Stanford HAI RegLab

RegLab

Adapting for AI

Sarah H. Cen

Stanford University | Dept. of Computer Science & Law School

(incoming Assistant Professor at CMU ECE & EPP)

AI has moved from the lab into our homes and institutions

Consequences of AI

AI intervenes on our **private sphere** (social media)

It has had profound **economic impact** (AI supply chains)

It affects our **rights & livelihoods** (employment algorithms)

It poses **societal risks** (privacy concerns)

It raises **existential questions** (AGI)

So, what?

Change is good! What's important is that we:

1. Adjust our understanding **of AI to its integration into society**
2. Adjust our understanding **of society as it integrates AI**

Today: We examine both perspectives

First, AI supply chains

Then, evidentiary burdens in legal cases against AI decisions

AI Supply Chains





Joint work at MIT with Aspen K. Hopkins, Andrew Ilyas, Isabella Struckman, Luis Videgaray, and Aleksander Madry

Ongoing work at Stanford with Jonathan Xue, Lindsey Gailmard, Daniel Ho, and Percy Liang

AI Supply Chains: The Complex Ecosystem of AI Actors

Aspen Hopkins* (EECS), Sarah H. Cen* (EECS), Andrew Ilyas (EECS),
Isabella Struckman (EECS), Luis Videgaray (Sloan), and Aleksander Mądry (EECS)

Massachusetts Institute of Technology

Abstract

The increasing complexity, accessibility, and outsourcing of AI systems has led to the emergence of AI supply chains: intricate networks of organizations contributing services, models, & datasets to AI development. This setting offers great potential yet is poorly understood. In this work, we model AI supply chains as directed graphs, providing two illustrative case studies on how AI supply chains can exacerbate issues of explainability and fairness. Specifically, we provide theoretical and empirical evidence showing that errors in local linear explanations increase with the width and depth of an AI supply chain, and that imposing conditions upstream (e.g., fairness) can propagate downstream in unexpected ways.

Outline

I. Introduction to AI supply chains
II. Case Study 1: Algorithmic fairness
    A. Theoretical result
    B. Experiments
III. Case Study 2: Explanations
    A. Theoretical result
    B. Experiments

What are AI supply chains?

AI supply chains are the complex network of AI products and services that integrate and produce AI

A canonical example is:

1. Org 1 produces a pre-trained base model M
2. Org 2 curates specialized data D
3. Org 3 fine-tunes M on specialized data D

A brief historical perspective

A tradition of "**outsourcing**" ML work developed over decades

- Began with data work (WordNet, ImageNet, Mechanical Turk, Scale)
- Extended to model training (transfer learning, AutoML)

For the most part, this was gradual until 2022

- With ChatGPT, there was an explosion of AI adoption
- This led to emergence of complex AI supply chains

Supply chains actually signal **healthy growth of AI industry**

- Improve efficiency, allow for specialization

Implications

AI supply chains have lots of implications!

Example: Copyright [Lee, Cooper & Grimmelman 2023]

How should *credit* be attributed?

How should *royalties* be distributed?

Who has *ownership* over an AI creation?

Implications

AI supply chains have lots of implications!

Example: Supply chain resilience

If an AI product or service goes down or suddenly decides to change how their product/service works, how does it affect others?

How should companies communicate to one another?

Implications

AI supply chains have lots of implications!

Example: Accountability [Widder & Nafus 2023]

- AI products & services combine in “non-modular” ways
- When so many entities contribute, how do we assign responsibility?

Implications

AI supply chains have lots of implications!

Example: Market concentration [**C**HISVM ’23 & ongoing work]

Where is there market concentration in the AI industry?

What are the implications of market concentration?

Previously, published short pieces at MIT.

At Stanford, we’re constructing the multilayered AI supply chain using public information (SEC filings, press releases, etc.)

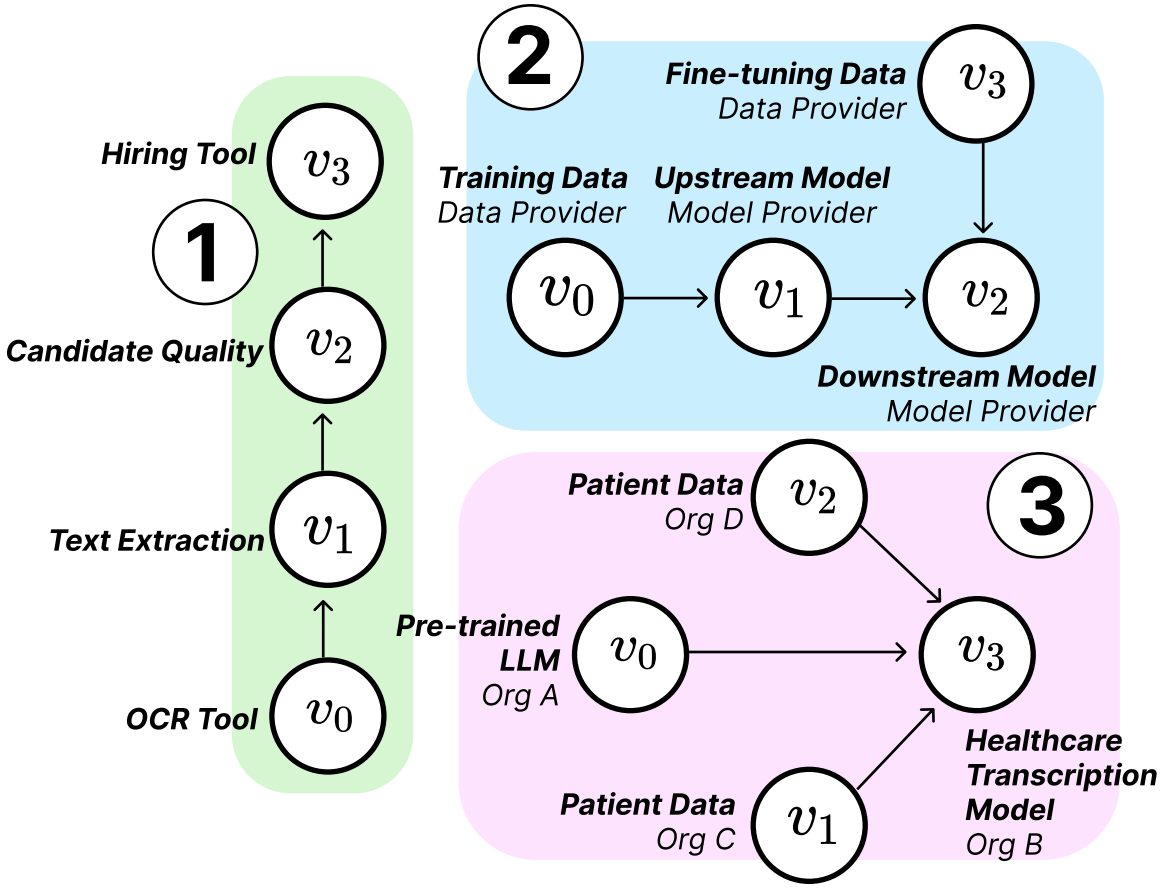
Implications

AI supply chains have lots of implications!

Example: Machine learning [**C**HISVM ’25] ← today!

