

Paths to AI Accountability

Making AI Fit for Humans Through Design, Measurement, and Regulation

Sarah H. Cen | MIT EECS | August 2, 2024

Thesis committee: Aleksander Mądry, Devavrat Shah, Manish Raghavan

AI algorithms are everywhere
(they affect all of us, whether we know it or not)

Healthcare

Hiring

Admissions

AI algorithms are everywhere

(they affect all of us, whether we know it or not)

Lending

Operations

How we interact
(social media)

What we buy
(targeted advertising)

Healthcare

Hiring

Admissions

AI algorithms are everywhere

(they affect all of us, whether we know it or not)

Lending

Operations

What we pay
(pricing algorithms)

How we vote
(news feeds)

How we receive info
(LLMs)

Reactions to AI

“AI should not intervene on consequential decisions”

Reasoning: AI will never fully understand the human experience

Reasoning: Codification of ethics and values is dangerous

Counterpoint: Even “minor” decisions are harmful in accumulation

So, should we require that AI not intervene on any decisions?

For better or worse, AI is here to stay

Reactions to AI

“AI is objective”

Reasoning: AI learns from data. Data is a true representation of reality.

Counterpoint: Data is inherently tied to the past

AI can propagate past human biases and idiosyncrasies

Counterpoint: Datasets are limited by what we chose to measure

For example, different performance metrics lead to different hiring recs

Regardless of whether AI is “objective,” it is inherently imperfect

Reactions to AI

there are some arguments
that AI is not a tool, but
we'll skip over it for this talk

"AI is a tool"

Reasoning: It carries out tasks on behalf of those who wield it: humans

If AI is a tool, then we can return to familiar concepts: **responsibility & liability**

Underlies area of **AI Accountability:**

1. **What** are we holding AI developers and deployers responsible for?
2. **How** do we hold them responsible (ensure that they uphold obligations)?

This thesis adds to body of work on AI accountability

AI Ethics

Moral and normative questions about AI's role in society

Machine Learning

Developing algorithms that align with society and the law

Paths to AI Accountability

Law & Policy

When and how the law can (and cannot) assist

Stats, Info & Games

Using stats, information, and game theory to analyze & improve AI systems

Thesis: 3 approaches to AI accountability

Recall: AI Accountability is holding AI developers & deployers responsible for their obligations to others

- I. Design:** Creating AI to be "responsible" from the ground up
- II. Measurement:** Determining how AI systems behave in practice
- III. Regulation:** Designing policies & laws to ensure responsibility

Design

User Strategization and Trustworthy Algorithms

Sarah H. Cen, Andrew Ilyas, and Aleksander Mądry
MIT
{shcen, ailyas, madry}@mit.edu

Abstract

Many human-facing algorithms—including those that power recommender systems or hiring decision tools—are trained on data provided by their users. The developers of these algorithms commonly adopt the assumption that the data generating process is *exogenous*: that is, how a user reacts to a given prompt (e.g., a recommendation or hiring suggestion) depends on the prompt and *not* on the algorithm that generated it. For example, the assumption that a person's behavior follows a ground-truth distribution is an exogeneity assumption. In prac-

TL;DR

Motivation

There is consistent calls for “**trustworthy**” algorithms

But we don’t know what “trustworthiness” means in practice!

At the same time, system designers aren’t incentivized to develop “trustworthy” algorithms if it interferes with performance

It appears in Biden’s executive order, the EU AI Act, and more

Our contribution

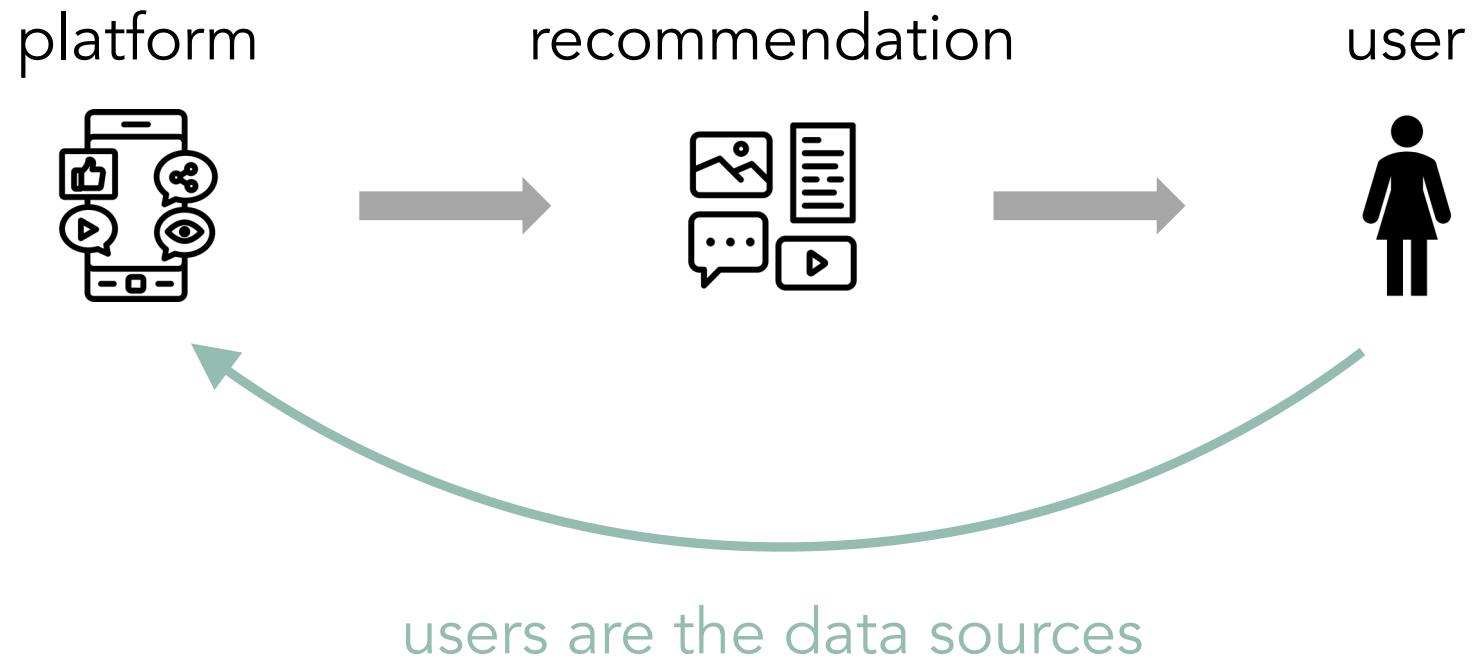
In this work, we

1. Provide a way to **formalize “trustworthiness”** (in a specific context)
2. Show that, when interacting with humans, **trustworthiness helps platforms!**

**Data-driven algorithms are
built on, well, data.**

Where does the data come from?

In many settings, the data comes from humans



In many settings, the data comes from humans



To make this work, typically assume that user behavior is **exogenous**

(i.e., if a *different* platform issues the same recommendation, the user would respond in the same way)

In many settings, the data comes from humans



In practice, users can **learn, adapt, and strategize**.

(i.e., they can respond to the same recommendation differently based on the algorithm that generated it!)

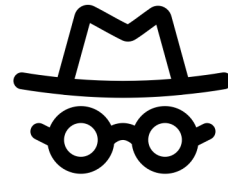
Strategization is common

Example 1: Social media users

User believes platform pays too much attention to their clicks.



Avoid clicking



Search links in private mode

“Sometimes I may like a song but not thumbs-up the song because I don't want my feed filled with similar artists/videos”

[Cen, Ilyas, Allen, Li & Madry, '23]

Strategization is common

Example 1: Social media users

User believes platform pays too much attention to their clicks.



Avoid clicking



Search links in private mode

"I avoid reading certain news stories on Google news because I know I will be bombarded with similar articles. Instead I switch to an untracked browser to read the story."

[Cen, Ilyas, Allen, Li & Madry, '23]

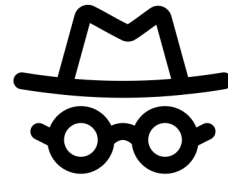
Strategization is common

Example 1: Social media users

User believes platform pays too much attention to their clicks.



Avoid clicking



Search links in private mode

“I have many YouTube accounts so my algorithm does not pick up a YouTube link a friend sends me to watch”

[Cen, Ilyas, Allen, Li & Madry, '23]

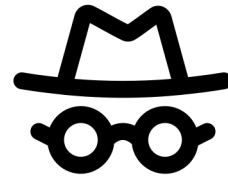
Strategization is common

Example 1: Social media users

User believes platform pays too much attention to their clicks.



