Algorithm Design: Fairness Versus Accuracy

Annie Liang[1] Jay Lu[2] Xiaosheng Mu[3]

[1]Northwestern

[2]UCLA [3]Princeton

Bravo Center (Brown)

background

algorithms are used to guide many high-stakes decisions

- who should receive a medical treatment?
- who should receive a loan?
- who should receive bail?
- who should receive employment?

background

algorithms are used to guide many high-stakes decisions

- who should receive a medical treatment?
- who should receive a loan?
- who should receive bail?
- who should receive employment?

recent empirical evidence that these algorithms often have errors that vary systematically across subgroups of the population

- patients assigned to same risk score have substantially different actual health risks depending on race (Obermeyer et al., 2019)
- false positive rate of algorithm used to predict criminal reoffense twice as high for Black defendants (Angwin and Larson, 2016)

fairness vs. accuracy

- algorithms increasingly optimized not only for accuracy but also "fairness" (equalizing error rates across groups)
- what is the tradeoff between fairness and accuracy?
- we introduce a "fairness-accuracy frontier" that ranges across a broad class of preferences/optimization criteria
- results characterize how this frontier depends on statistical properties of the inputs to the algorithm
  - whether the inputs reveal the group identity
  - whether the inputs are **group-balanced**
- in paper but will skip in talk: "input design" problem

framework

setup

- single designer and population of (non-strategic) subjects

setup

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

setup

- single designer and population of (non-strategic) subjects
- each subject is described by three variables:
  - **type** Y taking values in $\mathcal{Y}$
    (e.g. need for medical procedure)
  - **group** $G \in \mathcal{G} = \{r, b\}$
    (e.g. race)
  - **covariate** vector X taking values in $\mathcal{X}$
    (e.g. image scans, # past hospital visits, blood tests)
- X is observed by the designer, Y and G are not directly observed (but may be revealed by X)

setup

- single designer and population of (non-strategic) subjects
- each subject is described by three variables:
  - **type** Y taking values in $\mathcal{Y}$
    (e.g. need for medical procedure)
  - **group** $G \in \mathcal{G} = \{r, b\}$
    (e.g. race)
  - **covariate** vector X taking values in $\mathcal{X}$
    (e.g. image scans, # past hospital visits, blood tests)
- X is observed by the designer, Y and G are not directly observed (but may be revealed by X)
- in the population, $(Y, G, X) \sim \mathbb{P}$

algorithm

each subject receives a **decision** $d \in \mathcal{D} = \{0, 1\}$

(e.g. whether the procedure is recommended)

algorithm

each subject receives a **decision** $d \in \mathcal{D} = \{0, 1\}$

(e.g. whether the procedure is recommended)

the designer chooses an **algorithm**

$$a : \mathcal{X} \to \Delta(\mathcal{D})$$

for determining (distributions over) decisions based on the observed covariate vector

group errors

fix a **loss function** $\ell : \mathcal{D} \times \mathcal{Y} \times \mathcal{G} \to \mathbb{R}$

- can interpret ℓ as measure of inaccuracy or as the disutility of a given subject

group errors

fix a **loss function** $\ell : \mathcal{D} \times \mathcal{Y} \times \mathcal{G} \to \mathbb{R}$

- can interpret ℓ as measure of inaccuracy or as the disutility of a given subject

Definition

the **error** for group $g \in \mathcal{G}$ given algorithm a is

$$e_g(a) := \mathbb{E}_{D \sim a(X)} \left[\ell(D, Y, g) \mid G = g \right]$$

i.e., the average/expected loss for subjects in group g

group errors

fix a **loss function** $\ell : \mathcal{D} \times \mathcal{Y} \times \mathcal{G} \to \mathbb{R}$

- can interpret ℓ as measure of inaccuracy or as the disutility of a given subject

Definition

the **error** for group $g \in \mathcal{G}$ given algorithm a is

$$e_g(a) := \mathbb{E}_{D \sim a(X)} \left[\ell(D, Y, g) \mid G = g \right]$$

i.e., the average/expected loss for subjects in group g

- improving **accuracy**: lowering e_r and e_b
- improving **fairness**: lowering $|e_r - e_b|$

special cases

this approach nests several existing fairness metrics:

example: equality of false positive rates corresponds to $e_r(a) = e_b(a)$ with

$$\ell(d, y) = \begin{cases} 1 & \text{if } (d, y) = (1, 0) \\ 0 & \text{otherwise} \end{cases}$$

special cases

this approach nests several existing fairness metrics:

example: equality of false positive rates corresponds to $e_r(a) = e_b(a)$ with

$$\ell(d, y) = \begin{cases} 1 & \text{if } (d, y) = (1, 0) \\ 0 & \text{otherwise} \end{cases}$$

example: algorithm a satisfies **equalized odds** if

$$\mathbb{E}_Y \left[\ \mathbb{E}_X[a(X) \mid G = r, Y] - \mathbb{E}_X[a(X) \mid G = b, Y] \ \right] = 0.$$

special cases

this approach nests several existing fairness metrics:

example: equality of false positive rates corresponds to $e_r(a) = e_b(a)$ with

$$\ell(d, y) = \begin{cases} 1 & \text{if } (d, y) = (1, 0) \\ 0 & \text{otherwise} \end{cases}$$

example: algorithm a satisfies **equalized odds** if

$$\mathbb{E}_Y \left[\; \mathbb{E}_X[a(X) \mid G = r, Y] - \mathbb{E}_X[a(X) \mid G = b, Y] \; \right] = 0.$$

this corresponds to $e_r(a) = e_b(a)$ with

$$\ell(d, y, g) = \begin{cases} \dfrac{P(Y = y)}{P(Y = y \mid G = g)} & \text{if } d = 1 \\ 0 & \text{otherwise} \end{cases}$$

preferences

Definition (fairness-accuracy (FA) dominance)

let $>_{FA}$ be the partial order on $\mathbb{R}^2$ satisfying $(e_r, e_b) >_{FA} (e'_r, e'_b)$ if

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

with at least one of these inequalities strict

preferences

Definition (fairness-accuracy (FA) dominance)

let $>_{FA}$ be the partial order on $\mathbb{R}^2$ satisfying $(e_r, e_b) >_{FA} (e'_r, e'_b)$ if