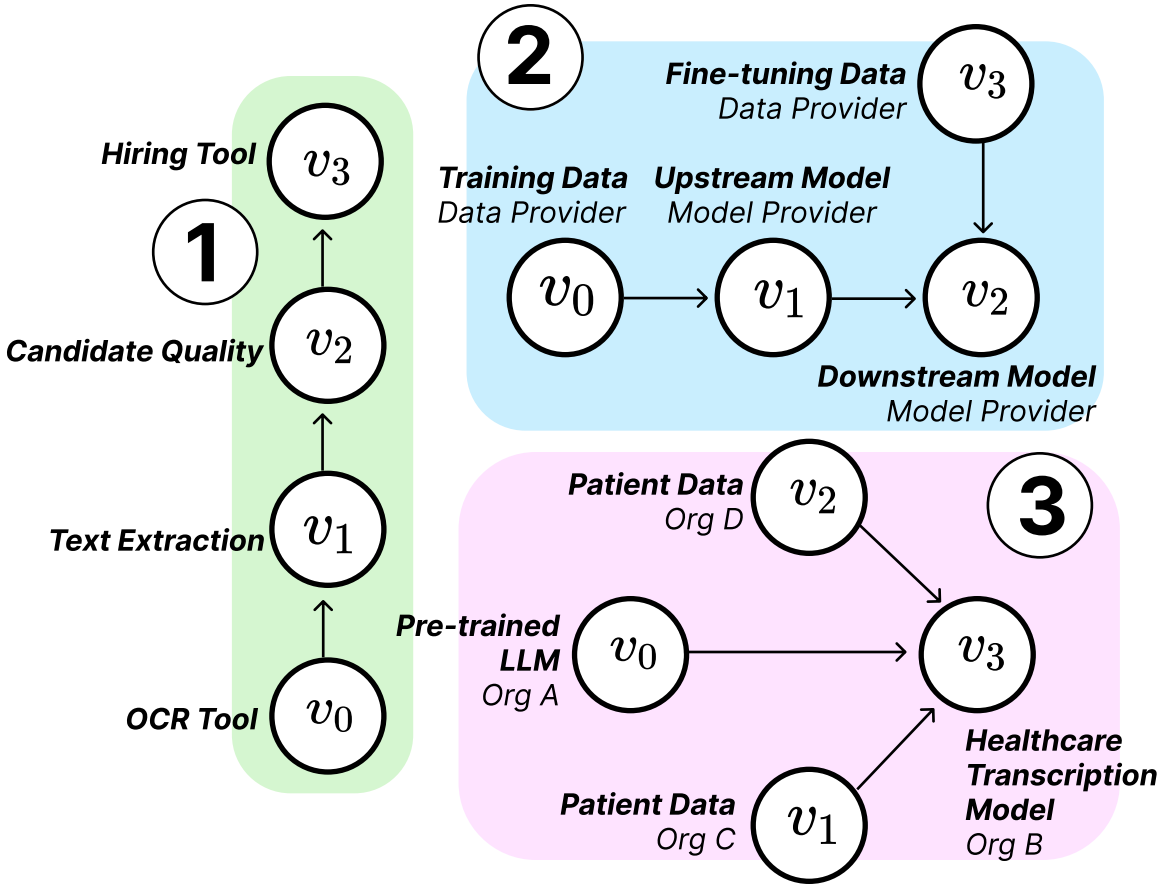
- How does the AI supply chain complicate ML development?
- How do ML decisions propagate through an AI supply chain?

AI supply chains are directed graphs



This follows tradition of supply chain literature long before AI

AI supply chains are directed graphs



Allows us to study events such as "dispersed control"

We formalize m-dispersed control as when changes to ancestors' operations $\{h_a\}$ within m hops of node v cannot be reversed by any changes made to h_v

Case Study 1: Upstream decisions have downstream consequences

Motivation

Upstream actors inevitably make design decisions

Downstream actors often have their own design criteria

- Downstream actors operate in more specialized industries

It is impossible for upstream actors to accommodate all possible downstream desiderata

How do upstream decisions affect downstream actors?

Related work

Unlearning can be reverse [Hu et al. 2024]

Upstream decisions can be “undone” [Salman et al. 2022]

No fair representation can guarantee fairness under any downstream data distribution [Lechner et al. 2021]

Fine-tuning can erase pre-training biases [Kirichenko et al. 2023, Qi et al 2023]

Setup

Input $\boldsymbol{x} \in \mathbb{R}^{\rho}$

Base model f_p

Fine-tuned model f_v

For simplicity, f_p and f_v take in $\boldsymbol{x}$'s and output scalars

e.g., both take in applicant information and output scores of some sort

Upstream constraint

Suppose f_p is trained w/ conditional independence (CI) constraint

$$f_p(X) \perp X_1 \mid Z$$

Conditional independence encompasses various types of structured learning (e.g., in causal inference)

It also includes notions of algorithmic fairness, like equalized odds ($Z = Y$) and demographic parity ($Z = \emptyset$) where X_1 is sensitive attribute

Suppose base model $f_p(\boldsymbol{x})$ is obtained by learning $\{\phi_i\}$ and $\{w_i\}$, where

$$f_p(x) = \sum_{i=1}^{N} \phi_i(x)^\top w_i$$

(Universal function approximator as $N \to \infty$ under sufficiently rich basis fns)

We model fine-tuning as learning linear model $\{v_i\}$ on embeddings

$$f_v(x) = \sum_{i=1}^{N} \phi_i(x)^\top v_i$$

Result

Theorem. Suppose $f_p(X) \perp X_1 \mid Z$ (w.r.t. some data distribution μ) and f_p is trained with L_1 sparsity regularizer on $\{\phi_i\}$. If basis functions are sufficiently rich and non-redundant*, then

$$\mathbb{E}_\mu[f_v(X)|X_1 = \beta, Z = \gamma] = \mathbb{E}_\mu[f_v(X)|Z = \gamma] \ , \qquad \forall \beta, \gamma$$

fine-tuned model

i.e., does not inherit CI, but 1st moment version (though equal sometimes)

*$g_i(\beta,\gamma) = \frac{\partial}{\partial \beta}\mathbb{E}[\phi_i(X)|\, X_1 = \beta, Z = \gamma]$ and assume g_i are linearly independent (as long as $g_i \neq 0$)

Experimental setup

Base model f_p (trained on data D_1)

Model f_v fine-tuned on f_p (with data D_2)