Avoid clicking



Search links in private mode

Example 2: Uber drivers

Driver learns that Uber represents their preferences as unimodal.

Uber

← for longer rides



for shorter rides →

lyft

Strategization is common

Example 1: Social media users

User believes platform pays too much attention to their clicks.

**What are the implications
of user strategization?**

Example
Driver lea

Uber

← for longer rides

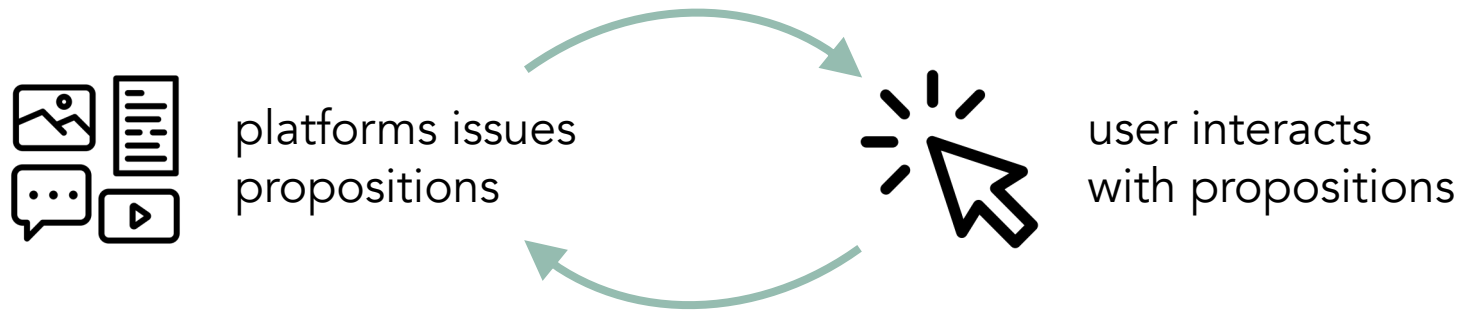


for shorter rides →

lyft

Contributions

Model: Repeated, two-player game



- Present a model that captures user strategization
- Find that strategization can help platform in short-term
- Find that strategization hurts platform by providing misleading data
- Show because humans are involved, **trustworthy design helps platforms**

Related work

Mechanism design & strategic behavior. [Myerson 1989; Nisan & Ronen 1999; Borgers & Kraemer 2015, ...]

Repeated, alternating games. [Roth et al. 2010; Fudenberg & Tirole 2005; Tuyls et al. 2018, ...]

Games & auctions with non-myopic users. [Amin et al. 2013; Liu et al. 2018; Abernethy et al. 2019; Haghtalab et al. 2022; Collina et al. 2024, ...]

Strategic classification. [Bruckner et al. 2012; Hardt et al. 2015; Levanon & Rosenfeld 2022; Zrnic et al 2021, ...]

Model

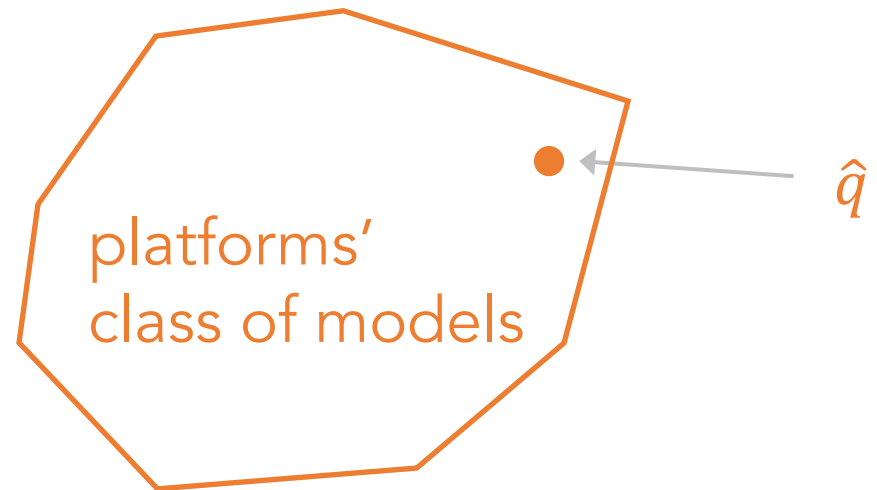
At each time step $t = 1, 2, \dots$

Platform generates propositions Z_t (recommendations, ride requests, diagnoses)

User responds with behavior $B_t \sim q(\cdot | Z_t)$ (engagement, admittance)

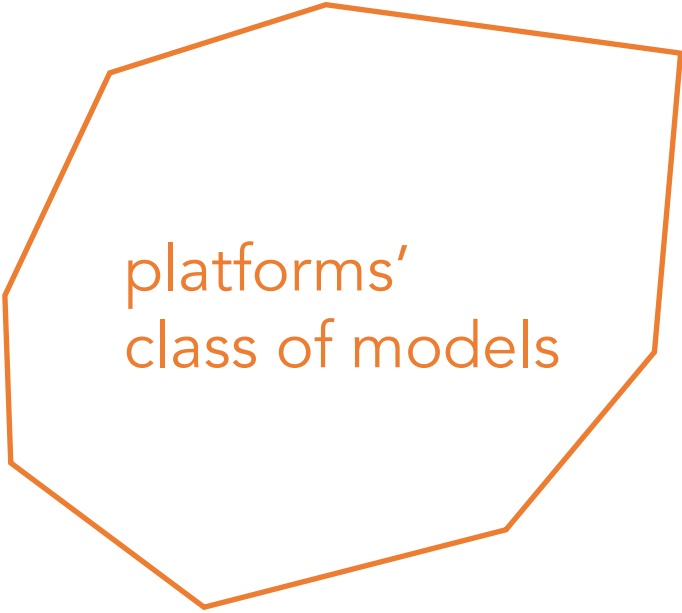
Platform and user collect payoffs $V(Z_t, B_t)$ and $U(Z_t, B_t)$.

To generate props, platform tries to learn model of user behavior q



Model

User behavior q ●



$$\hat{Q} = \{\hat{q}_i : i \in \Omega\}$$

Platform maintains belief about the user's behavior policy q

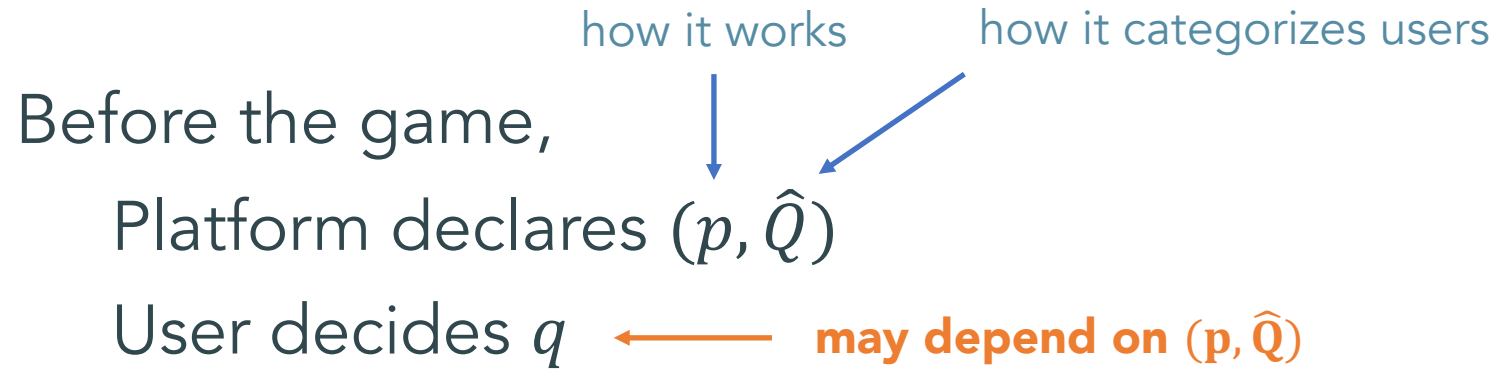
Platform (Bayesian) updates their belief at every t , using Z_t and B_t

Platform generates propositions Z_t using an algorithm p ←

e.g., if it believes you like cat videos, does it show you cat videos or animal videos

That is, $Z_t \sim p(\cdot ; \text{belief at } t)$

Model



At each time step $t = 1, 2, \dots$

Platform generates propositions $Z_t \sim p(\cdot; \mu_t)$

User responds with behavior $B_t \sim q(\cdot | Z_t)$

Platform and user collect payoffs $V(Z_t, B_t)$ and $U(Z_t, B_t)$.