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

with at least one of these inequalities strict

Definition

a **fairness-accuracy preference** $\succeq$ is any total order on $\mathbb{R}^2$ such that $e \succ e'$ whenever $e >_{FA} e'$

examples of FA preferences

1. **utilitarian/bayes-optimal:**

   $$w_u(e_r, e_b) = -p_r e_r - p_b e_b$$

   where p_r and p_b are the proportions of either group

examples of FA preferences

1. **utilitarian/bayes-optimal:**

   $$w_u(e_r, e_b) = -p_r e_r - p_b e_b$$

   where p_r and p_b are the proportions of either group

2. **egalitarian:** first order errors by $-|e_r - e_b|$, and then break ties using w_u

examples of FA preferences

1. **utilitarian/bayes-optimal:**
   $$w_u(e_r, e_b) = -p_r e_r - p_b e_b$$
   where p_r and p_b are the proportions of either group

2. **egalitarian:** first order errors by $-|e_r - e_b|$, and then break ties using w_u

3. **rawlsian/group DRO**: first order errors by $-\max\{e_r, e_b\}$, and then break ties using w_u

examples of FA preferences

1. **utilitarian/bayes-optimal:**

   $$w_u(e_r, e_b) = -p_r e_r - p_b e_b$$

   where p_r and p_b are the proportions of either group

2. **egalitarian:** first order errors by $-|e_r - e_b|$, and then break ties using w_u

3. **rawlsian/group DRO**: first order errors by $-\max\{e_r, e_b\}$, and then break ties using w_u

4. **constrained optimization** (e.g., Hardt et al., 2016):

   $$\min_{a \in \mathcal{A}_X} \quad p_r e_r(a) + p_b e_b(a) \quad \text{s.t. } |e_r(a) - e_b(a)| \leq \varepsilon$$

fairness-accuracy frontier

Definition

the **feasible set** given X is

$$\mathcal{E}(X) := \{(e_r(a), e_b(a)) : a \in \mathcal{A}_X\}$$

where $\mathcal{A}_X$ is the set of all algorithms $a : \mathcal{X} \to \Delta(\mathcal{D})$

fairness-accuracy frontier

Definition

the **feasible set** given X is

$$\mathcal{E}\left(X\right) := \left\{\left(e_r\left(a\right), e_b(a)\right) : \ a \in \mathcal{A}_X\right\}$$

where $\mathcal{A}_X$ is the set of all algorithms $a : \mathcal{X} \to \Delta(\mathcal{D})$

Definition

the **fairness-accuracy frontier** given X is

$$\mathcal{F}\left(X\right) := \left\{e \in \mathcal{E}\left(X\right) : \nexists \ e' \in \mathcal{E}\left(X\right) \text{ s.t. } e' >_{FA} e\right\}$$

- includes optimal points across all fairness-accuracy preferences

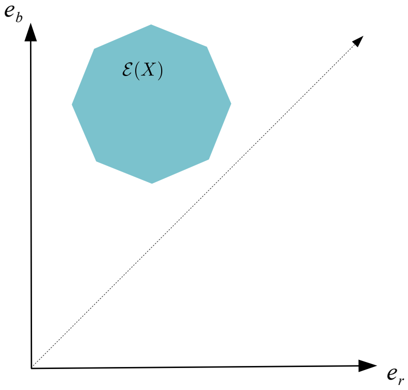
characterizing the

fairness-accuracy frontier

feasible set of group error pairs

lemma: for any X, the feasible set $\mathcal{E}(X)$ is compact and convex

(if $\mathcal{X}$ is finite, it is a convex polygon)



important points

group-optimal points:

$$R_X := \arg\min_{e \in \mathcal{E}(X)} e_r \qquad B_X := \arg\min_{e \in \mathcal{E}(X)} e_b$$

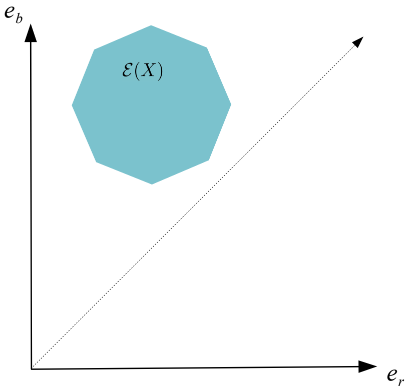
fairness-maximizing point:

$$F_X := \arg\min_{e \in \mathcal{E}(X)} |e_r - e_b|$$

(break all ties in favor of aggregate accuracy)

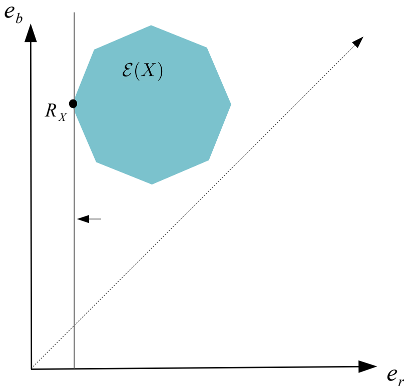
important points

easy to locate geometrically:



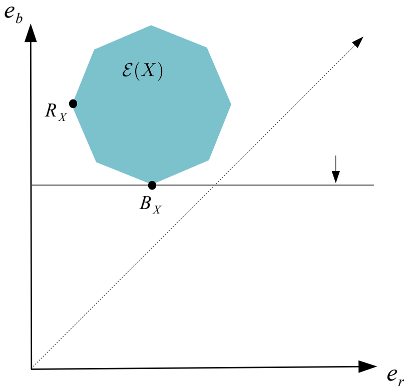
important points

easy to locate geometrically:



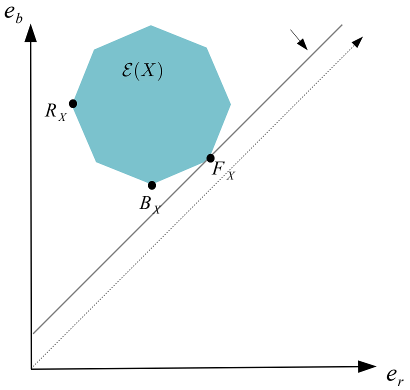
important points

easy to locate geometrically:



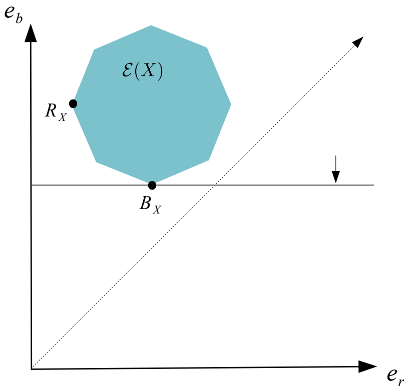
important points

easy to locate geometrically:



important points

easy to locate geometrically:



group-skewed vs group-balanced

Definition

covariate vector X is

- r-**skewed** if $e_r < e_b$ at R_X and $e_r \leq e_b$ at B_X

  "group r's error is lower both at group r's favorite point and also at group b's favorite point"

- b-**skewed** if $e_b < e_r$ at B_X and $e_b \leq e_r$ at R_X
- **group-balanced** otherwise

group-skewed vs group-balanced

Definition

covariate vector X is

- r-**skewed** if $e_r < e_b$ at R_X and $e_r \leq e_b$ at B_X

  "group r's error is lower both at group r's favorite point and also at group b's favorite point"

- b-**skewed** if $e_b < e_r$ at B_X and $e_b \leq e_r$ at R_X
- **group-balanced** otherwise

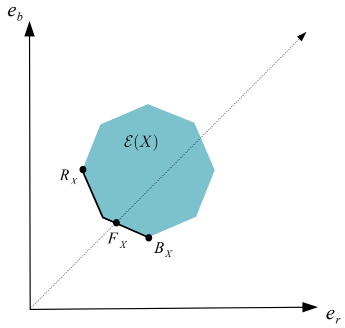
group-skew can emerge in practice (for example) if the inputs in X are systematically more informative about Y for one group

characterization of fairness-accuracy frontier

Theorem

$\mathcal{F}(X)$ is lower boundary of $\mathcal{E}(X)$ between

- *R_X and B_X if X is group-balanced*

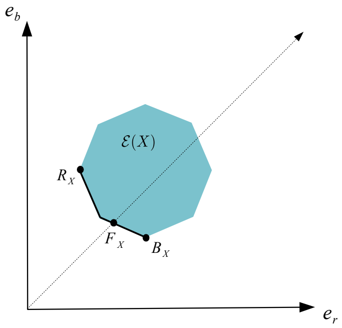


characterization of fairness-accuracy frontier

Theorem

$\mathcal{F}(X)$ is lower boundary of $\mathcal{E}(X)$ between

- *R_X and B_X if X is group-balanced* (= usual Pareto frontier!)

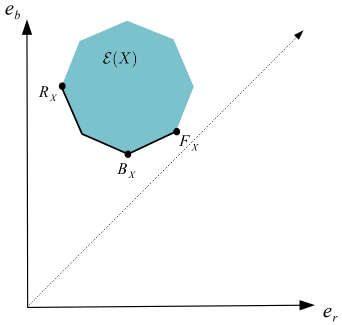


characterization of fairness-accuracy frontier

Theorem

$\mathcal{F}(X)$ is lower boundary of $\mathcal{E}(X)$ between

- *R_X and B_X if X is group-balanced*
- *G_X and F_X if X is g-skewed*

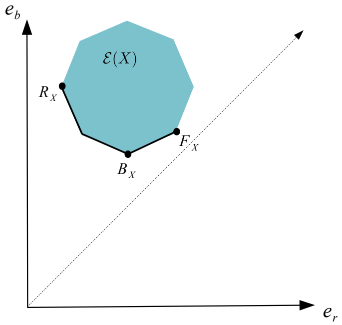


characterization of fairness-accuracy frontier

Theorem

$\mathcal{F}(X)$ is lower boundary of $\mathcal{E}(X)$ between

- *R_X and B_X if X is group-balanced*
- *G_X and F_X if X is g-skewed* (usual Pareto frontier + more)



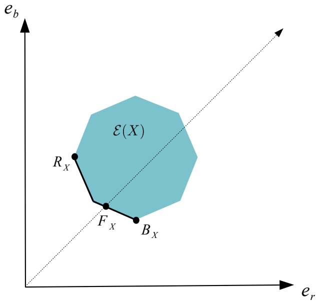
strong fairness-accuracy conflict

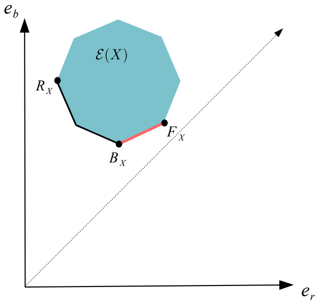
corollary: there exist $e, e' \in \mathcal{F}(X)$ satisfying

$$e_r \leq e'_r, \quad e_b \leq e'_b, \quad |e_r - e_b| > |e'_r - e'_b|$$

e.g., $e = (1/3, 1/4)$, $e' = (1/2, 1/2)$

if and only if X is group-skewed





POLICYMAKER

"I have a policy proposal, which would decrease accuracy for both groups, but increase fairness."

ACADEMIC

"Are the inputs to your algorithm group-balanced?"

POLICYMAKER

"I have a policy proposal, which would decrease accuracy for both groups, but increase fairness."

ACADEMIC

"Are the inputs to your algorithm group-balanced?"

POLICYMAKER

"**Yes**, they are **group-balanced**."

POLICYMAKER

"I have a policy proposal, which would decrease accuracy for both groups, but increase fairness."

ACADEMIC

"Are the inputs to your algorithm group-balanced?"

POLICYMAKER

"**Yes**, they are **group-balanced**."

ACADEMIC

"Your proposal is not optimal for you by your own preferences, **regardless** of how you tradeoff fairness and accuracy."

POLICYMAKER

“I have a policy proposal, which would decrease accuracy for both groups, but increase fairness.”

ACADEMIC

“Are the inputs to your algorithm group-balanced?”

POLICYMAKER

“**No**, they are **group-skewed**.”

POLICYMAKER

"I have a policy proposal, which would decrease accuracy for both groups, but increase fairness."

ACADEMIC

"Are the inputs to your algorithm group-balanced?"

POLICYMAKER

"**No**, they are **group-skewed**."