Model architecture: ResNet18

Dataset: Waterbirds

Common fairness dataset, background = sensitive attribute

Loss: $\mathcal{L}(f) = \mathrm{BCE}(f) + \alpha R_{\mathrm{fairness}}(f)$

Experimental setup

Loss: $\mathcal{L}(f) = \mathrm{BCE}(f) + \alpha R_{\mathrm{fairness}}(f)$

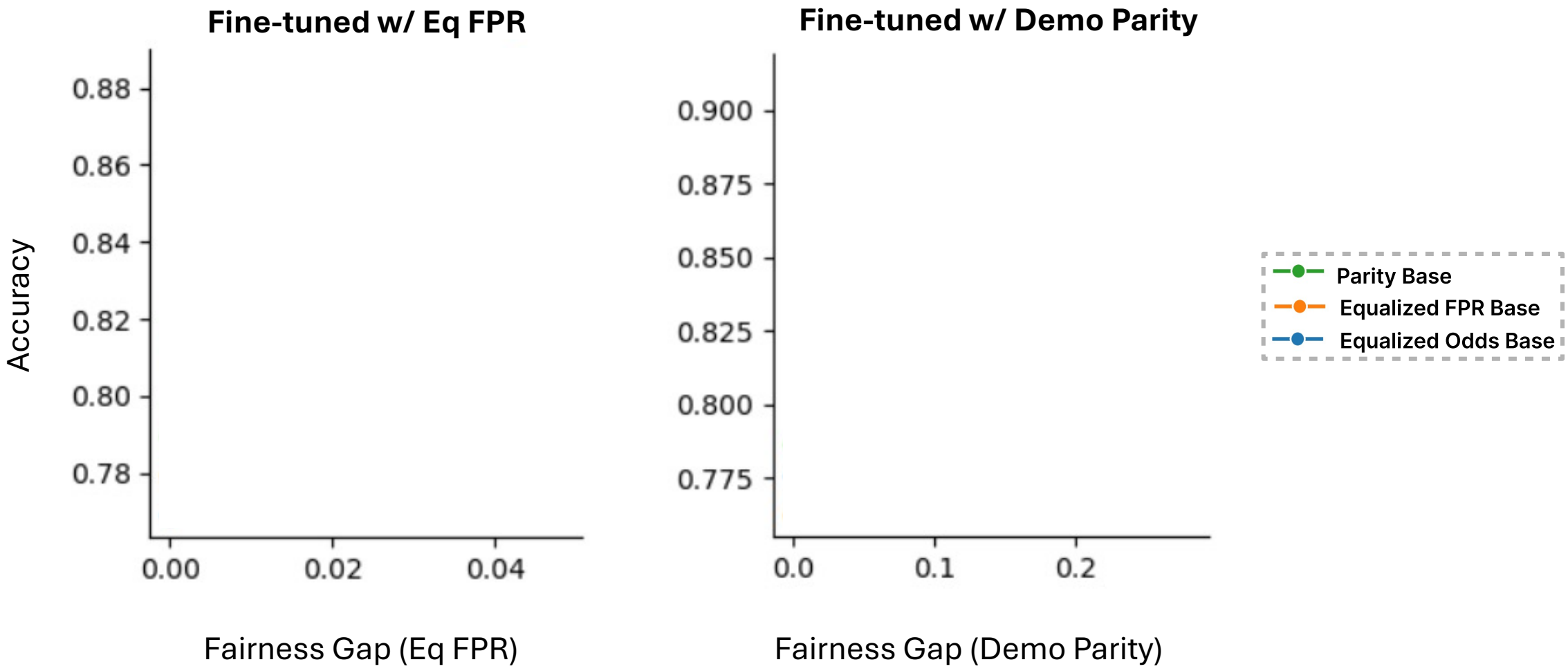
We use three types of fairness

- Demographic parity: selection rate for groups 1 and 2 is same
- Equalized FPR: FPR for groups 1 and 2 is same
- Equalized Odds: FPR and TPR for groups 1 and 2 is same

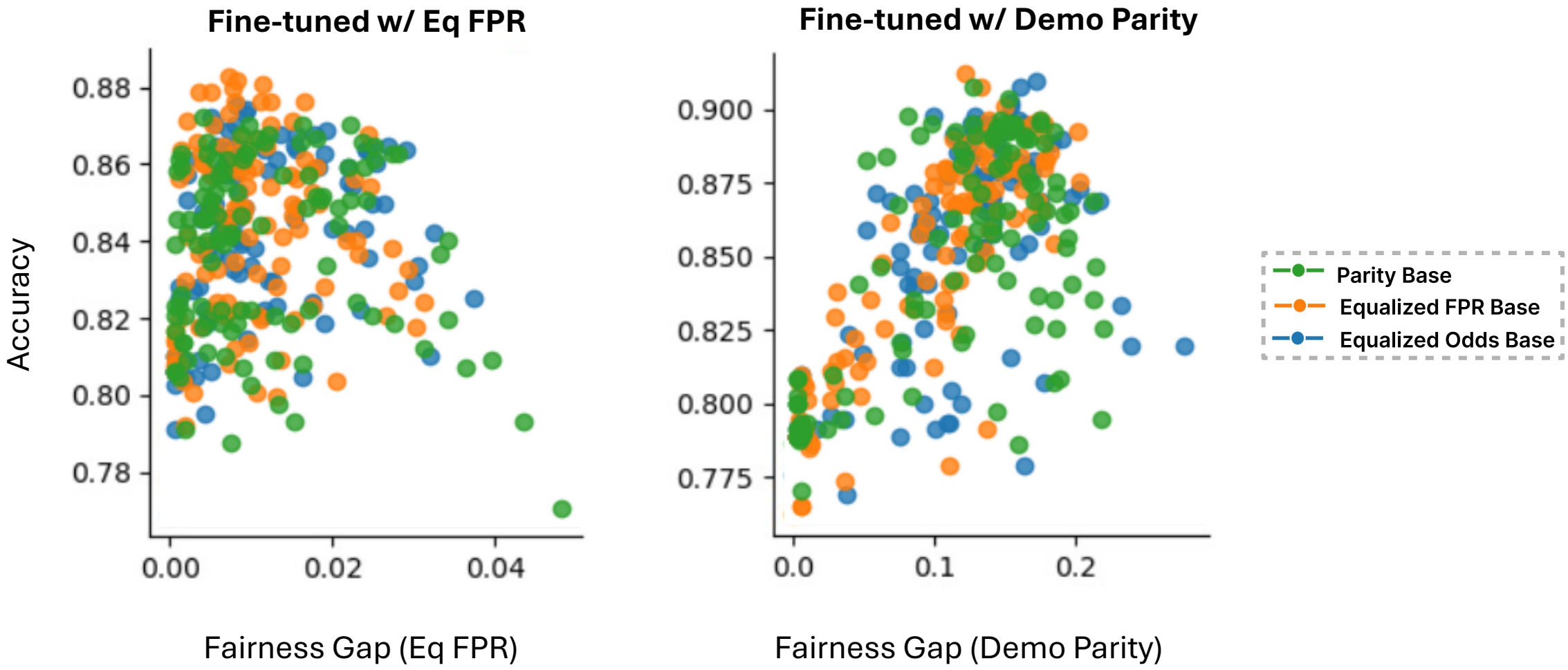
We vary the base α_{base} and fine-tuning α_{ft} regularization constant

