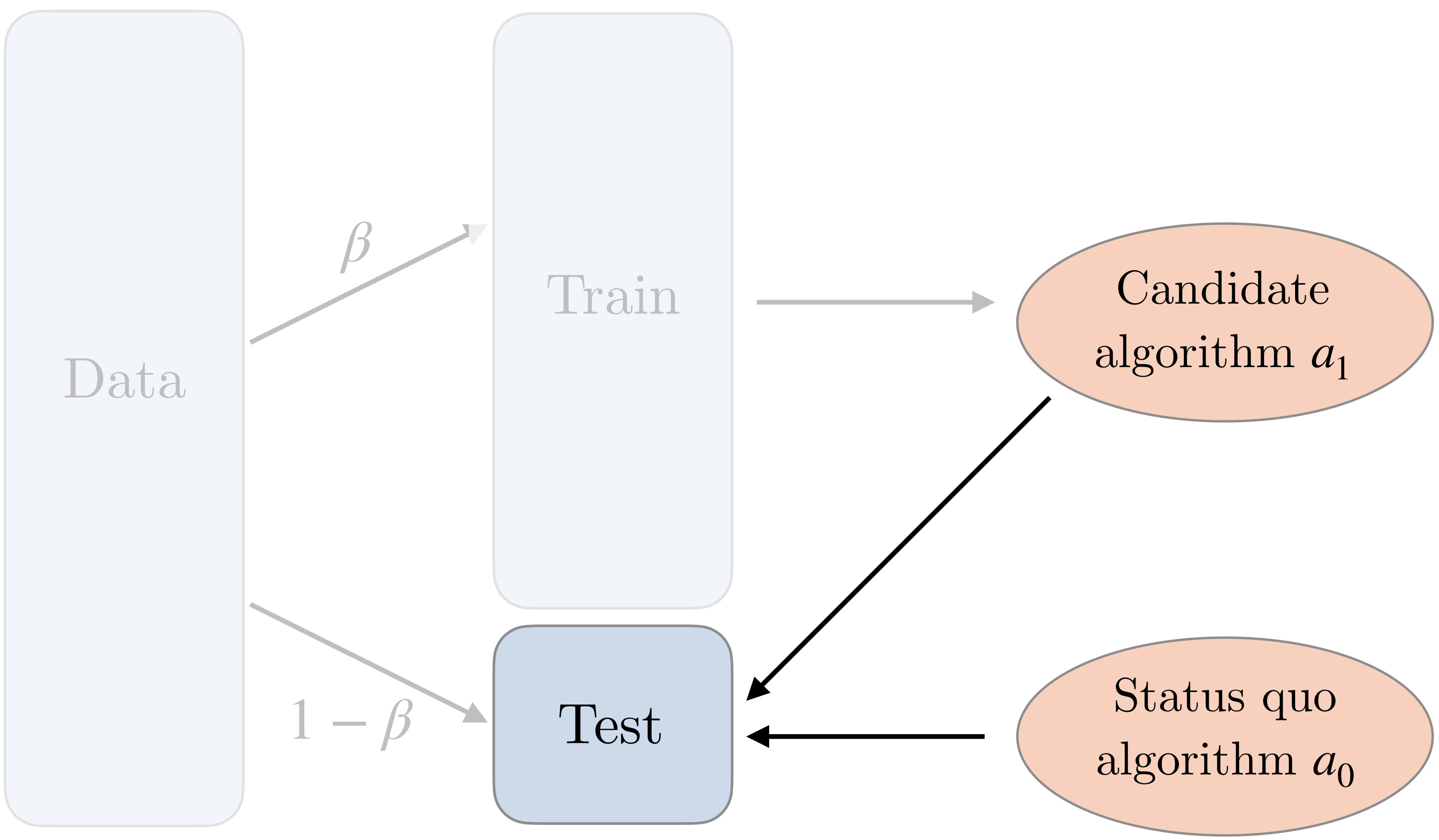
our proposed procedure



Step 2:

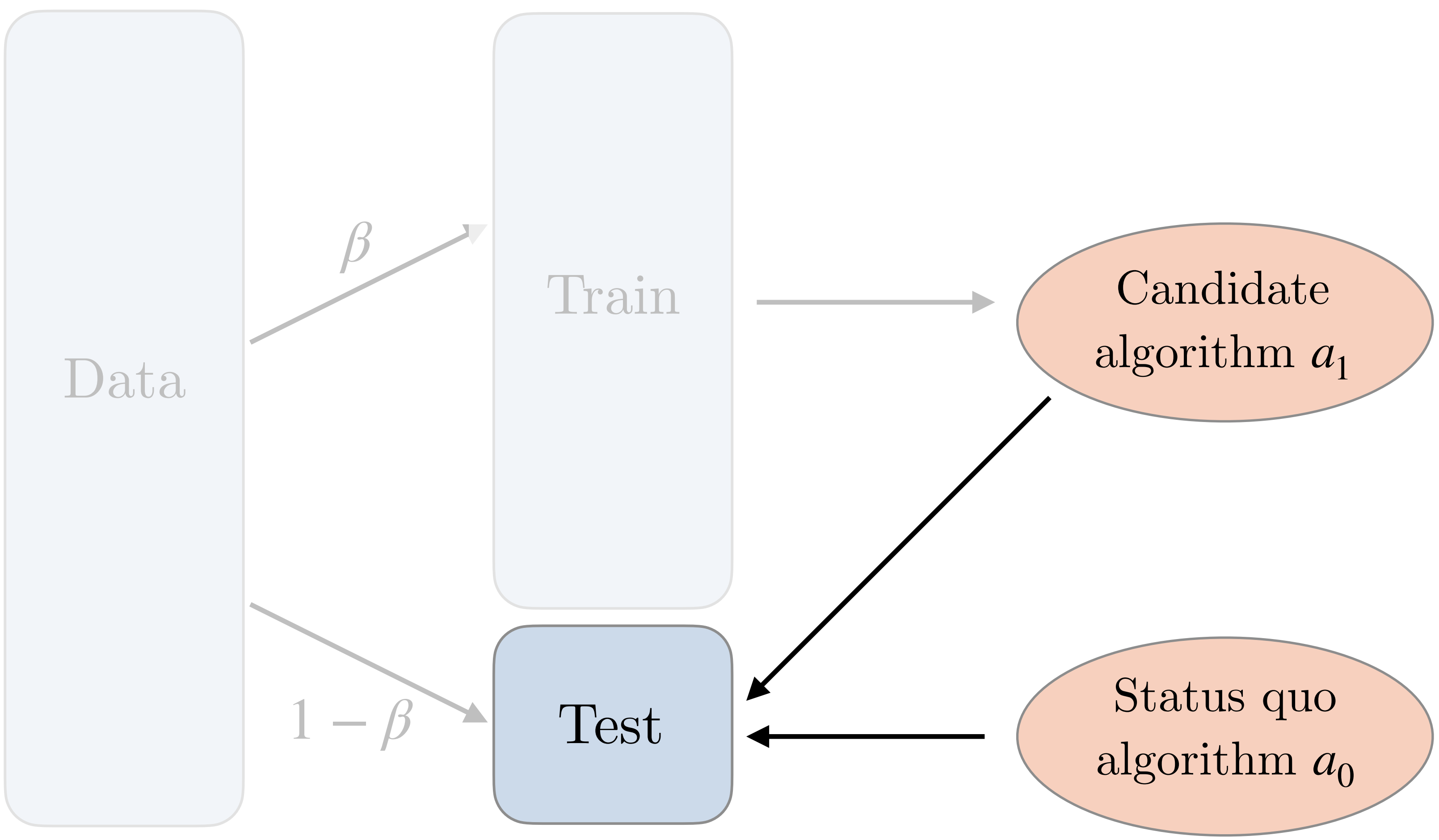
Identify a candidate algorithm a_1 using the training data and

our proposed procedure



Step 3: Test whether a_1 constitutes an $(\Delta_r, \Delta_b, \Delta_f)$-improvement on a_0