Naive user vs. Strategic user

Naive user: Behaves as if there is no personalization algorithm (does not think ahead)

$$q^{BR}(B|Z) \propto \mathbf{1}\{B = \arg \max_{B'} U(B', Z)\}$$

user action



user payoff



Naive user vs. Strategic user

Naive user: Behaves as if there is no personalization algorithm (does not think ahead)

Strategic user: Chooses behavior that benefits them in the long term (maximizes equilibrium payoff)

$$q^*(p, \hat{Q}) = \arg \max_q \underbrace{\liminf_{t \rightarrow \infty} E_{Z \sim p(\cdot; \mu^t), B \sim q(\cdot|Z)} [U(B, Z)]}_{\text{worst-case, limiting expected payoff under } q \text{ (& game dynamics)}}$$

user behavior

Effects of strategization

Theorem (informal). When platform and user payoffs are sufficiently aligned but platform is mis-specified, then user strategization increases the platform's payoff.

Effects of strategization

Theorem (informal). When platform and user payoffs are sufficiently aligned but platform is mis-specified, then user strategization increases the platform's payoff.

Theorem (informal). If a platform collects data under one algorithm, its estimate of its payoff under a different algorithm can be arbitrarily bad for strategic users.

Effects of strategization

Theorem (informal). When platform and user payoffs are sufficiently aligned but platform is mis-specified, then user strategization increases the platform's payoff.

Theorem (informal). If a platform collects data under one algorithm, its estimate of its payoff under a different algorithm can be arbitrarily bad for strategic users.

Theorem (informal). A platform's payoff can decrease when it expands its model family when the user is strategic.

Trustworthy algorithms

Definition. Platform is **κ -trustworthy** if (i) the user is not incentivized to strategize and (ii) her payoff without strategization at least κ .

Two components:

1. **User is not incentivized to strategize**

User **trusts** that platform looks out for their platform-related interests so that user does not have to do so themselves [Hardin]

Trustworthy algorithms

Definition. Platform is κ -trustworthy if (i) the user is not incentivized to strategize and (ii) her payoff without strategization at least κ .

Two

1

Trustworthiness helps the platform by
(i) eliciting representative data and
(ii) preserving participation

2

Takeaways

We can use game theory to formalize trustworthiness!

Concurrently, Roth et al. have also used best-response behavior to formalize “trust” in AI

Trustworthy design can be good for the user and platform!

Takeaways

We

Matrix Estimation for Individual Fairness

ness!

Tru

In recent years, fairness have ari fairness (IF), w are similar rece matrix estimatic paradigm for h values. In this w We show that p improve an algo formance. Spec ular ME metho

Regret, stability & fairness in matching markets with bandit learners

Sarah H. Cen
Massachusetts Institute of Technology

Devavrat Shah
Massachusetts Institute of Technology

Abstract

Making an informed decision—for example, when choosing a career or housing—requires knowledge about the available options. Such knowledge is generally acquired through costly trial and error, but this learn-

in the dark can only obtain knowledge through costly trial and error, which can be further frustrated by the presence of competition. For example, if members of a group are repeatedly denied opportunities in academia due to competition, they may never obtain the requisite knowledge to make an informed decision about pursuing a career in academia.

s to

h!

Takeaways

We

Matrix Estimation for Individual Fairness

ness!

Regret, stability & fairness in matching markets with bandit learners

What's next? So, we may have developed a way to characterize responsibility. But how do we enforce it?

formance. Spec
ular ME metho

Making an informed decision—for example, when choosing a career or housing—requires knowledge about the available options. Such knowledge is generally acquired through costly trial and error, but this learn-

trial and error, which can be further frustrated by the presence of competition. For example, if members of a group are repeatedly denied opportunities in academia due to competition, they may never obtain the requisite knowledge to make an informed decision about pursuing a career in academia.

Measurement

Regulating algorithmic filtering on social media

Sarah H. Cen

MIT EECS

shcen@mit.edu

Devavrat Shah

MIT EECS

devavrat@mit.edu

Abstract

By filtering the content that users see, social media platforms have the ability to influence users' perceptions and decisions, from their dining choices to their voting preferences. This influence has drawn scrutiny, with many calling for regulations on filtering algorithms, but designing and enforcing regulations remains challenging. In this work, we examine three questions. First, given a regulation, how would one design an audit to enforce it? Second, does the audit impose a performance

TL;DR

Motivation

Social media has demonstrated its incredible sociopolitical importance

How do we characterize and measure the effect of social media algorithms?

Our contribution

In this work, we

1. Provide a way to **audit social media algorithms**
2. Show that the audit **respects user privacy**, requires **minimal access** to the algorithm itself, and **does not impose performance cost**

Information is power

Social media platforms have a lot of it

Calls to Regulate

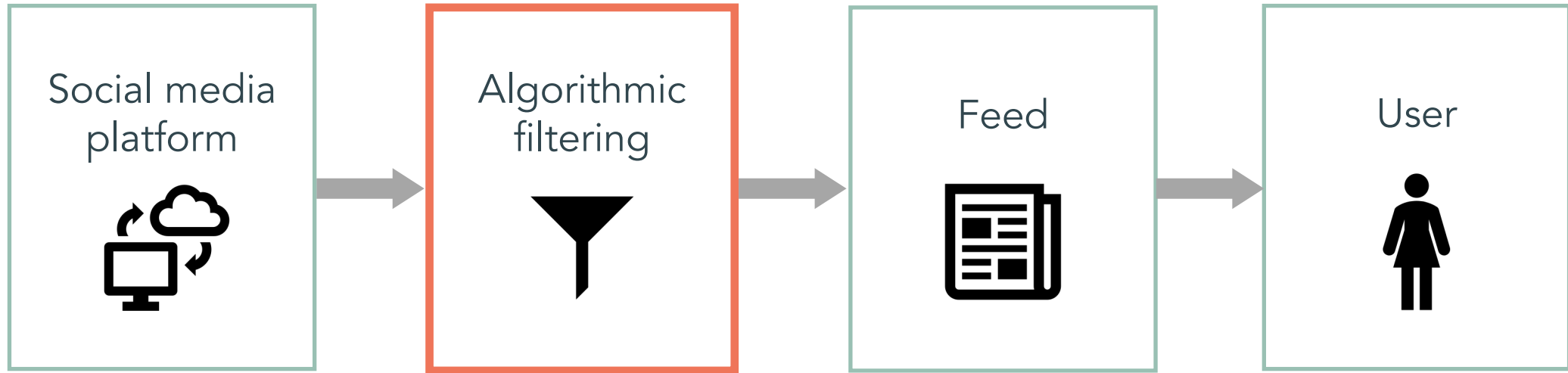
Ex 1: Ads not be based on user's sexual orientation

Ex 2: Info on public health (e.g., COVID-19) not reflect political affiliation

Ex 3: Not sway voting preferences beyond serving as a social network

Translating **desiderata** → **audit** is difficult

- Performance cost
- Removal of content (censoring)
- Privacy of users' personal data
- Access to algorithms is limited (e.g., due to trade secrets)