Trained over 10,000 models

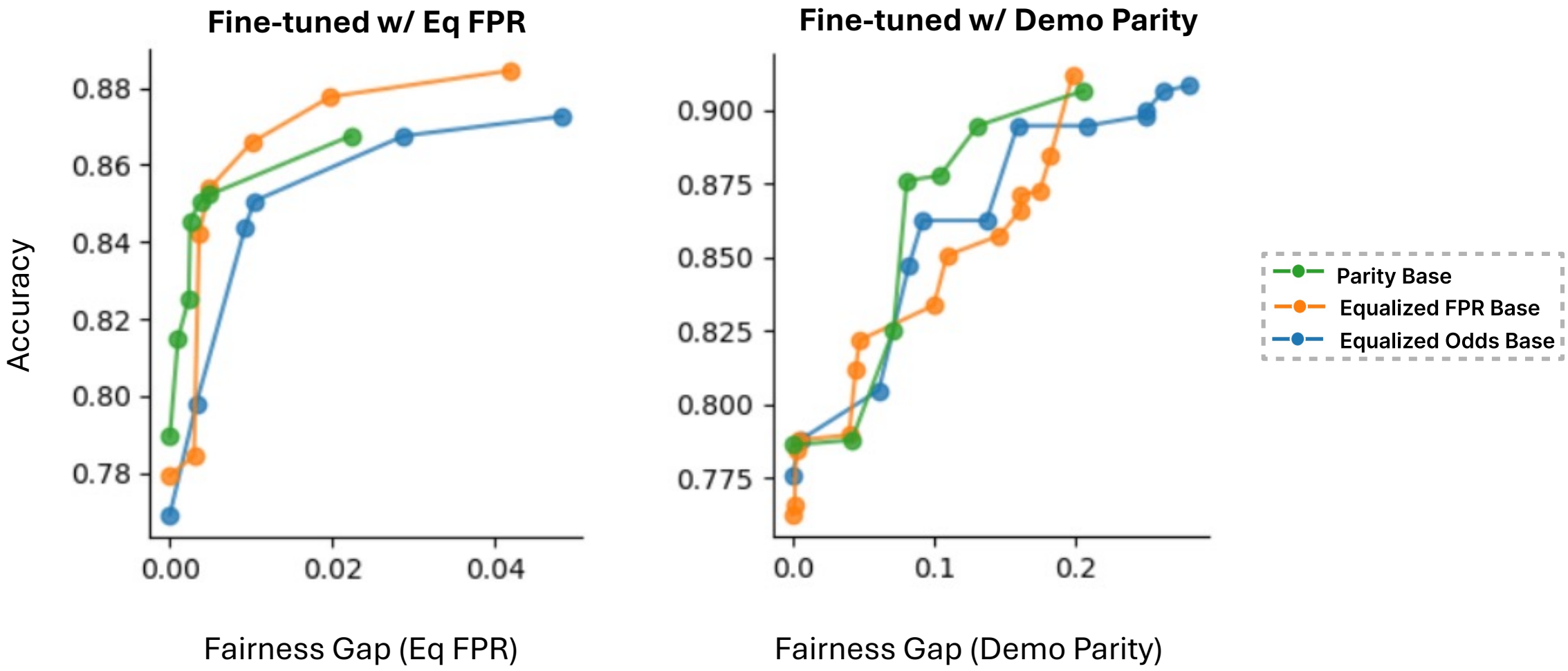
Results (fine-tuning on full network)



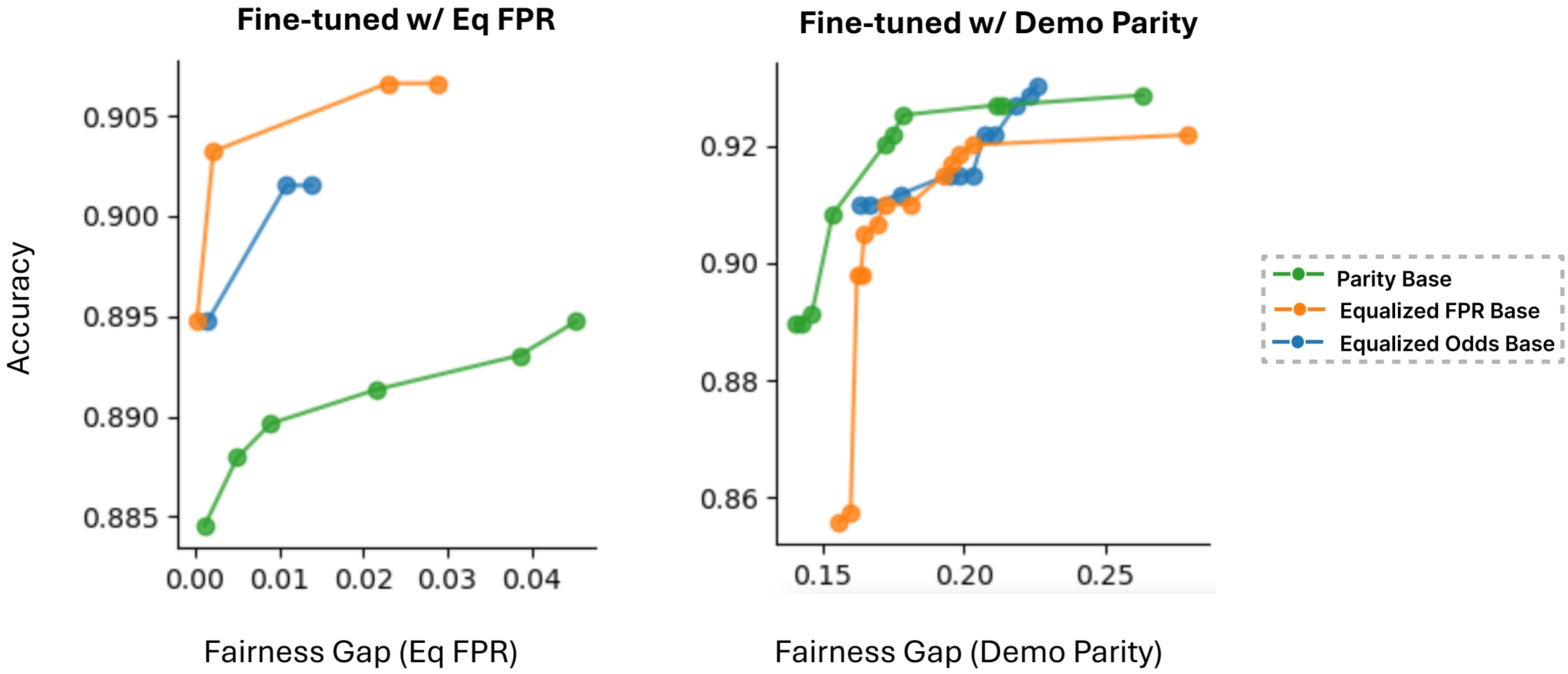
Results (fine-tuning on full network)



Results (fine-tuning on full network)



Results (fine-tuning on last layer)



Implications

Upstream constraints have downstream consequences!

It’s not straightforward: Downstream models don’t simply inherit upstream properties

In fact, you can “undo” or “remove” them …

… but they still leave a footprint

- In our case study, imposed a performance-fairness tradeoff

Case Study 2: Information Propagation in the AI supply chain

Motivation

Supply chains spread knowledge across multiple actors.

What are the implications of dispersed knowledge?