ACADEMIC

"If you care sufficiently about fairness relative to accuracy, then your proposal **may be optimal** for your goals."

which of group balance and group skew is more common?

difficult to anticipate without an empirical analysis

which of group balance and group skew is more common?

difficult to anticipate without an empirical analysis

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

- suppose X has a group-dependent meanings
- e.g., frequent moves signal high creditworthiness for high-income borrowers but low creditworthiness for low-income borrowers
- maximizing accuracy for the high-income group leads this group to have the lower error (and vice versa)

which of group balance and group skew is more common?

difficult to anticipate without an empirical analysis

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

- suppose X has a group-dependent meanings
- e.g., frequent moves signal high creditworthiness for high-income borrowers but low creditworthiness for low-income borrowers
- maximizing accuracy for the high-income group leads this group to have the lower error (and vice versa)

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

- suppose X is asymmetrically informative
- e.g., medical data is recorded more accurately for high-income patients than low-income patients
- best algorithm coincides for both groups and implies a lower error for high-income patients

generalizations

beyond absolute difference

- results extend when unfairness is measured as $|\phi(e_r) - \phi(e_b)|$ where ϕ is some continuous strictly increasing function
- if ϕ is log, then this corresponds to evaluating fairness using the ratio of errors rather than their difference

generalizations

beyond absolute difference

- results extend when unfairness is measured as $|\phi(e_r) - \phi(e_b)|$ where ϕ is some continuous strictly increasing function
- if ϕ is log, then this corresponds to evaluating fairness using the ratio of errors rather than their difference

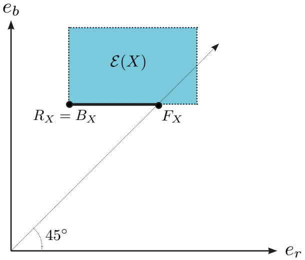
different loss functions for evaluating fairness and accuracy

- qualitative result extends whenever the two loss functions aren't "directly opposed"
- group-balance generalizes to whether F_X belongs to usual Pareto frontier
  - when this condition is satisfied, then the fairness-accuracy frontier is identical to the usual Pareto frontier
  - otherwise, the fairness-accuracy frontier is the union of the Pareto frontier and a positively-sloped sequence of lines

special case: when group identity is an input

Proposition

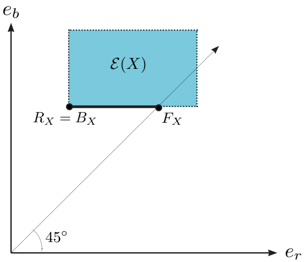
if G is an input in X, then $\mathcal{E}(X)$ is a rectangle with sides parallel to axes and $\mathcal{F}(X)$ is the line segment from $R_X = B_X$ to F_X



special case: when group identity is an input

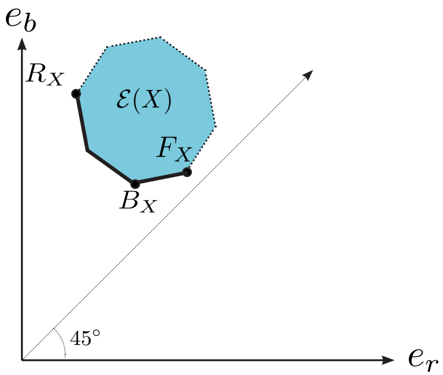
Proposition

if G is an input in X, then $\mathcal{E}(X)$ is a rectangle with sides parallel to axes and $\mathcal{F}(X)$ is the line segment from $R_X = B_X$ to F_X

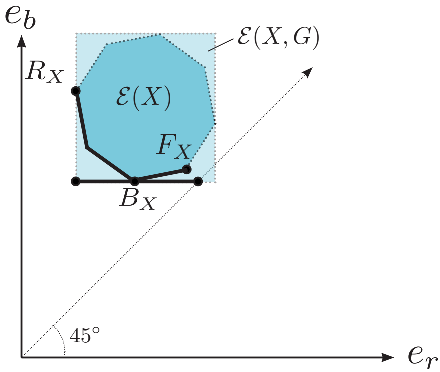


- frontier is **rawlsian**: worse-off group gets best feasible error (no matter which optimization problem we solve)

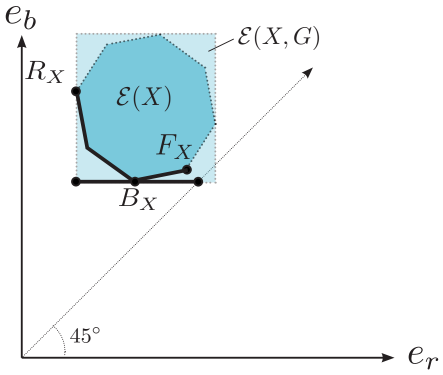
what happens when G is added as an input?



what happens when G is added as an input?



what happens when G is added as an input?



result (informal): for any designer preference (in our permitted class), access to G reduces the error for the worse-off group.

- not true for the other group

in paper: input design

- sometimes the algorithm is set by an agent cares only about accuracy, inputs are constrained by a regulator

in paper: input design

- sometimes the algorithm is set by an agent cares only about accuracy, inputs are constrained by a regulator
- we formulate a problem of "input design":
  - designer chooses a garbling of the covariates
  - decision-maker chooses an algorithm (based on this garbling) to maximize accuracy

in paper: input design

- sometimes the algorithm is set by an agent cares only about accuracy, inputs are constrained by a regulator
- we formulate a problem of "input design":
  - designer chooses a garbling of the covariates
  - decision-maker chooses an algorithm (based on this garbling) to maximize accuracy
- how limiting is this for the designer? are there garblings that can implement the designer's favorite (unconstrained) point?