Main contribution: auditing procedure

Strong statistical guarantees

Not necessarily a performance-audit trade-off

Requires only black-box access

Does not remove content or require personal user data

Bonus: Incentivizes platform to inject content diversity

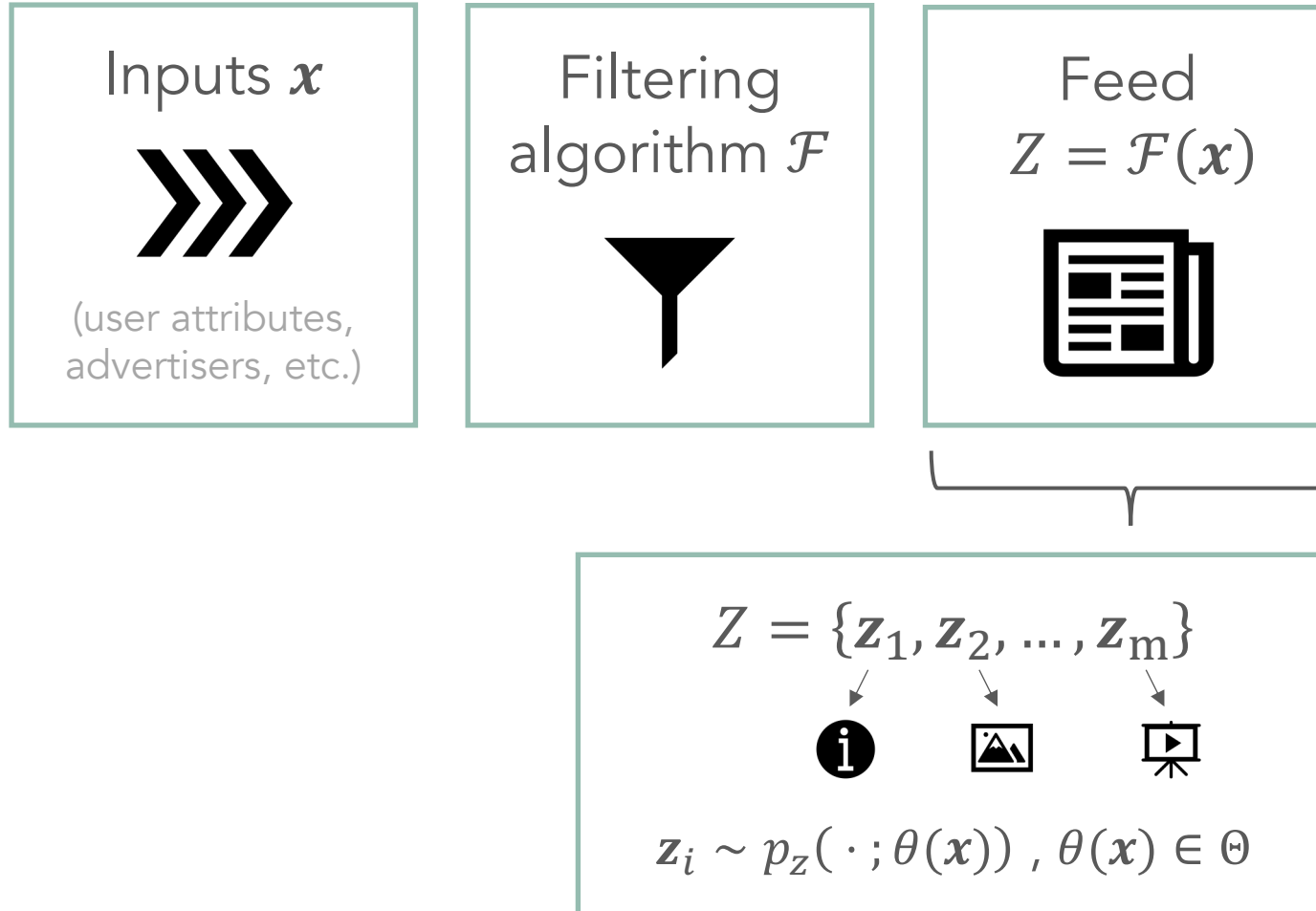
Related work

Social media trends & plights: [Mostagir & Siderius 2023] [Mostagir, Ozdaglar & Siderius 2022] [Haidt & Bail 2022] [Haidt & Twenge 2021] [Yang Mosleh Zaman Rand 2022] [Epstein Lin Pennycook Rand 2022] [Allcott & Gentzkow 2016] [Allcott Braghieri Eichmeyer Gentzkow 2019] [Milli Carroll Wang Pandey Zhao Dragan 2023] [González-Bailón & Lelkes 2023] [Kulshrestha Eslami Messias Zafar Ghosh Gummadi Karahalios 2019]

Social media regulation: [Keller 2021] [Laufer & Nissenbaum 2023] [Brannon & Holmes 2021] [Cobbe & Singh 2019] [Gillespie 2022] [Vese 2022] [Balkin 2021]

Auditing: [Metaxa Park Robertson Karahalios Wilson Hancock Sandvig 2021] [Robertson Lazer Wilson 2018] [Bartley Abeliuk Ferrara Lerman 2021] [Bandy & Diakopoulos 2020]

Problem setup



Why black box?

1. Minimal access
2. Algorithm agnostic
3. Prospective

Black-box access:
Run \mathcal{F} on $\{x_j\}$
and observe $\{Z_j\}$.

Auditor's task

Given a criterion & black-box access to \mathcal{F} ,
check if platform complies.



What do regulations have in common?

Insight: Most regulation don't ask for a **global definition** of a "good" outcome. They ask for **similarity** between outcomes.

What do regulations have in common?

Insight: Most regulation don't ask for a **global definition** of a "good" outcome. They ask for **similarity** between outcomes.

Example 1: Election interference

-  Global: Election-related content must come from trusted sources
-  Comparative: Let $\mathcal{F}(\mathbf{x}_{\setminus\{\text{untrusted election}\}})$ is the feed filtered if election-related content comes whitelisted sources. How different is the information in $\mathcal{F}(\mathbf{x})$ vs. $\mathcal{F}(\mathbf{x}_{\setminus\{\text{untrusted election}\}})$?

What do regulations have in common?

Insight: Most regulation don't ask for a **global definition** of a "good" outcome. They ask for **similarity** between outcomes.

Example 2: Discriminatory advertising

- ✗ Global: All users see same distribution of employments ads
- ✓ Comparative: If two users are identical except for race, their employment ads should be "similar," e.g., $\mathcal{F}(x, \text{White})$ and $\mathcal{F}(x, \text{Black})$ should contain similar employment ads.

What do regulations have in common?

Many regulations require “similar” behavior under two conditions

Recall: Inputs x include user attributes & content universe

Counterfactual criteria

“Algorithm \mathcal{F} must behave similarly under x and x' for all $(x, x') \in S$ ”

Counterfactual criteria

“Algorithm \mathcal{F} must behave similarly under x and x' for all $(x, x') \in S$ ”

In the election interference example,

x = some content universe & user

x' = same, except only trusted sources for election-related content

Counterfactual criteria

“Algorithm \mathcal{F} must behave similarly under \mathbf{x} and \mathbf{x}' for all $(\mathbf{x}, \mathbf{x}') \in S$ ”

In the discriminatory advertising example,

\mathbf{x} = some content universe & user

\mathbf{x}' = same, except user's race is changed

S are synthetic scenarios to test (no distributional assumptions!)

Counterfactual criteria

"Algorithm \mathcal{F} must behave similarly under x and x' for all $(x, x') \in S$ "

In

What is an appropriate notion of "similarity" ?

x' = same, except user's race is changed

S are synthetic scenarios to test (no distributional assumptions!)

Information is power

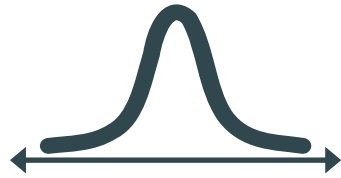
If we didn't think content affected votes and employment,
we wouldn't regulate social media to begin with

Our approach

Alice



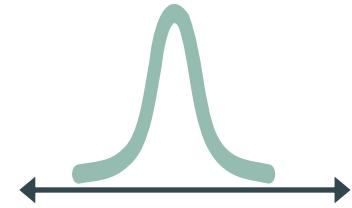
Initial belief $\hat{\theta}_0$



$Z = \mathcal{F}(x)$



New belief $\hat{\theta}_{\text{new}}$



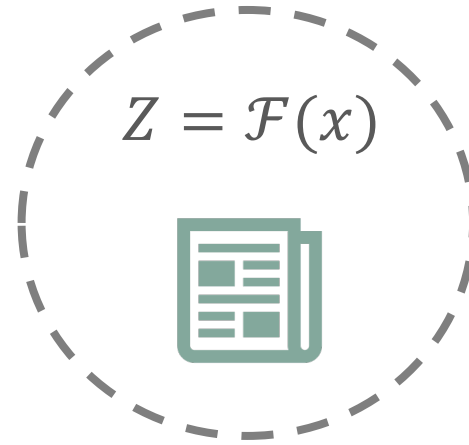
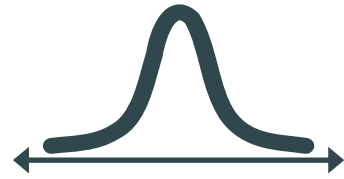
$$\hat{\theta}_{\text{new}} = \mathcal{L}^{\text{Alice}}(\hat{\theta}_0, Z)$$

Our approach

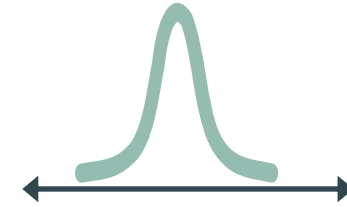
Alice



Initial belief $\hat{\theta}_0$



New belief $\hat{\theta}_{\text{new}}$



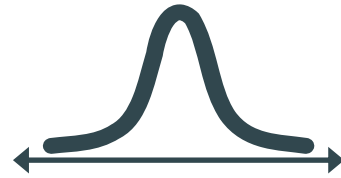
What would happen
if Alice was shown
 Z' instead?