Case study: Explanations of AI decisions

- Suppose a company must provide explanations of its model's decisions
- The company's model is built on an AI supply chain
  - e.g., it uses the outputs of other models as inputs to its own model
- The company must generate end-to-end explanations
  - e.g., explain why an applicant was rejected (not just how it used other models)
- However, the company does <u>not</u> have access to upstream AI models

Setup

Applicant $\boldsymbol{x} \in \mathbb{R}^{\rho}$

Organization v

Decision $f_v(\boldsymbol{x})$

Setup

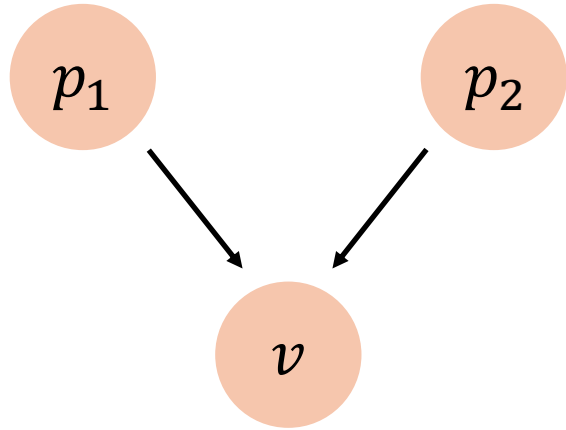
Organization v receives applicant $\boldsymbol{x}$

v sends $\boldsymbol{x}$ to p_1 and p_2

p_1 processes $\boldsymbol{x}$ → sends z_1 back

p_2 processes $\boldsymbol{x}$ → sends z_2 back

v uses z_1 and z_2 to produce output $f_v(\boldsymbol{x})$



Setup

Applicant $\boldsymbol{x} \in \mathbb{R}^{\rho}$

Organization v

Decision $f_v(\boldsymbol{x})$

Organization v uses third-party tools, i.e.,

$$f_v(\boldsymbol{x}) = h_v(\boldsymbol{x}, f_{p_1}(\boldsymbol{x}), \ldots, f_{p_n}(\boldsymbol{x}))$$

where $p_1, \ldots, p_n$ are v's parents.

Locally linear explanations

A δ-explanation at for model g at $\boldsymbol{z} \in \mathbb{R}^m$ is $E_\delta(g, \boldsymbol{z})$ where

$$E_\delta(g, \boldsymbol{z}) \in \underset{M}{\operatorname{argmin}}\, \mathbb{E}_{\boldsymbol{u}} \| g(\boldsymbol{z} + \delta \boldsymbol{u}) - g(\boldsymbol{z}) - M^\top \boldsymbol{u} \|_2^2$$

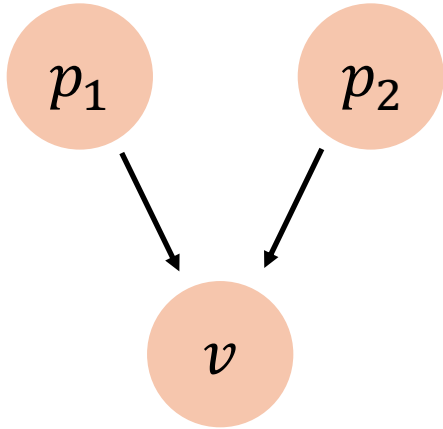
where $\boldsymbol{u}$ is drawn uniformly at random from unit ball in $\mathbb{R}^m$.

This type of explanation encompasses popular approaches, such as LIME [Ribeiro, Singh, Guestrin SIGKDD'16]

Passing explanations along the AISC

Organization v receives applicant $\boldsymbol{x}$

v sends $\boldsymbol{x}$ to p_1 and p_2



p_1 processes $\boldsymbol{x}$ → sends z_1 + explanation $E_\delta(f_{p_1}, \boldsymbol{x})$

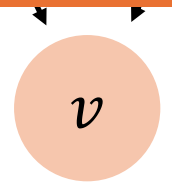
p_2 processes $\boldsymbol{x}$ → sends z_2 + explanation $E_\delta(f_{p_2}, \boldsymbol{x})$

v uses z_1 and z_2 to produce output $f_v(\boldsymbol{x})$

v uses $E_\delta(f_{p_1}, \boldsymbol{x})$ and $E_\delta(f_{p_2}, \boldsymbol{x})$ to produce explanation $E_\delta(f_v, \boldsymbol{x})$

Passing explanations along the AISC

The issue is: Explanations are usually empirically approximated



p_1 processes $\boldsymbol{x}$ → sends z_1 + explanation $\hat{E}\{_\delta(f_{p_1}, \boldsymbol{x})$

p_2 processes $\boldsymbol{x}$ → sends z_2 + explanation $\hat{E}_\delta(f_{p_2}, \boldsymbol{x})$

v uses z_1 and z_2 to produce output $f_v(\boldsymbol{x})$

v uses $\hat{E}_\delta(f_{p_1}, \boldsymbol{x})$ and $\hat{E}_\delta(f_{p_2}, \boldsymbol{x})$ to produce explanation $\hat{E}_\delta(f_v, \boldsymbol{x})$

Explanation accuracy degrades

Goal. Downstream org v must generate explanation at $\boldsymbol{x}$

Supply chain. Ancestor tree of v is m-regular with depth d

Information sharing. Each ancestor a computes an explanation with Δ_a error (entries are independent with variance ε)

Theorem. There exist mappings $\{h_a : \text{ancestors } a\}$ such that

$$\mathbb{E}_{\{\Delta_a\}} \left\| \hat{E}_\delta(f_v, \boldsymbol{x}) - E_\delta(f_v, \boldsymbol{x}) \right\|_{\mathrm{F}} = \Omega\left(\varepsilon m^d\right).$$

Related work

Bullwhip effect [Lee et al. 1997]

Error propagation in numerical analysis [Gautschi & Klein 1967]

Differences in explanation fidelity can lead to unfairness [Balagopalan et al. 2022]

Experimental setup

10-D features, sampled from multivariate Gaussian

Labels generated using noisy linear model

Linear supply chain (from length 1 to 5)

Each model learns independently

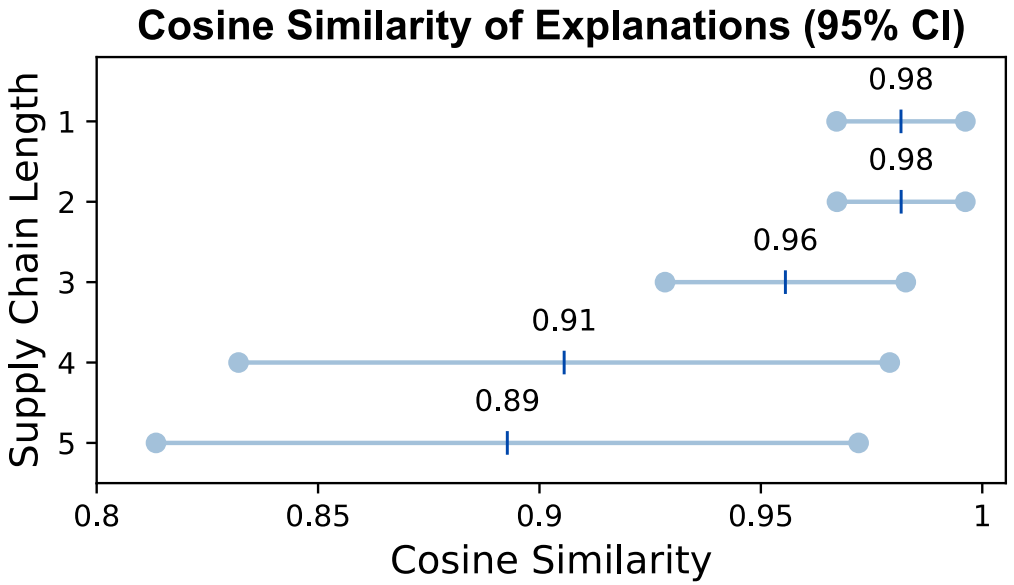
- Depends on the predictions of its parent, creating step-by-step regression