in paper: input design

- sometimes the algorithm is set by an agent cares only about accuracy, inputs are constrained by a regulator
- we formulate a problem of "input design":
  - designer chooses a garbling of the covariates
  - decision-maker chooses an algorithm (based on this garbling) to maximize accuracy
- how limiting is this for the designer? are there garblings that can implement the designer's favorite (unconstrained) point?
- ask whether the optimal garbling could involve completely banning a covariate (such as group identity)
  - if X is group-balanced, then banning group identity make every designer strictly worse off

related literature

huge literature on algorithmic fairness in CS

- Dwork et. al (2012), Hardt et al. (2016), Kleinberg et al. (2017), Chouldechova (2017), Roth and Kearns (2019), and many more

related literature

huge literature on algorithmic fairness in CS

- Dwork et. al (2012), Hardt et al. (2016), Kleinberg et al. (2017), Chouldechova (2017), Roth and Kearns (2019), and many more

key differences/contributions

1. characterize the fairness-accuracy frontier (in contrast to focusing on a specific optimization problem)

related literature

huge literature on algorithmic fairness in CS

- Dwork et. al (2012), Hardt et al. (2016), Kleinberg et al. (2017), Chouldechova (2017), Roth and Kearns (2019), and many more

key differences/contributions

1. characterize the fairness-accuracy frontier (in contrast to focusing on a specific optimization problem)
2. formulate problem of choosing inputs as one of "information design" (final input-design section)
    - Kamenica & Gentzkow (2011), Bergemann & Morris (2019)

related literature

huge literature on algorithmic fairness in CS

- Dwork et. al (2012), Hardt et al. (2016), Kleinberg et al. (2017), Chouldechova (2017), Roth and Kearns (2019), and many more

key differences/contributions

1. characterize the fairness-accuracy frontier (in contrast to focusing on a specific optimization problem)
2. formulate problem of choosing inputs as one of "information design" (final input-design section)
    - Kamenica & Gentzkow (2011), Bergemann & Morris (2019)

recent empirical work in economics:

- Obermeyer, Powers, Vogeli, Mullainathan (2019), Arnold, Dobbie, and Hull (2021), Fuster, Goldsmith-Pinkham, Ramadorai, Walther (2021)

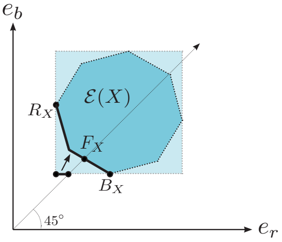
Conclusion

- framework for evaluating the accuracy/fairness tradeoffs of algorithms
- characterized the fairness-accuracy frontier over different designer preferences for how to trade off these criteria
- explained how certain statistical properties of the algorithm's inputs impact the shape of this frontier
- in some cases (e.g., when the inputs are group-balanced), there are conclusions/policy recommmendations that hold **for all** designer preferences in a broad class

thank you

compare X to (X, G)

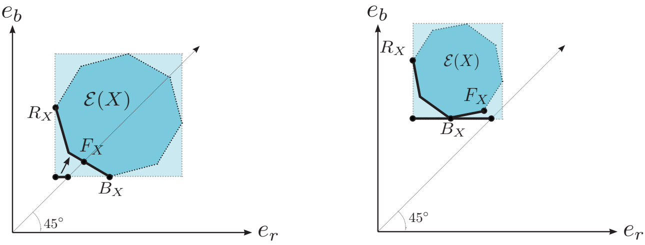
result (informal): banning G uniformly worsens the Pareto frontier if and only if X is group-balanced.



(a) group-balanced X

compare X to (X, G)

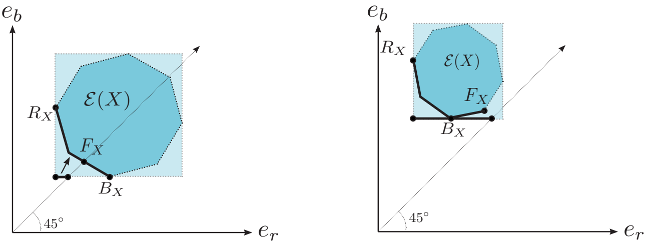
result (informal): banning G uniformly worsens the Pareto frontier if and only if X is group-balanced.



(a) group-balanced X (b) group-skewed X

compare X to (X, G)

result (informal): banning G uniformly worsens the Pareto frontier if and only if X is group-balanced.



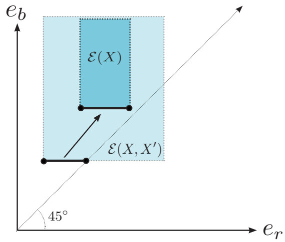
(a) group-balanced X (b) group-skewed X

takeaways:

- when X is group-balanced, all designers strictly benefit from allowing the algorithm to condition on G
- this is true even for an Egalitarian designer: disparate treatment may be necessary to remove disparate impact

compare X to (X, X') when X reveals G

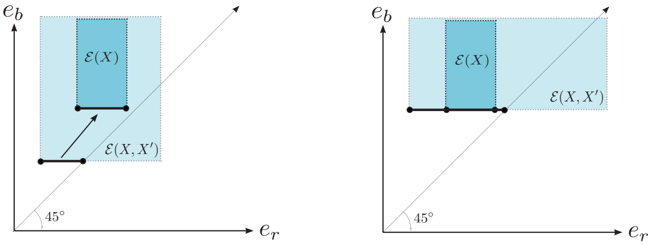
result (informal): banning X' uniformly worsens the Pareto frontier if and only if X' reduces the error for the worse-off group.



(a) X' reduces group b's error

compare X to (X, X') when X reveals G

result (informal): banning X' uniformly worsens the Pareto frontier if and only if X' reduces the error for the worse-off group.

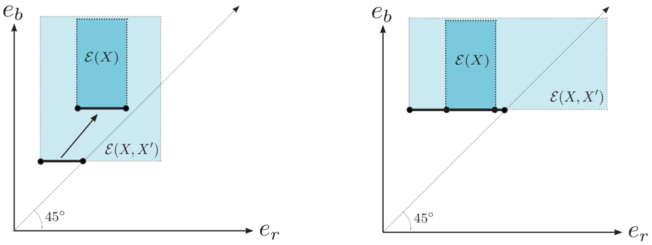


(a) X' reduces group b's error

(b) X' does not reduce group b's error

compare X to (X, X') when X reveals G

result (informal): banning X' uniformly worsens the Pareto frontier if and only if X' reduces the error for the worse-off group.



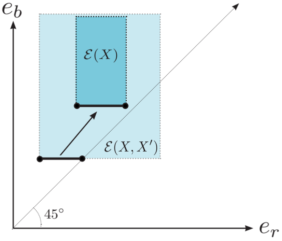
(a) X' reduces group b's error

(b) X' does not reduce group b's error

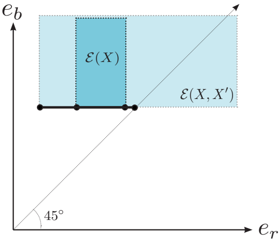
takeaway: active policy debate regarding whether to ban test scores in admissions decisions.

compare X to (X, X') when X reveals G

result (informal): banning X' uniformly worsens the Pareto frontier if and only if X' reduces the error for the worse-off group.