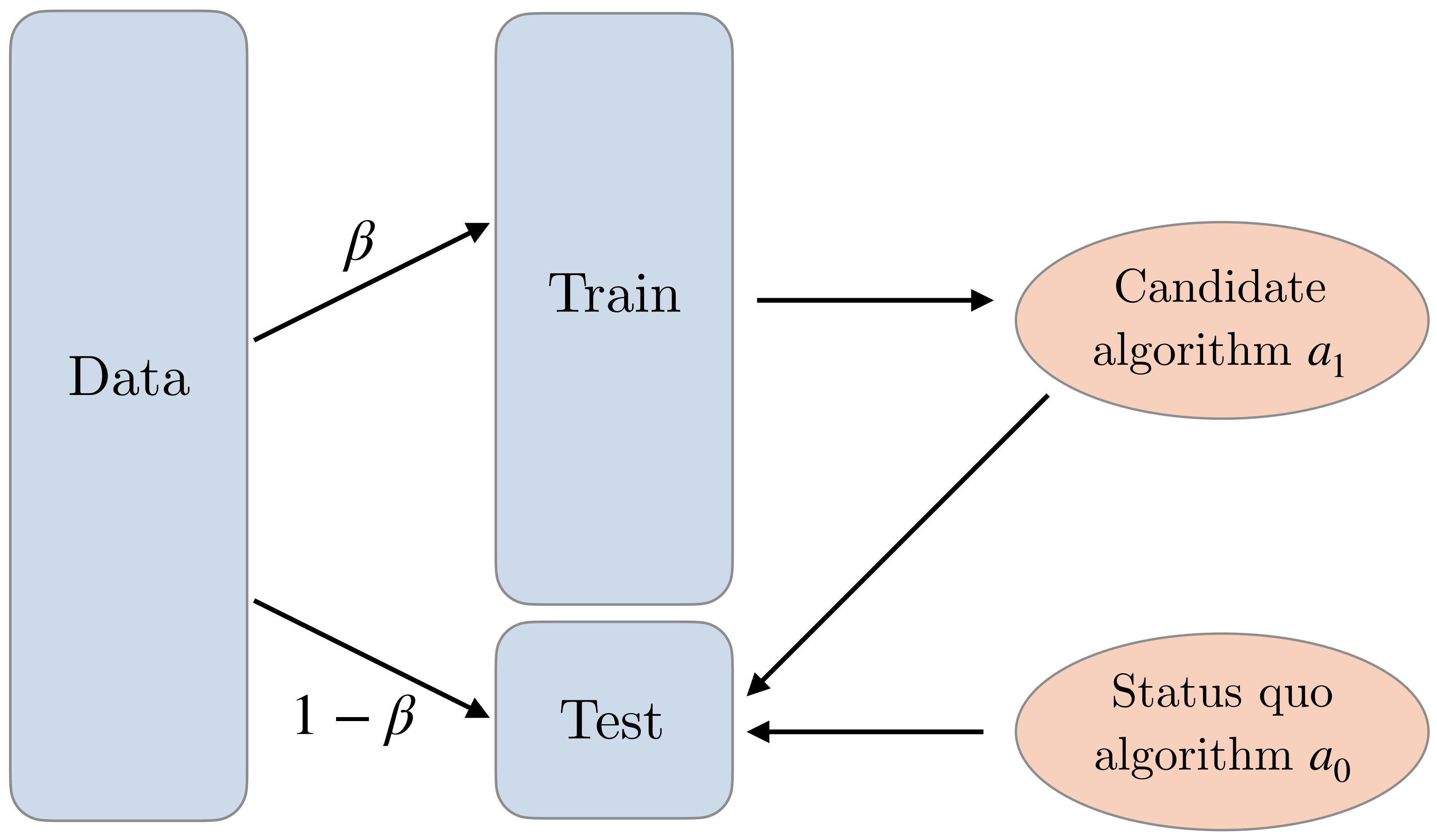
our proposed procedure



$\underbrace{\textbf{Step 3:} \text{ Test whether } a_1 \text{ constitutes an } (\Delta_r, \Delta_b, \Delta_f)\text{-improvement on } a_0}$

plug in $(\Delta_r, \Delta_b, \Delta_f) = (0, 0, \delta)$ or $(\Delta_r, \Delta_b, \Delta_f) = (\delta, \delta, 0)$ depending on which is the desired null

our proposed procedure



Step 4: Repeat steps 1-3 K times, and obtain $(p_1, \ldots, p_K)$.