Each model is MLP with 3 fully connected layers and ReLU activation

- Trained using MSE and Adam optimizer, 0.001 LR, 20,000 epochs

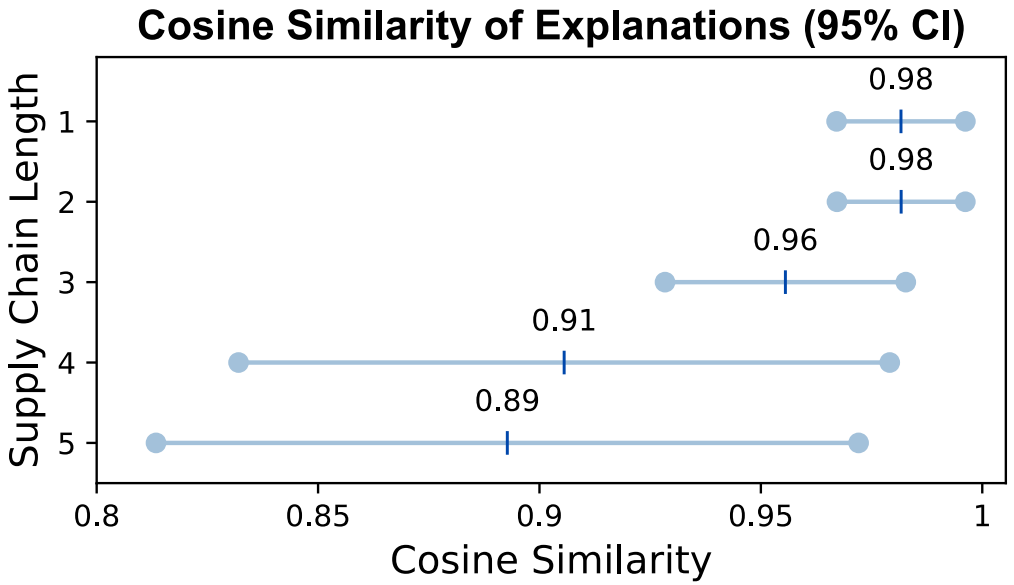
Explanations generated using LIME

Results

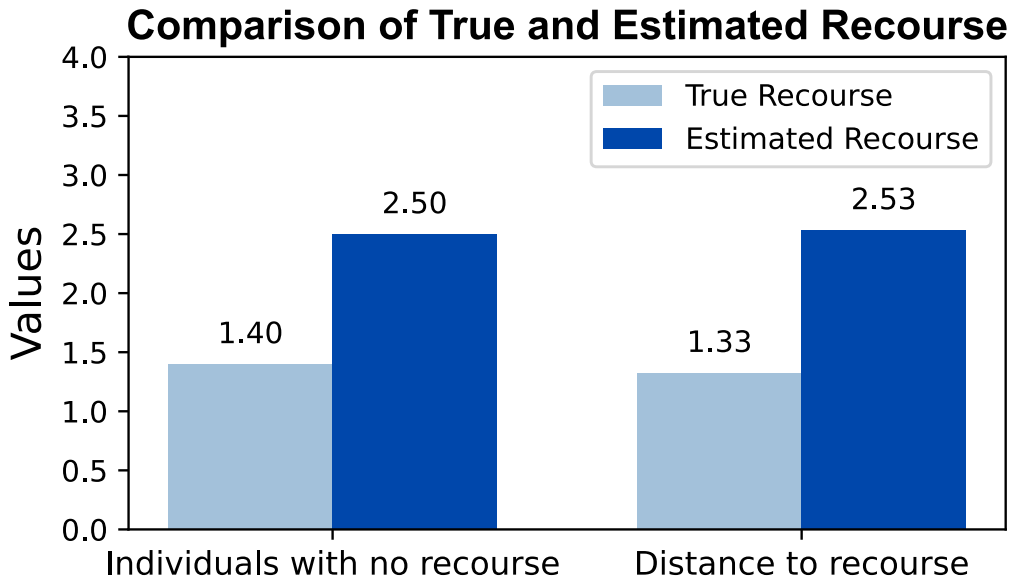


Cosine similarity between estimated explanation and true explanation (true = if v could probe all upstream models)

Results



Cosine similarity between estimated explanation and true explanation
(true = if v could probe all upstream models)



Recourse = how far applicant would need to move along direction indicated by explanation to flip prediction

[Wacther, Mittelstadt, Russell HJLT'17]

Evidentiary burdens in legal cases against AI decisions



Ongoing work with Ananya Karthik, Daniel E. Ho, and Percy Liang

Evidentiary burdens

Our legal systems operate under burden of proof

For instance, criminal cases apply the well-known "guilty beyond reasonable doubt" standard

In a previous paper on AI Auditing, Rohan Alur and I connected this to hypothesis testing [CA EAAMO'24]

Evidentiary burdens

Our legal systems operate under burden of proof

For instance, criminal cases apply the well-known "guilty beyond reasonable doubt" standard

Can we close the gap on evidentiary burdens?

Let's look at Title VII as a case study…

Brief intro to Title VII

Title VII of US CRA “prohibits employment discrimination based on race, color, religion, sex and national origin”

Brief intro to Title VII

Title VII of US CRA “prohibits employment discrimination based on race, color, religion, sex and national origin”

There is a complex procedure, but we can simplify as follows:

Brief intro to Title VII

Title VII of US CRA "prohibits employment discrimination based on race, color, religion, sex and national origin"

There is a complex procedure, but we can simplify as follows:

1. The plaintiff must first establish **disparate impact**

    This typically involves showing that, e.g., female applicants receive worse outcome on average than male applicants

    Quantifying disparate impact is the subject of 10+ years of research

Brief intro to Title VII

Title VII of US CRA "prohibits employment discrimination based on race, color, religion, sex and national origin"

There is a complex procedure, but we can simplify as follows:

1. The plaintiff must first establish **disparate impact**
2. The defendant can respond by showing **business necessity**

    Discrimination is inevitable for business purposes

    e.g., the police must be sufficiently strong and agile

Brief intro to Title VII

Title VII of US CRA "prohibits employment discrimination based on race, color, religion, sex and national origin"

There is a complex procedure, but we can simplify as follows:

1. The plaintiff must first establish **disparate impact**
2. The defendant can respond by showing **business necessity**
3. The plaintiff can then prove there is **a "less discriminatory alternative"** ← the burden of proof is on the plaintiff