(a) X' reduces group b's error

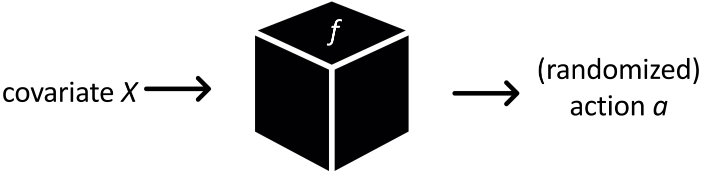


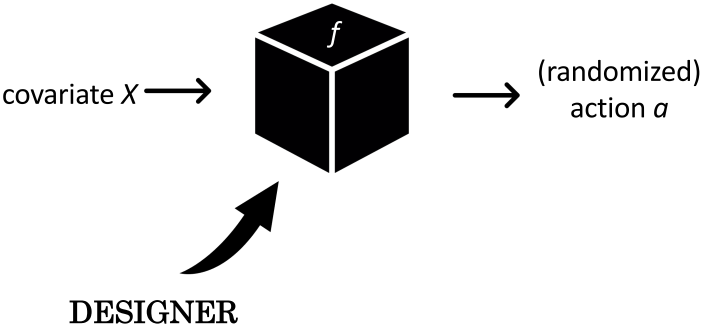
(b) X' does not reduce group b's error

takeaway: active policy debate regarding whether to ban test scores in admissions decisions.

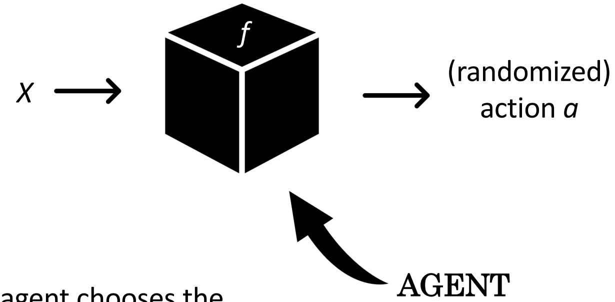
- so long as G is permissible, then excluding test scores makes all designers worse off
- if G is not a permitted input (e.g., California), then it can be strictly optimal to ban X' (see example in paper)

what happens when the designer only controls the inputs?


covariate X
f
(randomized)
action a

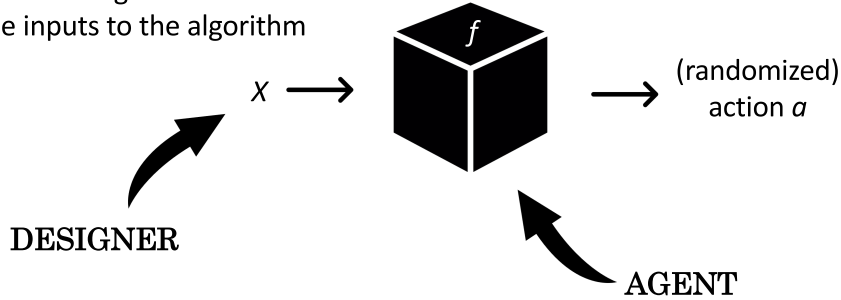


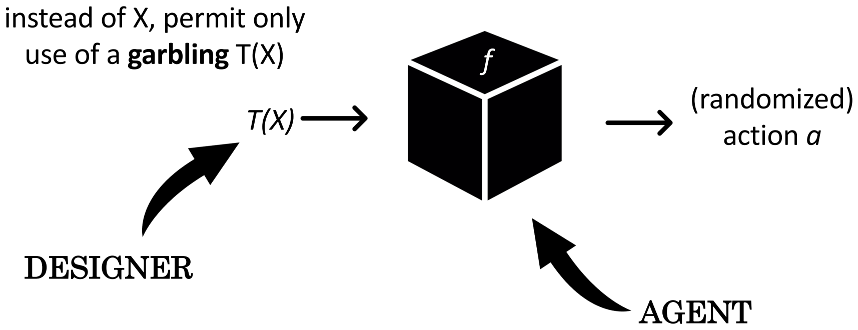
the designer chooses
the algorithm



another agent chooses the algorithm

but the designer can control the inputs to the algorithm





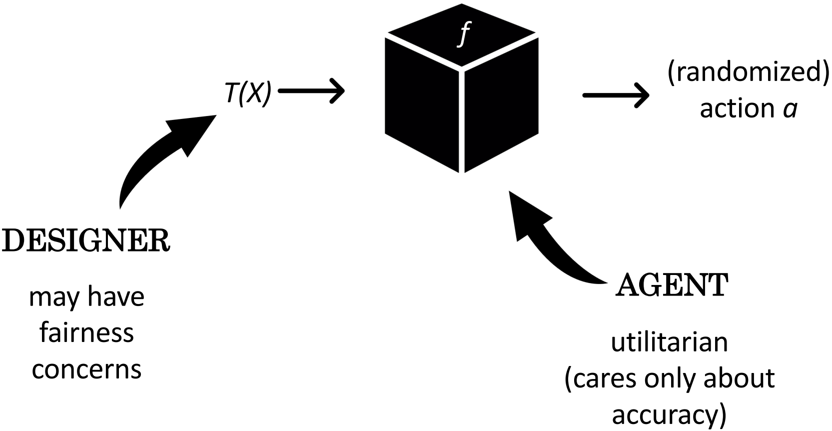
instead of X, permit only use of a **garbling** T(X)
T(X)
f
(randomized) action *a*
DESIGNER
AGENT

example: drop an input

example: coarsening

Boalt Hall, UC Berkeley's law school, attracted national media attention in 1997 when its entering class of 268 included only one African-American.[17] The following year, administrators made a number of changes to their admissions policy. The new policy no longer assigns candidates Academic Index Scores—previously a function of undergraduate GPA (weighted by the quality of the candidate's undergraduate institution) and LSAT score. Indeed, it no longer adjusts candidates' GPAs to account for the quality of their undergraduate institutions. Nor does it consider candidates' exact LSAT scores; instead, LSAT scores are partitioned into intervals, and the admissions committee only learns which interval contains the candidate's score.