Define $p = \text{median}\left\{p_1, \ldots, p_K\right\}$ and reject if $p < \dfrac{\alpha}{2}$.

guarantees for this procedure (informal)

recall the null hypothesis

H_0: algorithm a_0 is not δ-fairness (or accuracy) improvable within class $\mathscr{A}$

- under regularity conditions, this procedure is **asymptotically valid**
  - i.e., for any desired guarantee α, the probability of rejecting (under the null) is no more than α (in the limit as the sample grows large)

- when the approach for finding a candidate algorithm is "sufficiently powerful," then the procedure is also **consistent**
  - i.e., if the null is false, then it will be rejected with probability converging to 1 as the sample grows large

empirical application

- we already introduced the Obermeyer et al., (2019) data
  - X is a patient's medical profile
  - G is whether the patient is White or Black
  - Y is the patient's number of active chronic illnesses in the next year
  - D is a decision of whether to automatically enroll the patient in a care management program

- the status quo algorithm is the hospital's algorithm (assigns 3% of patients to care)

- we apply our approach to evaluate the improvability of this algorithm within the class of algorithms $a : \mathscr{X} \to \{0,1\}$ that also enrolls 3% of patients

accuracy and fairness

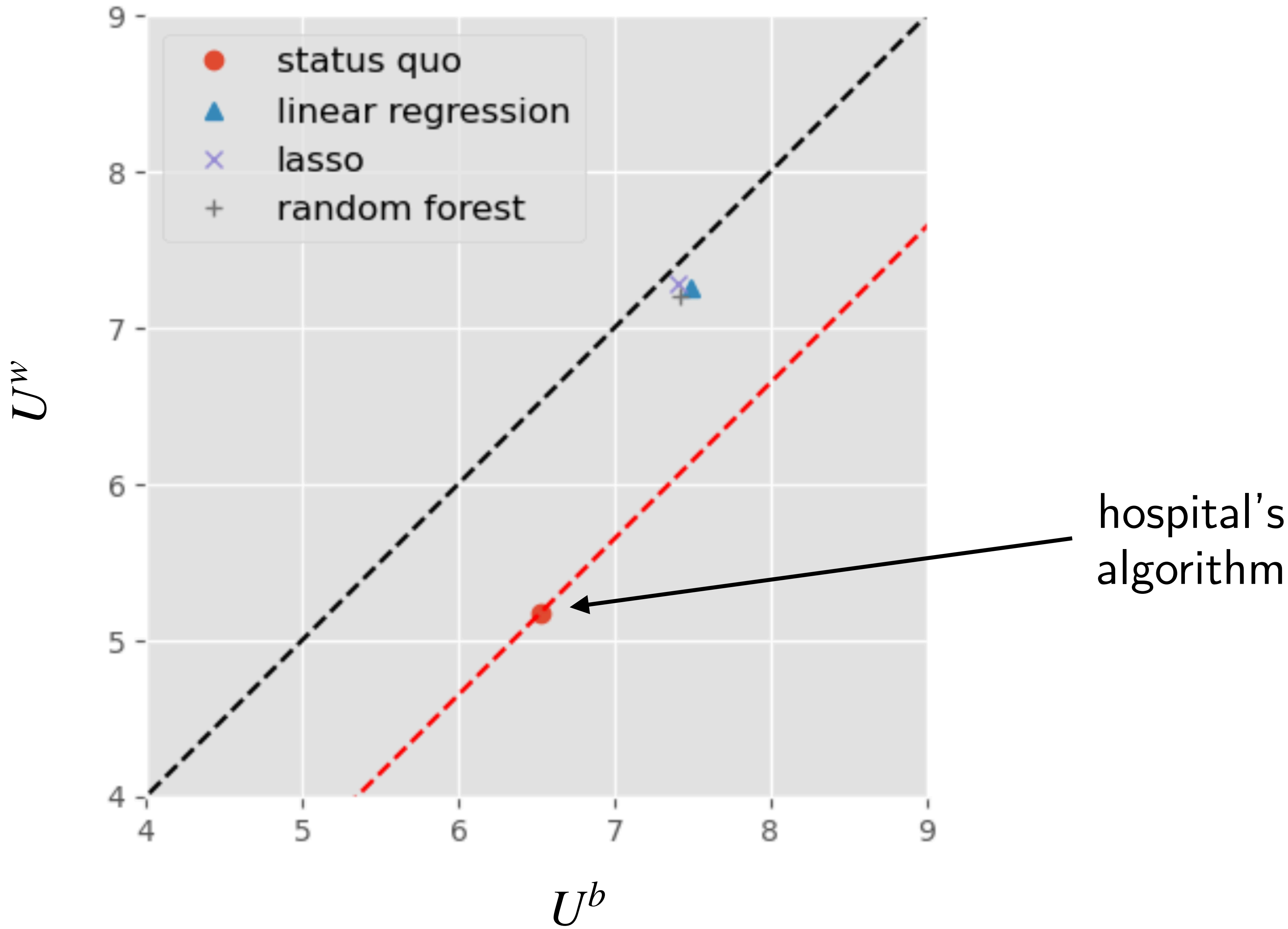
- following Obermeyer et al., (2019), let