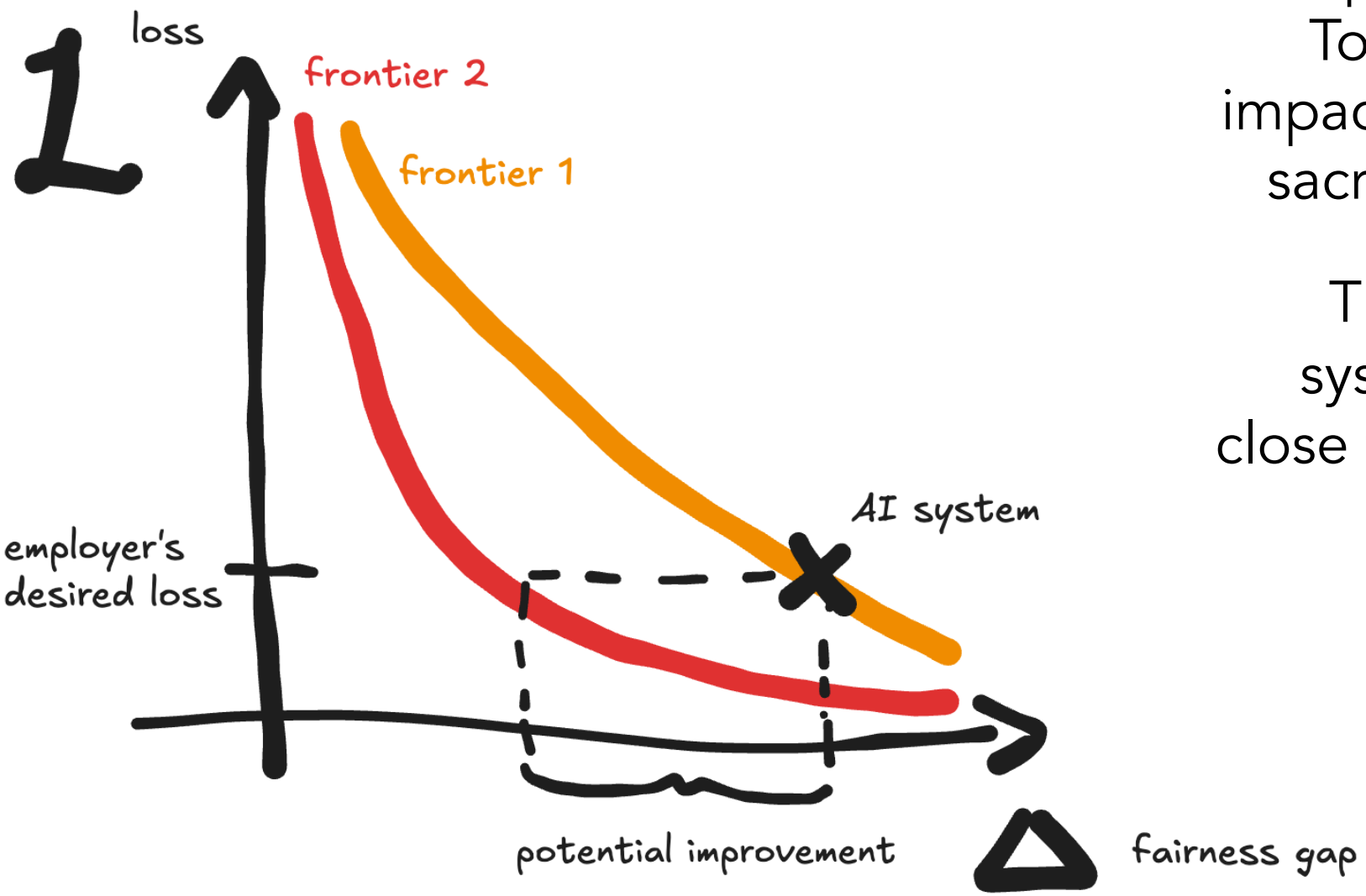
"Less discriminatory alternative"

This is a tall order for AI systems! A straightforward interpretation is that the plaintiff produce an algorithm that performs just as well but is less discriminatory.

Why not shift the burden of proof? It's unlikely longstanding statute on procedure will change

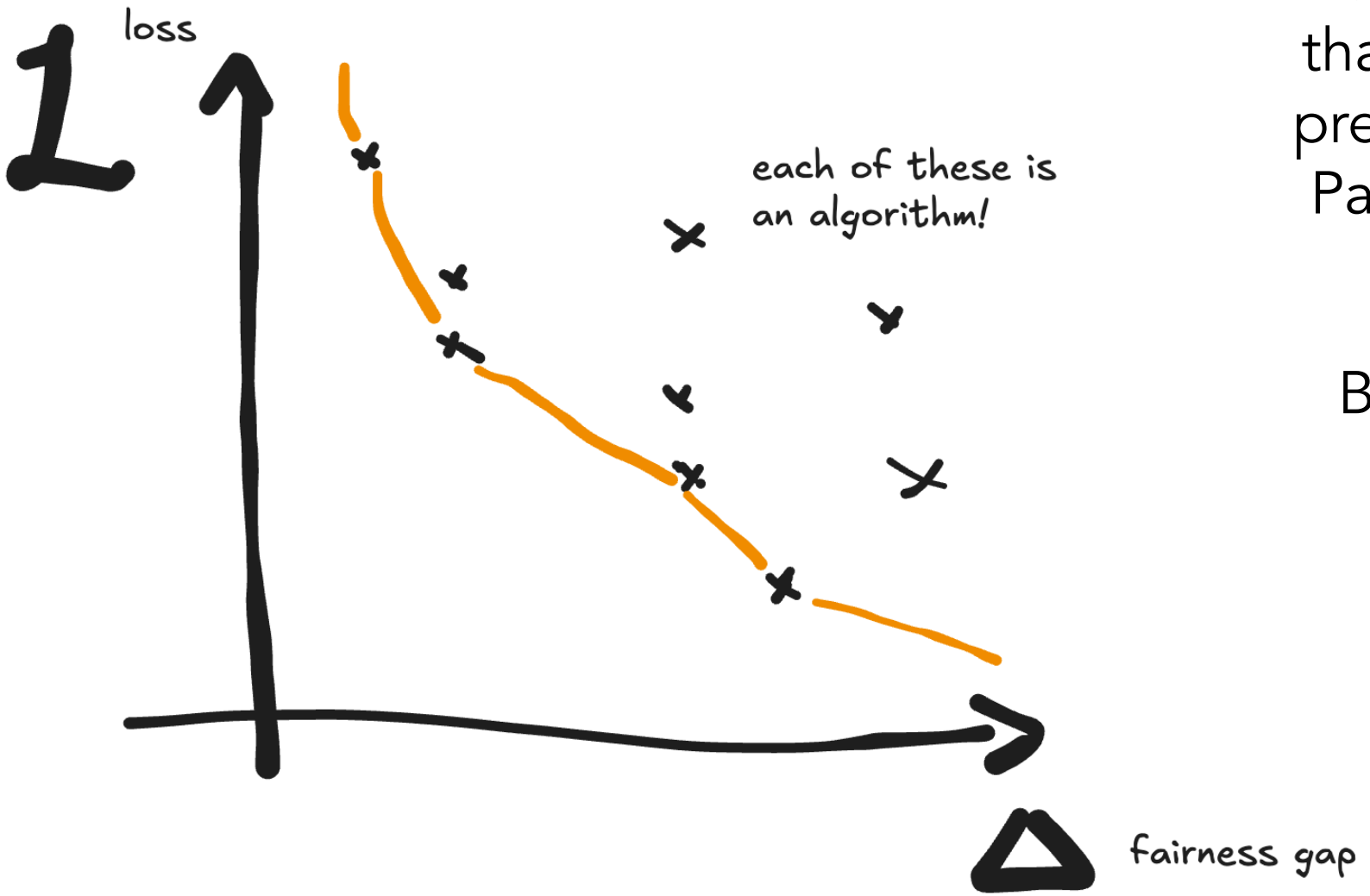
Two reasons why this burden is typically high:

1. The plaintiff has **limited knowledge** of the AI system
2. The plaintiff has **limited expertise and resources**



Employer's argument: To reduce disparate impact, AI system must sacrifice performance

The claim is that AI system is sufficiently close to Pareto frontier!