(Chan and Eyster, 2003)



f
T(X)
(randomized)
action a
DESIGNER
may have
fairness
concerns
AGENT
utilitarian
(cares only about
accuracy)

input design: feasible and pareto sets

let f_T denote the utilitarian-optimal algorithm given T

Definition

the **feasible set** under **input design** given X is

$$\mathcal{E}^*(X) := \{e(f_T) \ : \ T \text{ is a garbling of } X\}$$

input design: feasible and pareto sets

let f_T denote the utilitarian-optimal algorithm given T

Definition

the **feasible set** under **input design** given X is

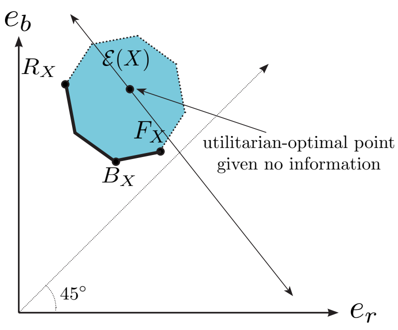
$$\mathcal{E}^*(X) := \{e(f_T) \ : \ T \text{ is a garbling of } X\}$$

the **fairness-accuracy frontier** under **input design** given X is

$$\mathcal{F}^*(X) := \left\{ e \in \mathcal{E}^*(X) \ : \ \underbrace{\text{no } e' \in \mathcal{E}^*(X) \text{ s.t. } e' \succ_{FA} e}_{e \text{ is FA-undominated in } \mathcal{E}^*(X)} \right\}$$

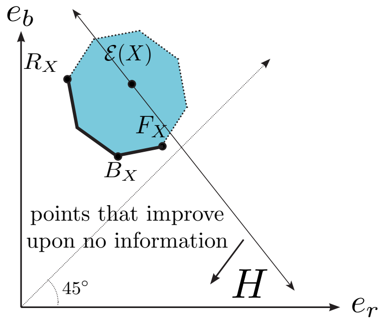
how powerful is input design?

- let e_0 be the utilitarian's best payoff given no information



how powerful is input design?

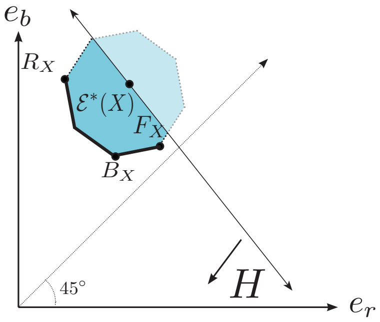
- let e_0 be the utilitarian's best payoff given no information
- define $H := \{(e_r, e_b) : p_r e_r + p_b e_b \leq e_0\}$



how powerful is input design?

- let e_0 be the utilitarian's best payoff given no information
- define $H := \{(e_r, e_b) : p_r e_r + p_b e_b \leq e_0\}$
- **lemma:** $\mathcal{E}^*(X) = \mathcal{E}(X) \cap H$
  (see also Alonso and Camâra, 2016)

  $\Longrightarrow$ any point that is feasible given X and in the halfspace H can be implemented using some garbling of X

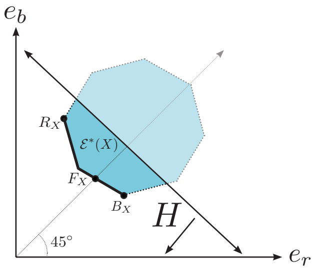


how powerful are informational constraints?

Proposition

(a) *If X is group-balanced, then $\mathcal{F}(X) = \mathcal{F}^*(X)$ iff $R_X, B_X \in H$*

(b) *If X is g-skewed, then $\mathcal{F}(X) = \mathcal{F}^*(X)$ iff $G_X, F_X \in H$*



takeaway: under weak conditions, designer can implement favorite (unconstrained) outcome by designing the algorithmic inputs will the designer want to exclude inputs?

add/ban covariates?

- regulatory question: should certain inputs be banned?
  - some group identities are already banned (e.g. race, religion for hiring or bank loans)
  - other covariates increasingly prohibited due to fairness concerns (e.g. universities excluding test scores)

add/ban covariates?

- regulatory question: should certain inputs be banned?
  - some group identities are already banned (e.g. race, religion for hiring or bank loans)
  - other covariates increasingly prohibited due to fairness concerns (e.g. universities excluding test scores)

- we can study this using our framework
  - X is the permitted part, X' is the input in question

add/ban covariates?

- regulatory question: should certain inputs be banned?
  - some group identities are already banned (e.g. race, religion for hiring or bank loans)
  - other covariates increasingly prohibited due to fairness concerns (e.g. universities excluding test scores)

- we can study this using our framework
  - X is the permitted part, X' is the input in question

- **question:** how does the input-design frontier $\mathcal{F}^*(X)$ compare to $\mathcal{F}^*(X, X')$?

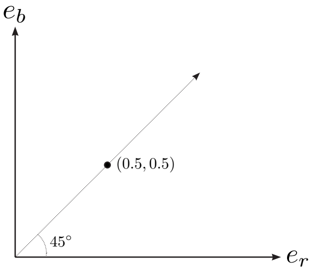
excluding X' can be strictly optimal

- $Y \in \{0, 1\}$ with $P(Y = 1 \mid G = g) = 1/2$ for both groups g
- X is a null signal, while $X' \in \{0, 1\}$ is a binary signal where
  - $X' = Y$ with probability 1 if $G = r$
  - $X' = Y$ with probability 0.6 if $G = b$

  so X is more informative about type for group r
- the designer is Egalitarian (payoff is $-|e_r - e_b|$)

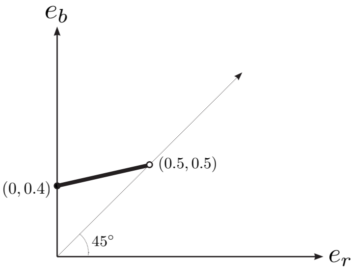
excluding X' can be strictly optimal

- sending the null signal X leads to a payoff of $|0.5 - 0.5| = 0$.



excluding X' can be strictly optimal

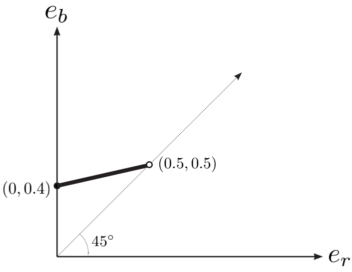
- sending the null signal X leads to a payoff of $|0.5 - 0.5| = 0$.