$$U_A^g(a) = U_F^g(a) = E[Y \mid a(X) = 1, G = g]$$

  i.e., expected number of illnesses for patients in group g who are assigned to the program

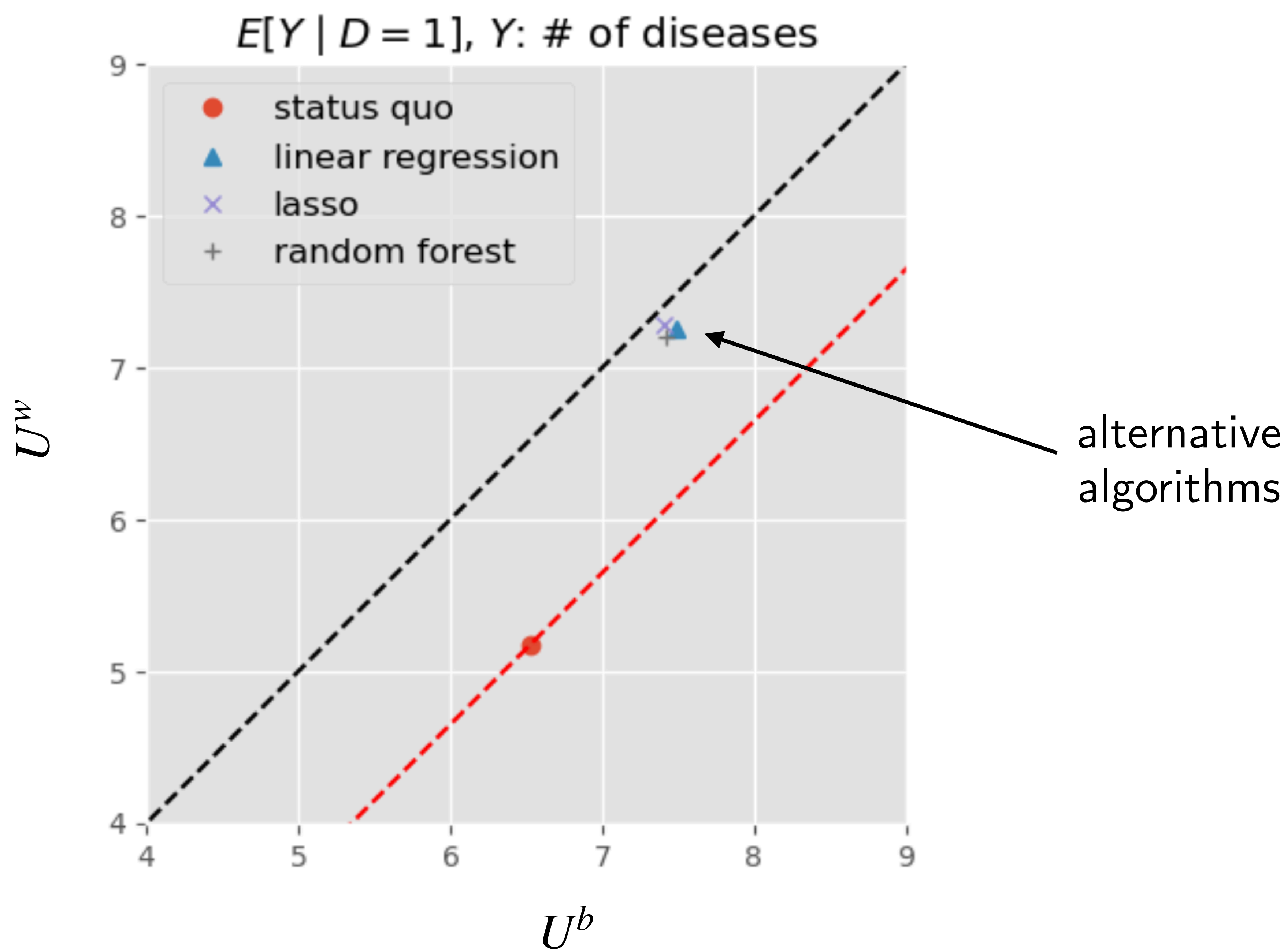
- an algorithm is:
  - **more accurate** if the expected number of health conditions is higher among both Black and White patients assigned to the program
  - **more fair** if it reduces the disparity in the expected number of health conditions among Black and White patients assigned to the program