Our approach

Alice



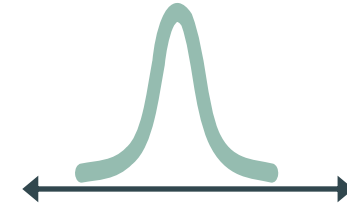
Initial belief $\hat{\theta}_0$



$Z = \mathcal{F}(x)$



New belief $\hat{\theta}_{\text{new}}$



What would happen
if Alice was shown
 Z' instead?

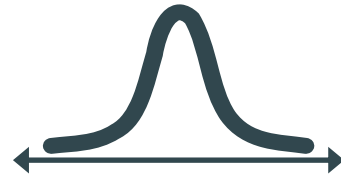
Election interference: If we swap Alice's personalized feed for a "trusted-sources-only" feed, would her beliefs change much?

Our approach

Alice



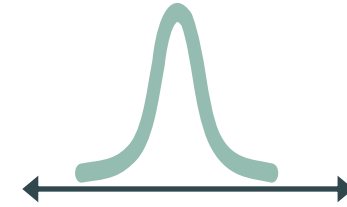
Initial belief $\hat{\theta}_0$



$$Z = \mathcal{F}(x)$$

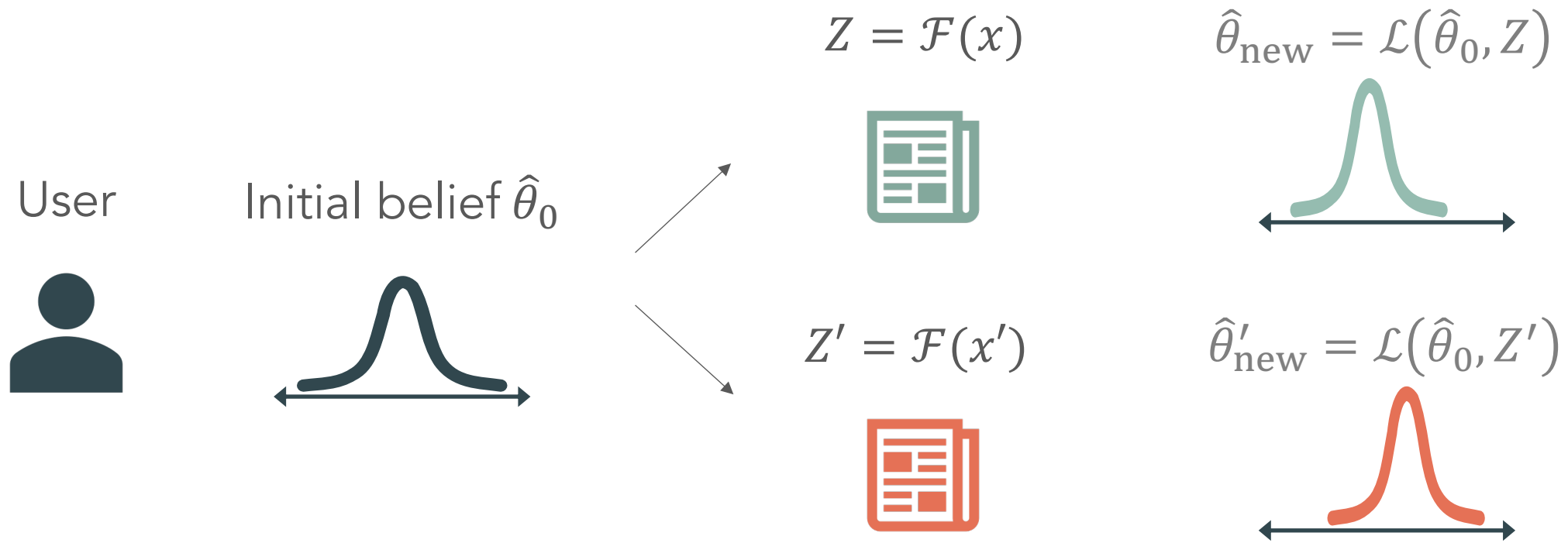


New belief $\hat{\theta}_{\text{new}}$



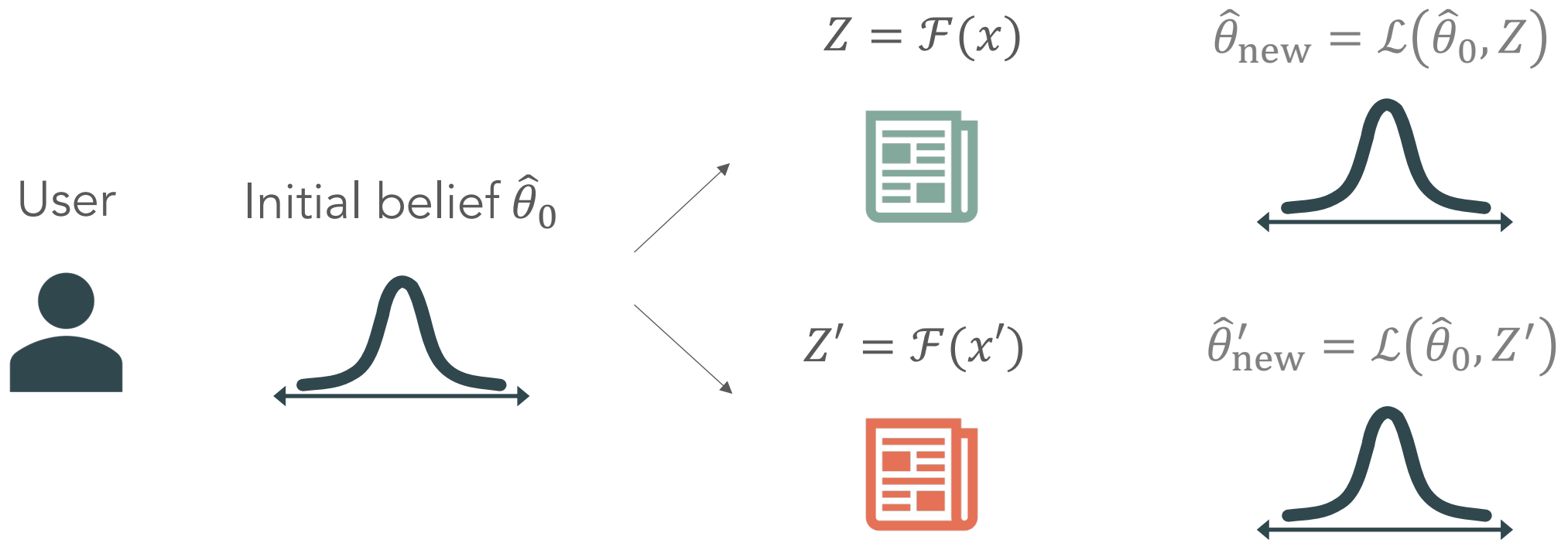
What would happen
if Alice was shown
 Z' instead?

Discriminatory advertising: If we swap Alice's ads for Alice's ads if her race was changed, would her beliefs about employment change much?



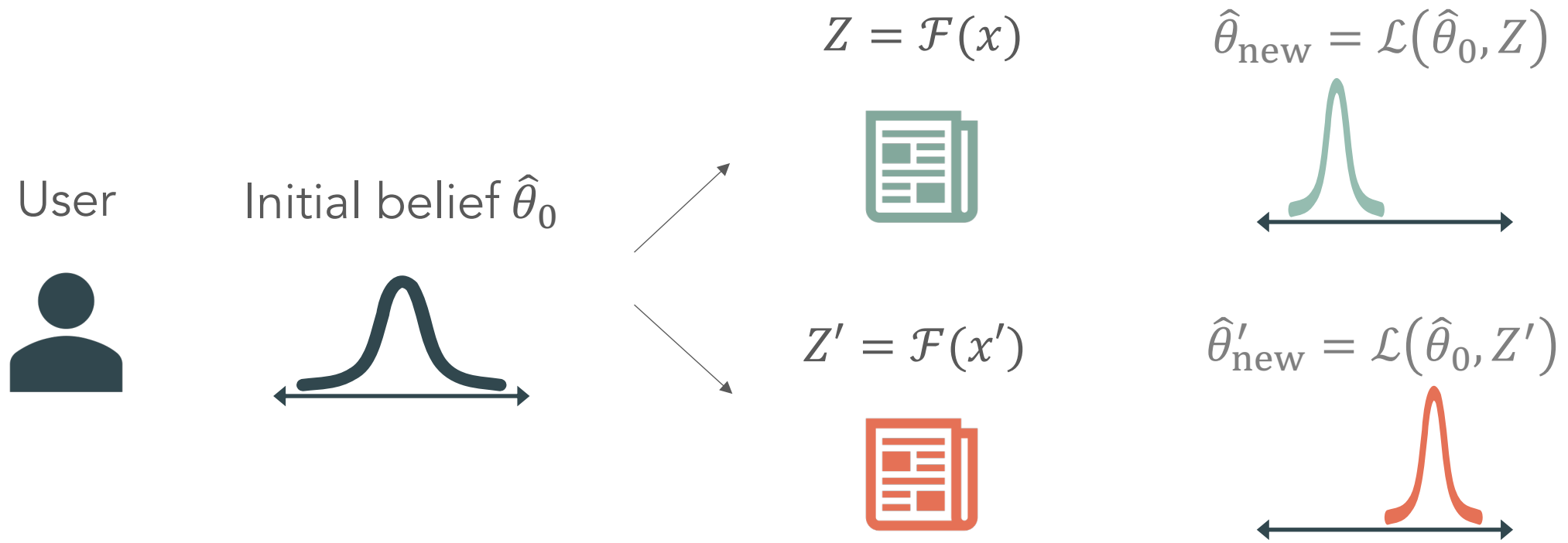
What is the effect of Z vs. Z' on user beliefs?

