POSITION PAPER · COMPUTATIONAL JUSTICE

Beyond Sensitive Attributes, ML Fairness Should Quantify Structural Injustice via Social Determinants

Zeyu Tang[1] **Alex John London**[2] **Atoosa Kasirzadeh**[2] **Sarah Stewart de Ramirez**[3] **Peter Spirtes**[2†] **Kun Zhang**[2,4†] **Sanmi Koyejo**[1†]

[1] Stanford University [2] Carnegie Mellon University [3] OSF HealthCare [4] Mohamed bin Zayed University of Artificial Intelligence · † Equal senior authorship

Stanford University · Carnegie Mellon University · OSF HealthCare · MBZUAI

Our Paper

WHY IT MATTERS

Algorithmic fairness research has largely framed unfairness as discrimination along **sensitive attributes**.

ML fairness should quantify structural injustice through social determinants, and audit it before mitigation.

- **Sensitive attribute:** a trait of the individual
- **Social determinant:** a character of contextual environment

01 Structural Justice: The Reframe

SENSITIVE ATTRIBUTES

Traits that belong to a person: race, sex, age, disability status, etc. Legally protected, relatively stable individual-level traits.

SOCIAL DETERMINANTS

Conditions of a place or institution: school funding, clean air, access to care, local policy. Shared by **everyone** in that context.

- **Same** environment, **different** identities → **similar** barriers.
- **Same** identity, **different** environments → **different** barriers.

Variable	Context-level definition?	Social-structural content?	Exogenous stratification?	Categorization
Applicant's race	**No**	–	–	**Sensitive attribute**
Zip code	Yes	**No**	Yes	Not a social determinant
Racial composition (w.r.t. HOLC-redlined district)	Yes	Yes	**No**	**Proxy for sensitive attribute**
School funding per pupil (zip code area)	Yes	Yes	Yes	**Social determinant**

02 How Current Tools Miss It

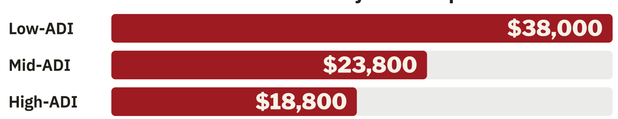
1 **Contextual environment gets thrown away.**
Contextual variables are routinely dropped from fairness datasets, stripping the structural injustice signal.

2 **Individual's identity is stable; context isn't.**
A fixed label like "Black women" can't track how disadvantage shifts across neighborhoods and clinics.

3 **Causal models omit contextual environment.**
Causal fairness graphs draw edges from individual attributes, not the surrounding contextual environment.

03 The Evidence

Median income of Black women by area deprivation:



Same race and sex; higher ADI means more deprivation. U.S. census.

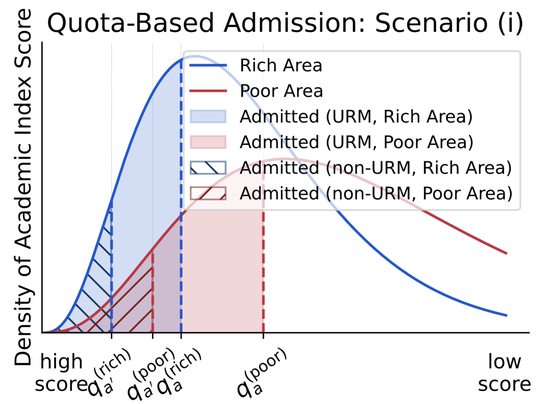
White women in poorer areas start breast-cancer screening **3 to 5 years later** under one guideline. Adopt the richer area's pattern and detections rise, at no extra cost:

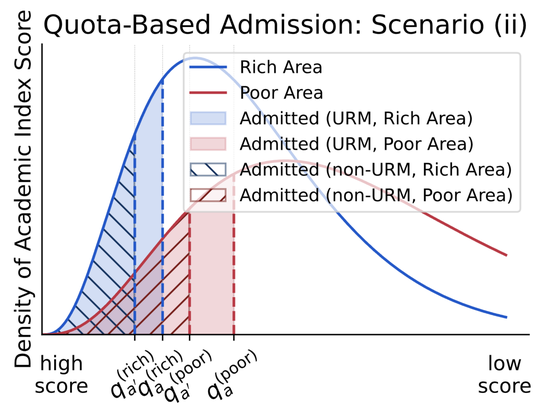
under semi-synthetic simulation

1,367 → 1,461 early detections / 10k ppl

04 A Paradox

The **backfire** is least likely when injustice is **most severe**, and **grows as things get fairer**.





TAKEAWAY

Fairness mitigation centered around sensitive attributes can create new structural injustice.

05 Alternative Views

"Just treat it as another sensitive attribute."

No. These are structural forces to audit, not noise to normalize away.

"The causal effect already covers social determinants."

No. An identity's effect isn't intervening on a school, clinic, or policy. Place often isn't downstream of identity.

06 What to Do?

PILLAR 1 · GOVERN DATA

- Keep raw identifiers locked and audited
- Give models privacy-preserving measures of place: deprivation, pollution, access

PILLAR 2 · REDESIGN METRICS

- Measure fairness against conditions, not identity alone
- **Social Determinant Parity** tracks disparities over time

PILLAR 3 · MODEL LEVERS

- Model outcomes over both people and their contexts
- Test which structural changes actually help

OUR POSITION

Beyond **sensitive attributes**, ML fairness should quantify **structural injustice via social determinants**.