a first look



status quo
linear regression
lasso
random forest
hospital's algorithm
U^w
U^b
4
5
6
7
8
9

applying our procedure



our procedure yields $p < 0.001 \rightarrow$ reject the null (that the status quo algorithm is not FA-dominated) for $\alpha = 0.05$

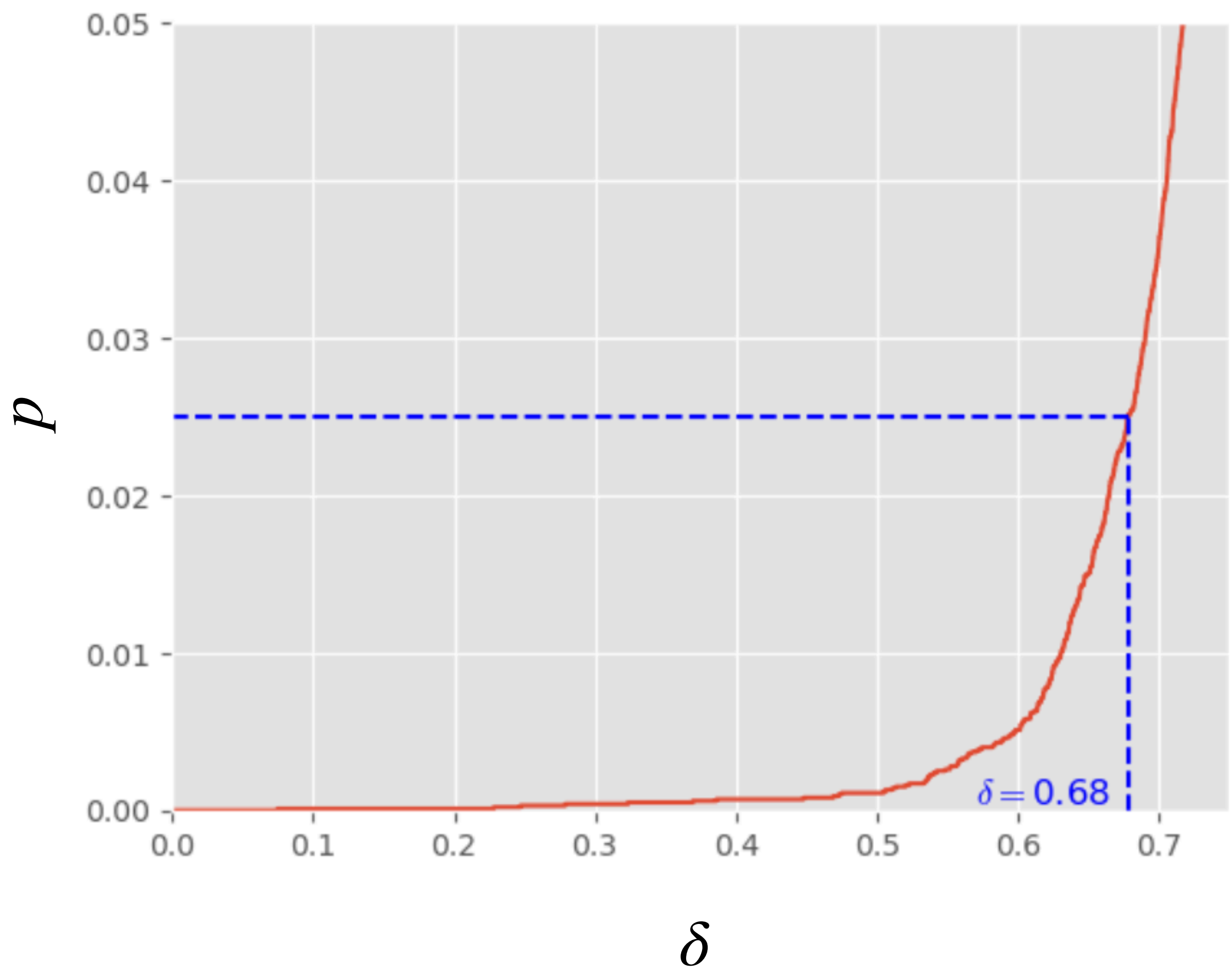
more detail

δ-fairness improvability

can further explore the tradeoff between improvements in accuracy