It depends on the user's learning behavior \mathcal{L}



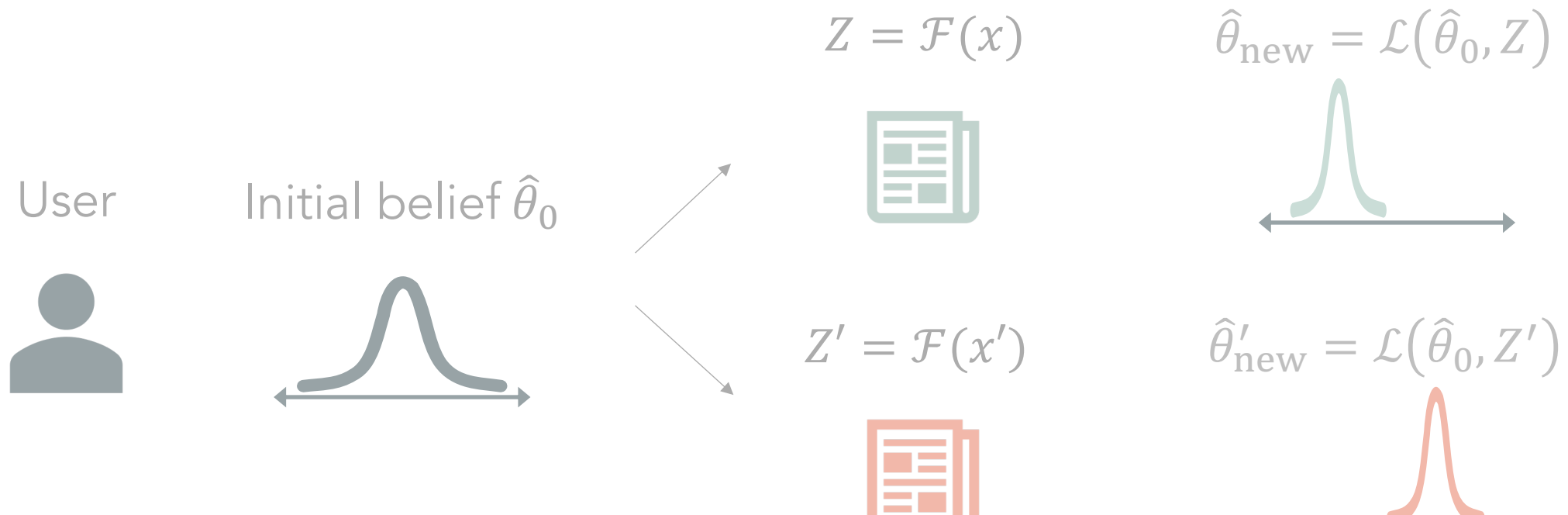
Suppose the user is stubborn \rightarrow nothing affects their belief

Then, $\hat{\theta}_{\text{new}}$ and $\hat{\theta}'_{\text{new}}$ are identical



Suppose the user is gullible \rightarrow they're very easily influenced

Then, $\hat{\theta}_{\text{new}}$ and $\hat{\theta}'_{\text{new}}$ are far apart



Every user learns differently. What should we do?

1. User studies \rightarrow good but costly
2. Pick representative user \rightarrow bad coverage

Every user learns differently. What should we do?

1. User studies \rightarrow good but costly
2. Pick representative user \rightarrow bad coverage

The auditor wants $\mathcal{L}(Z) \approx \mathcal{L}(Z')$ over all \mathcal{L}

So, let's upper bound the difference btw $\mathcal{L}(Z)$ and $\mathcal{L}(Z')$!

Every user learns differently. What should we do?

1. User studies \rightarrow good but costly
2. Pick representative user \rightarrow bad coverage

The auditor wants $\mathcal{L}(Z) \approx \mathcal{L}(Z')$ over all \mathcal{L}

So, let's upper bound the difference btw $\mathcal{L}(Z)$ and $\mathcal{L}(Z')$!

$$\max_{\mathcal{L}} d(\mathcal{L}(Z), \mathcal{L}(Z'))$$

Every user learns differently. What should we do?

1. User studies \rightarrow good but costly
2. Pick representative user \rightarrow bad coverage

The auditor wants $\mathcal{L}(Z) \approx \mathcal{L}(Z')$ over all \mathcal{L}

So, let's upper bound the difference btw $\mathcal{L}(Z)$ and $\mathcal{L}(Z')$!

$$\max_{\mathcal{L}} d(\mathcal{L}(Z), \mathcal{L}(Z')) < \delta \implies d(\mathcal{L}(Z), \mathcal{L}(Z')) < \delta \text{ for all } \mathcal{L}$$

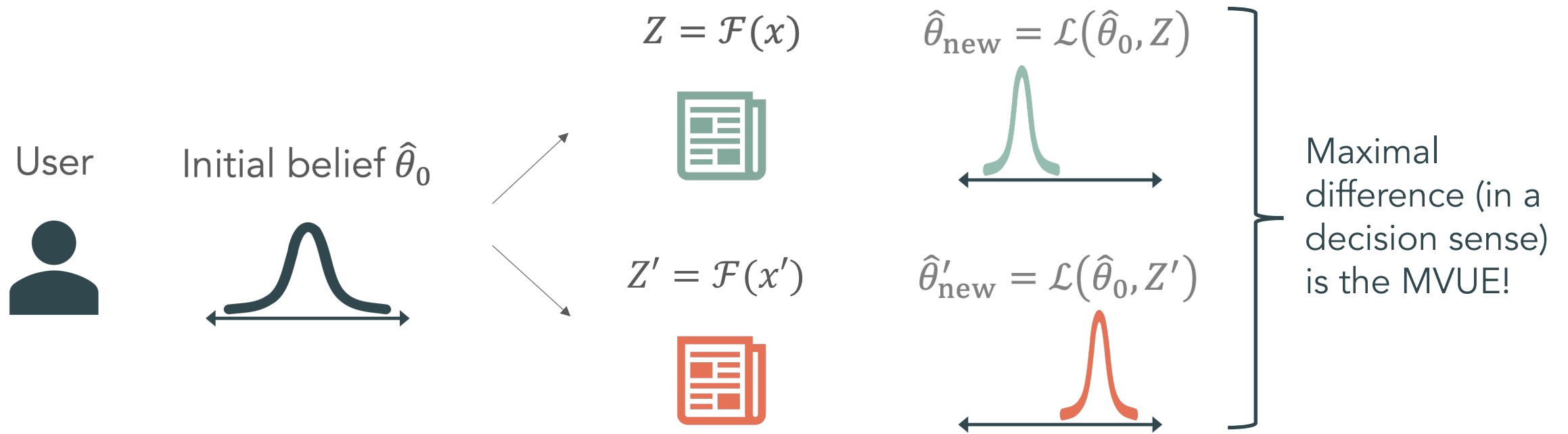
“Most gullible user”

Of all users, the learning behavior \mathcal{L} that results in the maximal difference between $\hat{\theta}_{\text{new}} = \mathcal{L}(Z)$ and $\hat{\theta}'_{\text{new}} = \mathcal{L}(Z')$ is the MVUE!

Proposition (informal)

- Suppose a user has a finite number of choices (e.g., candidates)
- Suppose the user has some default option (e.g., votes for Democrat)
- Then, the user who is **most likely to switch from the default** after seeing some feed Z is a user whose learning behavior \mathcal{L} is the **MVUE**

"Most gullible user"



Auditing procedure

“Algorithm \mathcal{F} must behave similarly under \mathbf{x} and \mathbf{x}' for all $(\mathbf{x}, \mathbf{x}') \in S$.”

Auditing procedure

“Algorithm \mathcal{F} must behave similarly under x and x' for all $(x, x') \in S$.”