- any information provided about X' will be used by the agent to improve accuracy
- but this information decreases r's error more than b's error, contributing to a larger gap

excluding X' can be strictly optimal

- sending the null signal X leads to a payoff of $|0.5 - 0.5| = 0$.



- any information provided about X' will be used by the agent to improve accuracy
- but this information decreases r's error more than b's error, contributing to a larger gap
- the designer's payoffs are strictly negative when any information about X' is provided to the agent

uniform worsening of the frontier

at the other extreme...

Definition

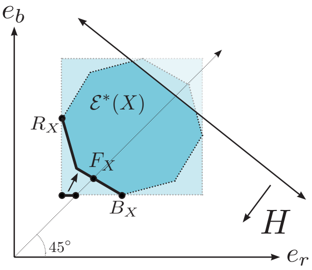
say that excluding X' given X **uniformly worsens the frontier** if every point in $\mathcal{F}^*(X)$ is FA-dominated by a point in $\mathcal{F}^*(X, X')$

- $\hookrightarrow$ every designer strictly prefers to send information about X'
- $\hookrightarrow$ excluding X' cannot be justified by a fairness-accuracy preference

excluding group identity

first compare X to (X, G)

result: suppose $R_X, B_X \in H$. excluding G uniformly worsens the frontier if and only if X is group-balanced

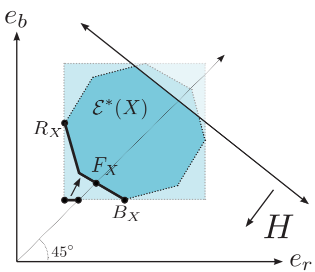


(a) group-balanced X

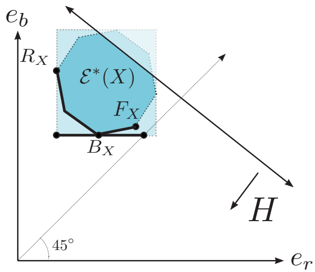
excluding group identity

first compare X to (X, G)

result: suppose $R_X, B_X \in H$. excluding G uniformly worsens the frontier if and only if X is group-balanced



(a) group-balanced X



(b) group-skewed X

takeaways

- when X is group-balanced, all designers benefit from sending some information about G
- conditioning on G means applying an information policy that is asymmetric across groups
- result suggests that **disparate treatment** may be necessary to preclude **disparate impact**
- echos previous findings in the statistical discrimination literature (e.g., Chan and Eyster, 2003)

excluding a covariate when group identity is known

compare X to (X, X') when X reveals G

excluding a covariate when group identity is known

compare X to (X, X') when X reveals G

definition: say that X' is **decision-relevant over X for group g** if there are realizations (x, x') and $(x, \tilde{x}')$ of (X, X') where

$$\{1\} = \operatorname*{argmin}_{d \in \mathcal{D}} \mathbb{E}[\ell(a, Y, g) \mid X = x, X' = x', G = g]$$

while

$$\{0\} = \operatorname*{argmin}_{d \in \mathcal{D}} \mathbb{E}[\ell(a, Y, g) \mid X = x, X' = \tilde{x}', G = g]$$

excluding a covariate when group identity is known

compare X to (X, X') when X reveals G

definition: say that X' is **decision-relevant over X for group g** if there are realizations (x, x') and $(x, \tilde{x}')$ of (X, X') where

$$\{1\} = \operatorname*{argmin}_{d \in \mathcal{D}} \mathbb{E}[\ell(a, Y, g) \mid X = x, X' = x', G = g]$$

while

$$\{0\} = \operatorname*{argmin}_{d \in \mathcal{D}} \mathbb{E}[\ell(a, Y, g) \mid X = x, X' = \tilde{x}', G = g]$$

this is a weak condition:

- says only that the additional information in X' can change the optimal assignment for some individual in group g

excluding a covariate when group identity is known

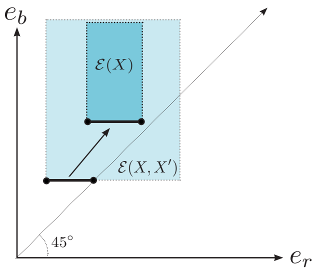
result: suppose X reveals G, and let X be g-skewed.

excluding X' given X uniformly worsens the frontier $\iff X'$ is decision-relevant over X for group $g' \neq g$.

excluding a covariate when group identity is known

result: suppose X reveals G, and let X be g-skewed.

excluding X' given X uniformly worsens the frontier $\iff$ X' is decision-relevant over X for group $g' \neq g$.

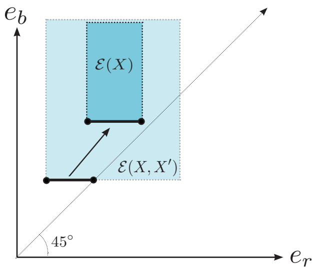


(a) X' reduces group b's error

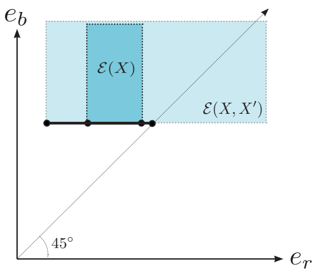
excluding a covariate when group identity is known

result: suppose X reveals G, and let X be g-skewed.

excluding X' given X uniformly worsens the frontier $\iff$ X' is decision-relevant over X for group $g' \neq g$.



(a) X' reduces group b's error



(b) X' does not reduce group b's error

takeaways

there is an active debate regarding whether to ban test scores in admissions decisions

since test scores are likely to be decision-relevant for both groups, our result suggests that:

- so long as G is permissible, then excluding test scores makes all designers worse off
- if G is not a permitted input (as is the case in California), then it can be strictly optimal to ban X' (as in previous example)

nuances/qualifications

- our results depend critically on our assumption that the designer has access to a **fully flexible** garbling of the inputs X
- do not imply a ranking between sending X' (un-garbled) versus excluding it