and fairness by subsequently testing for δ-fairness improvability, where we allow δ to vary

- i.e., is it possible to improve on fairness by at least δ percent without compromising on accuracy?

p-value in a test of δ-fairness improvability



can reject the null for all $\delta \leq 0.68$

↓

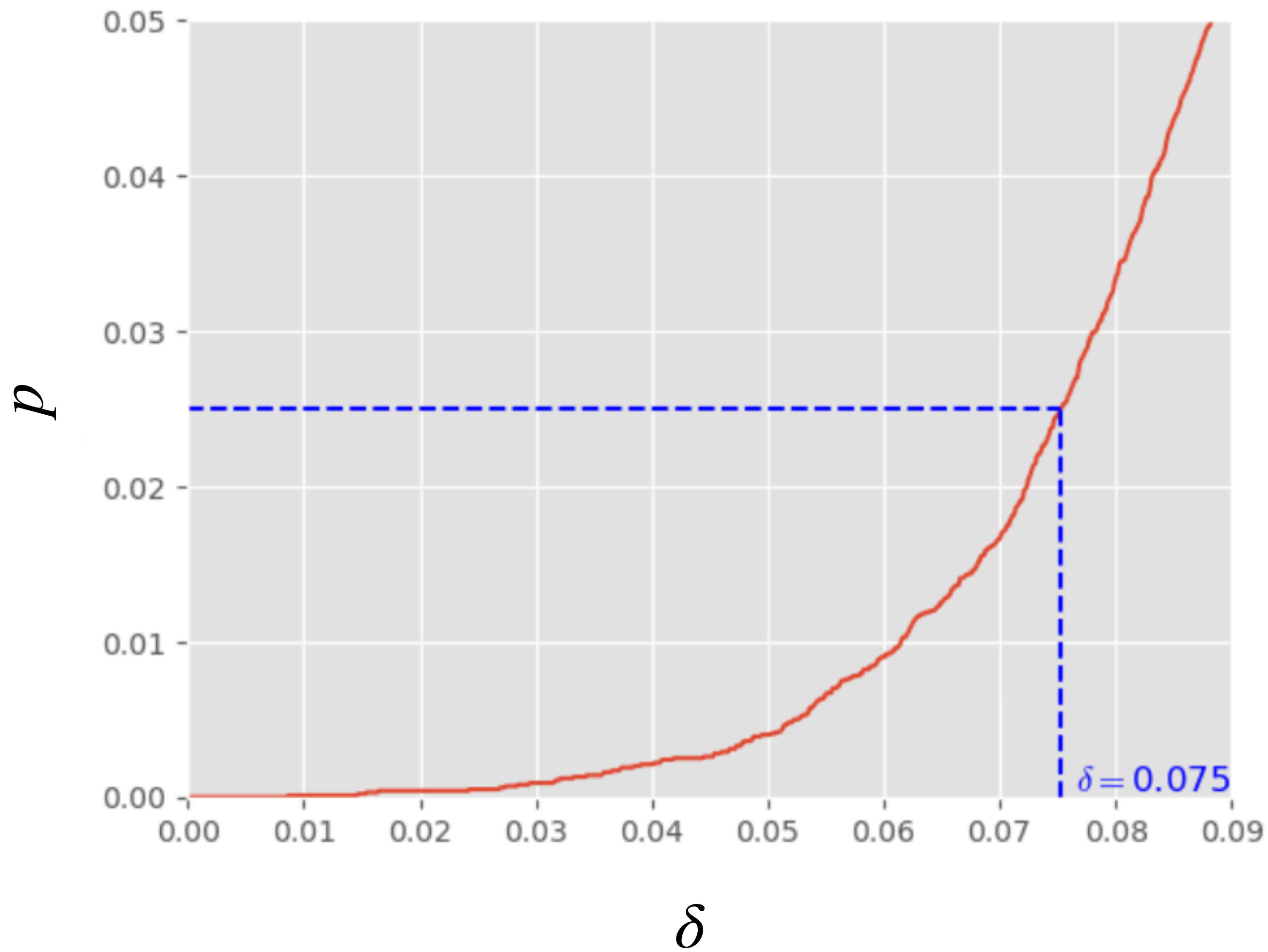
possible to **halve** disparate impact without compromising on accuracy!