Inputs:

\mathcal{F}

x

x'

Θ

ϵ

Auditing procedure

“Algorithm \mathcal{F} must behave similarly under \mathbf{x} and \mathbf{x}' for all $(\mathbf{x}, \mathbf{x}') \in S$.”

Inputs:

\mathcal{F}

\mathbf{x}

\mathbf{x}'

Θ

ϵ

1 $\tilde{\theta} \leftarrow \mathcal{L}^+(\mathcal{F}(\mathbf{x}));$

2 $\tilde{\theta}' \leftarrow \mathcal{L}^+(\mathcal{F}(\mathbf{x}'));$

Minimum-variance
unbiased estimator (MVUE)

Auditing procedure

"Algorithm \mathcal{F} must behave similarly under \mathbf{x} and \mathbf{x}' for all $(\mathbf{x}, \mathbf{x}') \in S$."

Inputs:

\mathcal{F}

\mathbf{x}

\mathbf{x}'

Θ

ϵ

- 1 $\tilde{\boldsymbol{\theta}} \leftarrow \mathcal{L}^+(\mathcal{F}(\mathbf{x}));$
 - 2 $\tilde{\boldsymbol{\theta}}' \leftarrow \mathcal{L}^+(\mathcal{F}(\mathbf{x}'));$
 - 3 **if** $(\tilde{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}}')^\top I(\tilde{\boldsymbol{\theta}})(\tilde{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}}') \geq \frac{2}{m} \chi_r^2(1 - \epsilon)$ **then**
 - 4 | Does not pass the test for $(\mathbf{x}, \mathbf{x}')$;
 - 5 **end**
 - 6 Passes the test for $(\mathbf{x}, \mathbf{x}')$;
- Minimum-variance unbiased estimator (MVUE)

Advantages

Intuitive tunable parameter. ϵ is false positive rate

Trade secret law. Only needs black-box access to \mathcal{F}

Modular. Can scale up for any (x, x') pairs

Privacy law. Do not need access to users or personal data

First amendment. Requires similarity btw outcomes \rightarrow unlike harmful content approach, there is no accidental stifling of speech

Section 230. Our approach audits \mathcal{F} , not the legality of content

Limitations

Choosing S . Auditor must choose what inputs to audit

Although out of scope of this work, we are exploring this direction

Asymptotic guarantee. We provide an asymptotic guarantee

Can likely improve the audit (and get finite-sample guarantees)

Type of criteria. Audit only works for counterfactual criteria

Audit is the UMP* (best hypothesis test)

The audit is the (approximate) UMPU as $m \rightarrow \infty$

That's what we wanted! It's the "best" possible audit

Theorem (informal). The audit is guaranteed to have a false positive rate (FPR) $\leq \epsilon$ as $m \rightarrow \infty$. Moreover, under regularity conditions, the audit is the UMP* test of all ϵ -significant tests.

Takeaway: Tune strictness using $\epsilon \in [0, 1]$ = allowable false positive rate

Is there a performance cost?

Not always.

- Audit does not require that Z and Z' are identical
- It requires that the information they convey is similar
- This can be achieved by adding content diversity!

Theorem (informal). Consider a finite feed. If performance is independent of elements in θ that can increase the Fisher information and the available content is holistic, then there is no regulation-performance trade-off.

Takeaways

We can audit algorithms using a counterfactual perspective

This helps when there is no obvious “reference” outcome

The audit requires only black-box access, does not violate user privacy, and does not always impose a cost on platforms

These considerations are important from viability in practice

Takeaway

We

The
use

Measuring Strategization in Recommendation: Users

Ad

Departme

Departme

Departme

Most mo

observin

a piece o

that gen

Network Synthetic Interventions: A Causal Framework for Panel Data Under Network Interference

Anish Agarwal

Department of Industrial Engineering and Operations Research, Columbia University, New York City, New York 10027,
aa5194@columbia.edu

Sarah H. Cen

Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge,
Massachusetts 02139, shcen@mit.edu

Devavrat Shah

Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge,
Massachusetts 02139, devavrat@mit.edu

Christina Lee Yu

Operations Research and Information Engineering, Cornell University, Ithaca, New York 14853, cleeyu@cornell.edu

We propose a generalization of the synthetic controls and synthetic interventions methodology to incorporate network interference. We consider the estimation of unit-specific potential outcomes from panel data in the presence of spillover across units and unobserved confounding. Key to our approach is a novel latent factor model that takes into account network interference and generalizes the factor models typically used

ive

3

orms

Takeaways

Network Synthetic Interventions:

We can
The
De
ve

Measuring Strategization in Recommendation: Users Adapt Their Behavior to Shape Future Content

Sarah H. Cen

Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, shcen@mit.edu

What's next? We have seen ways of formalizing and measuring responsibility. But what responsibilities exist?

Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, madry@mit.edu

Most modern recommendation algorithms are data-driven: they generate personalized recommendations by observing users' past behaviors. A common assumption in recommendation is that how a user interacts with a piece of content (e.g., whether they choose to "like" it) is a reflection of the content, but *not* of the algorithm that generated it. Although this assumption is convenient, it fails to capture user strategization: that users

Regulation

Winter 2023 ▾

Published on Feb 27, 2023

DOI 10.21428/2c646de5.a15f7255

SHOW DETAILS ↕

The Right to Be an Exception to a Data-Driven Rule

CITE [#]

SOCIAL ↩

DOWNLOAD ↓

CONTENTS ☰

Data-driven tools are increasingly used to make consequential decisions. In recent years, they have begun to advise employers on which job applicants to interview, judges on which defendants to grant bail, lenders on which homeowners to give loans, and more. In such settings...

by *Sarah H. Cen and Manish Raghavan*



last released
1 year ago

ABSTRACT

Data-driven tools are increasingly used to make consequential decisions. In recent years, they have begun to advise employers on which job applicants to interview, judges on which defendants to grant bail, lenders on which homeowners to give loans, and more. In such settings, different data-driven

TL;DR

Motivation

We often treat AI as an “oracle”

But even if AI is 99.99% accurate, there are always **exceptions**

Because AI is highly complex, current legal mechanisms of relief for people who are the “exceptions”

Our contribution

We propose that individuals have the “**right to be an exception**”

We detail this right (which emphasizes the importance of **uncertainty**)

We make sense of our world through **rules**.

But, to every rule, there are **exceptions**.

**What happens to individuals on
which the rule fails?**

Sentencing decisions

Mandatory minimum sentences (1970s)

Standardized set of rules

Intended to improve fairness, predictability, & consistency

Lockett v. Ohio (1978)

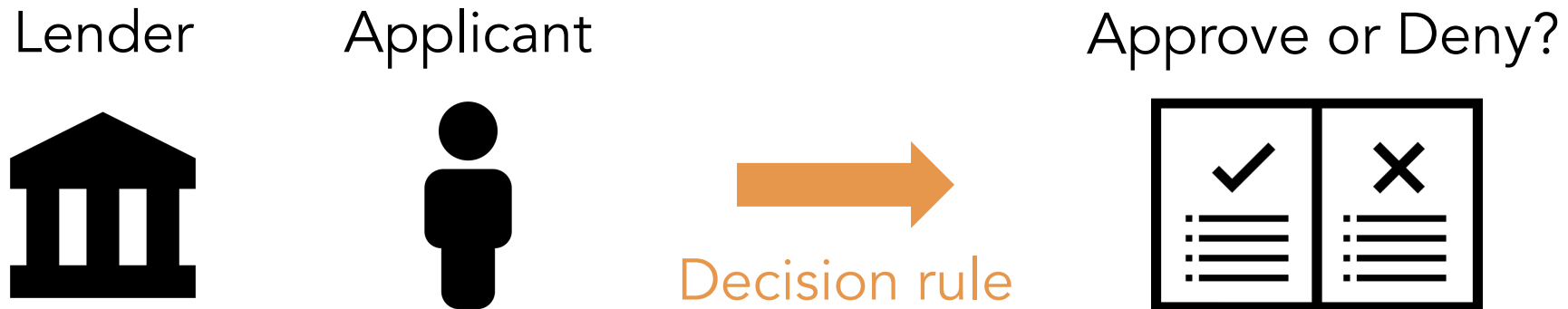
No mandatory minimum sentences for capital cases

Requires consideration of a case's particular circumstances

Due to "seriousness and irrevocability of the death penalty"

Data-driven exceptions

Data-driven rule: decision rule behind data-driven decision aid.



An applicant may be approved under some rules but not others.

Exceptions are natural.

Data-driven exceptions matter because:

1. ML \leftrightarrow statistical **averages**
2. ML can be applied **rapidly** and **repeatedly**
3. Data-driven rules are **non-intuitive**

Example: Exceptions in healthcare



Common cold



Fatal disease

Treated as average
of statistically similar
individuals?

vs.