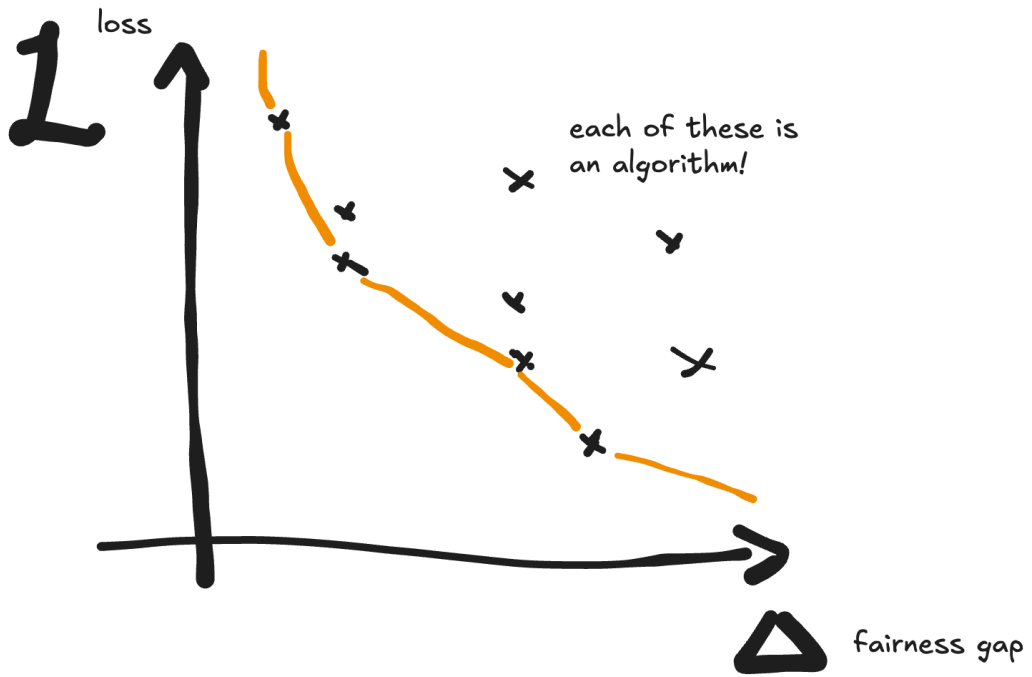


If the plaintiff argues that the AI system is pretty "far" from the Pareto frontier, then they'd be done

But how would the plaintiff find the Pareto frontier?

We can do better!

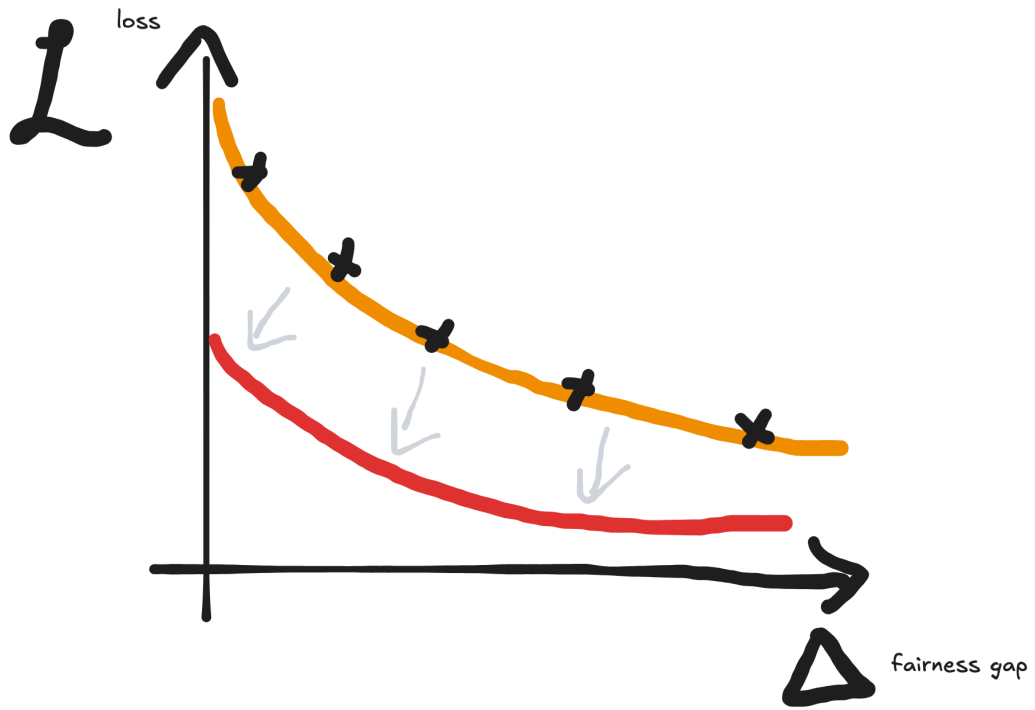
We can characterize the shape of the Pareto frontier



This is important b/c Pareto frontiers do not need to be convex/concave!

We can do better!

We show one can extrapolate the true Pareto frontier from easy tasks (at least, much easier than training state-of-the-art systems!)



Main result

X, A, Y = covariates, sensitive attribute, outcome RVs

$\mathrm{BCE}(p, \hat{f})$ = binary cross entropy of $\hat{f} \in \mathcal{F}$ on p

D = training dataset

Δ = fairness gap (demographic parity gap for our analysis)

Source of unfairness: Assume unfairness is due to selection bias

Main result

X, A, Y = covariates, sensitive attribute, outcome RVs

$\mathrm{BCE}(p, \hat{f})$ = binary cross entropy of $\hat{f} \in \mathcal{F}$ on p

D = training dataset

Δ = fairness gap (demographic parity gap for our analysis)

Theorem (informal):

$$\mathrm{BCE}(p, \hat{f}) \approx C_1(p) + C_2 \log\left(\frac{C_3}{\Delta}\right) + C_4(\mathcal{F}, D)(\Delta + C_5)$$

Why is this useful?

Theorem (informal):

$$\mathrm{BCE}\left(p, \hat{f}\right) \approx C_1(p) + C_2 \log\left(\frac{C_3}{\Delta}\right) + C_4(\mathcal{F}, D)(\Delta + C_5)$$

There are 5 constants! 4 do not depend on $\mathcal{F}, D$!

That means we can fit 4 constants on small models/datasets

Helps to address the resource & expertise problem plaintiffs face!

For the C_4, use scaling laws!

Extensions

This is an ongoing work and this is our first main result, but we anticipate extending the Pareto frontier calculation to other types of fairness and other sources of unfairness.

Ongoing experiments w/ promising results (coming soon!)

Thoughts & feedback are very welcome

Wrapping up

Two directions as AI adoption rises:

1. Adjust our understanding **of AI to its integration into society**
2. Adjust our understanding **of society as it integrates AI**

Today: We examined both perspectives

AI supply chains → how ML targets change in AISCs

Evidentiary burdens → reducing burden of proof in AI cases

Thank you!

shcen@stanford.edu