δ-accuracy improvability

now conduct same exercise, but for δ-accuracy improvability

- i.e., is it possible to improve on accuracy by at least δ percent for both groups without compromising on fairness?

p-value in a test of δ-accuracy improvability



can only reject the null for $\delta \leq 0.075$

binding dimension turns out to be accuracy for Black patients

takeaways

in this application:

- it is possible to simultaneously strictly improve on the accuracy and the fairness of the status quo algorithm
- large improvements in fairness are possible without compromising on accuracy, while the reverse is not true

Algorithmic Fairness and Social Welfare

Annie Liang
(Northwestern)

Jay Lu
(UCLA)

summary

- the CS literature often formulates fairness metrics similar to the ones we've been looking at so far, or sometimes in the even more stringent form

$$\max \textbf{accuracy}$$
$$\text{subject to } \underbrace{G \perp D}_{\text{demographic parity}}$$

summary

- the CS literature often formulates fairness metrics similar to the ones we've been looking at so far, or sometimes in the even more stringent form

$$\begin{aligned} &\text{max } \textbf{accuracy} \\ &\quad \text{subject to } \underbrace{G \perp D \mid Y}_{\text{equalized odds}} \end{aligned}$$

summary

- the CS literature often formulates fairness metrics similar to the ones we've been looking at so far, or sometimes in the even more stringent form

$$\max \textbf{accuracy} \quad \text{subject to } \textbf{(statistical condition)}$$

summary

- the CS literature often formulates fairness metrics similar to the ones we've been looking at so far, or sometimes in the even more stringent form

$$\text{max } \textbf{accuracy} \quad \text{subject to } \textbf{(statistical condition)}$$

- a long tradition in moral philosophy and economics instead measures social welfare by aggregating across individuals in society
  - fairness considerations stem from contemplating how an individual would choose to structure society prior to the realization of own identity (i.e., "behind the veil")
  - the individual's ex-ante payoffs are $E[\phi(U_i)]$, where U_i is the ex-post utility for an individual with identity i, and the expectation is with respect to randomness in the realization of this identity
  - concave ϕ returns a preference for fairness

summary

- can the CS perspective be motivated as the choices of someone from behind the veil of ignorance?
- we formalize a sense in which the answer is **no**
- does not necessarily suggest that the CS perspective is misguided, but does suggest that novel justifications would be required (open question)

thank you