Rule out exceptional
(high-risk) cases?

Individual rights

Rights in the age of AI

- Right to be forgotten (EU, 2014)
- Right to reasonable inferences (Wachter, 2019)
- Right to rectification (GDPR, 2016)
- Right to access (GDPR, 2016)
- ...

Goal: redistribute power back to decision subjects.

Right to be an exception to a data-driven rule

*When the risk of harm is high, a data-driven decision-maker must adopt the presumption that the subject **may be an exception** to the data-driven rule.*

*They must inflict harm only if they have applied the appropriate **care and diligence in ruling out the possibility** that the decision-subject is a data-driven exception.*

Moving away from averages

Are there protections for individuals who fall through the cracks?
Surprisingly few.

Most still rely on average-based notions.

Ex: Some believe improving accuracy justifies a method.

But accuracy is an average-based notion!

Loomis v. Wisconsin (2017)

Loomis v. Wisconsin

From the ruling:

1. Although algorithm is secret, no relevant information is hidden from Loomis because he **knows inputs and outputs**.
2. Use of gender by algorithm was not discriminatory and **promoted accuracy to the benefit of defendants**.

Loomis: algorithm is secret → violates right to due process

If argue on basis of accuracy (*average* notion), an *individual* will always lose.

Need new language: **harm, individualization, uncertainty!**

Right to be an exception

Has three ingredients

1. Harm
2. Individualization
3. Uncertainty

Element #1: Harm

Measurement stick: **What level of care, skill & diligence required?**

Weighs right against other stakeholder interests.

Ex: Individualized sentencing vs. judicial economy.

How to measure harm?

“Significant effects” (Kaminski & Urban, 2021)

“High-risk inferences” (Wachter & Middelstadt, 2019)

“Risk methodology” (EU AI Act, 2021)

Element #2: Individualization

Individualization: tailoring a rule to specific circumstances.

Shifts from **aggregate to individual**.

An information concept → considering totality of circumstances.

Limitations to individualization in data-driven rules.

Even if a data-driven rule were **fully individualized** (incorporated all relevant features), would this be enough?

Element #3: Uncertainty (Part I)

Exceptions defy general rules.

So, is can we just improve individualization? **No.**

(This is where we differ from existing proposals.)

Why? Always sources of uncertainty.

Ex: Suppose individualized by incorporating more info.

The more tailored, the less data (i.e., **less evidence**).

Even if sufficient data, **unremovable sources of doubt**.

Element #3: Uncertainty (Part II)

Two types of uncertainty:

1. **Epistemic**: reducible uncertainty from lack of knowledge.
2. **Aleatoric**: irreducible uncertainty from “unknowability”
e.g., randomness or too many factors

Individualization **reduces epistemic, but not aleatoric, uncertainty.**

Ex: College student’s performance is not predetermined.

Computational irreducibility (Wolfram, 2002)

Takeaways

The law as it stands now cannot protect against all AI harms

Exceptions are natural in data-driven decision-making.

**We define a framework to protect those who are AI
"exceptions" which are (definitionally) difficult to detect**

Does not imply every individual is an exception. But when a decision inflicts harm, consider the possibility the subject may be an exception

Three components: harm, individualization, and uncertainty

Takeaways

The law as it stands now cannot protect against all AI harms

Except

Making Sense of AI Risk Scores

Chapt
Sideri



AI is in
human
Accou

We defi
"except

Does
inflicts

Three compo

Subcommittee on Human Rights and the Law
Hearing: "Artificial Intelligence and Human Rights"

Written Statement of Aleksander Mądry¹

June 13th, 2023

Chairman Ossoff, Ranking Member Blackburn and Members of the Committee, thank you for inviting me to testify. Much has already been said and written about how AI may transform society, both about the opportunities and risks—from AI's potential to enhance our productivity, creativity, and overall quality of life to its ability to perpetuate discrimination, drive economic inequality, and pose an existential risk.

I will not reprise those conversations here. Instead, I will focus my testimony on one issue that I find particularly salient, time-sensitive and extremely worrisome: *how AI could erode central tenets that enable our society to function, including our ability to carry out democratic decision-making.*

Specifically, I will discuss how AI is poised to fundamentally transform mechanisms for the

Conclusion

Thesis: 3 approaches to AI accountability

Recall: AI Accountability is holding AI developers & deployers responsible for their obligations to others

- I. Design:** Creating AI to be "responsible" from the ground up
- II. Measurement:** Determining how AI systems behave in practice
- III. Regulation:** Designing policies & laws to ensure responsibility

Many opportunities in AI accountability

National Strategies & Innovation

US AI Initiative Act (2021)
Japan's AI Strategy (2019)
South Korea's National AI Strategy (2019)
Australia's AI Action Plan (2021)

Data Protection & Privacy

EU GDPR (2016)
South Korea's Data 3 Act (2020)
California Consumer Privacy Act (2018)
Japan's APPI (2017)

Ethical Guidelines & Responsible AI

Biden's Executive Order (2023)
State-level regulation (discrimination)
EU's Guidelines for Trustworthy AI (2019)

Regulatory & Compliance Frameworks

EU AI Act (2023)
US Algorithmic Accountability Act (2023)
FTC & FDA rules

Many opportunities in AI accountability

The collage features several key elements:

- EU Artificial Intelligence Act:** A logo with a starburst pattern and the text "EU Artificial Intelligence Act".
- White House:** The text "THE WHITE HOUSE" in blue serif font, accompanied by a colorful circular logo of the White House.
- Handwritten Note:** A whiteboard with blue and red ink. It contains two entries: "118: lecturer name: Jane dept: CS" and "name: Jane dept: IT". There are also some numbers in brackets like [3, 3] and [4, 4].
- New York Times Article:** The masthead "The New York Times" and "Account" with a dropdown arrow. The main headline is ***A Hiring Law Blazes a Path for A.I. Regulation***. Below it, the sub-headline reads: "New York City's pioneering, focused approach sets rules on how companies use the technology in work force decisions."

Acknowledgements

Thank you to my amazing thesis committee!



Aleksander Mądry



Devavrat Shah



Manish Raghavan

Thank you to my past mentors!



Naomi Leonard



Paul Newman



Karl Tuyls



Yaron Rachlin



Vijay Kumar



Vaibhav Srivastava

Thank you to my collaborators!

Anish Agarwal



Cosimo Fabrizio



Aspen Hopkins



Andrew Ilyas



Hannah Li



Martha Minow



Asu Ozdaglar



Chara Podimata



Jennifer Allen



Rohan Alur



James Siderius



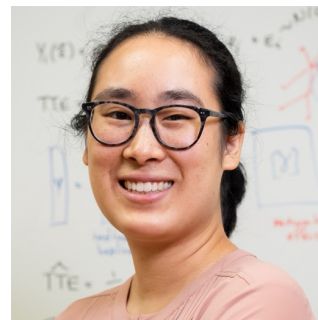
Isabella Struckman



Luis Videgaray



Christina Lee Yu



Cindy Zhang



Thank you to friends ❤️🧡💛



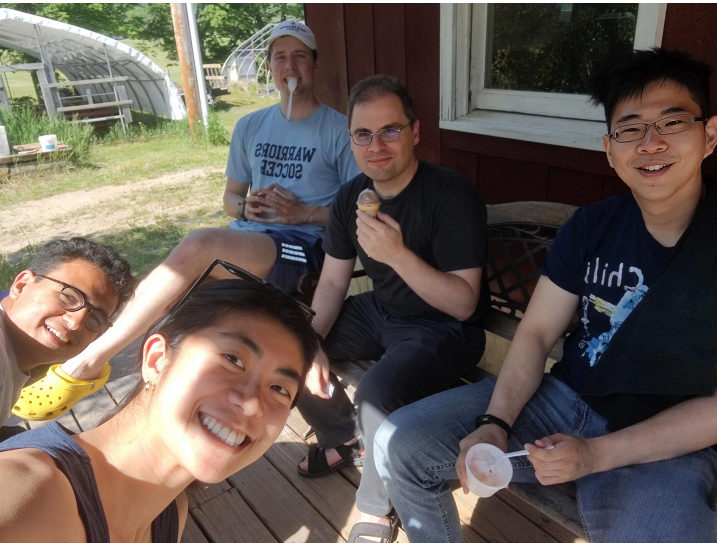
Thank you to friends ❤️🧡💛



Thank you to friends ❤️❤️❤️



Thank you to my labs ❤️❤️❤️



Thank you to my loved ones ❤️🧡💛



Thank you!

Questions?