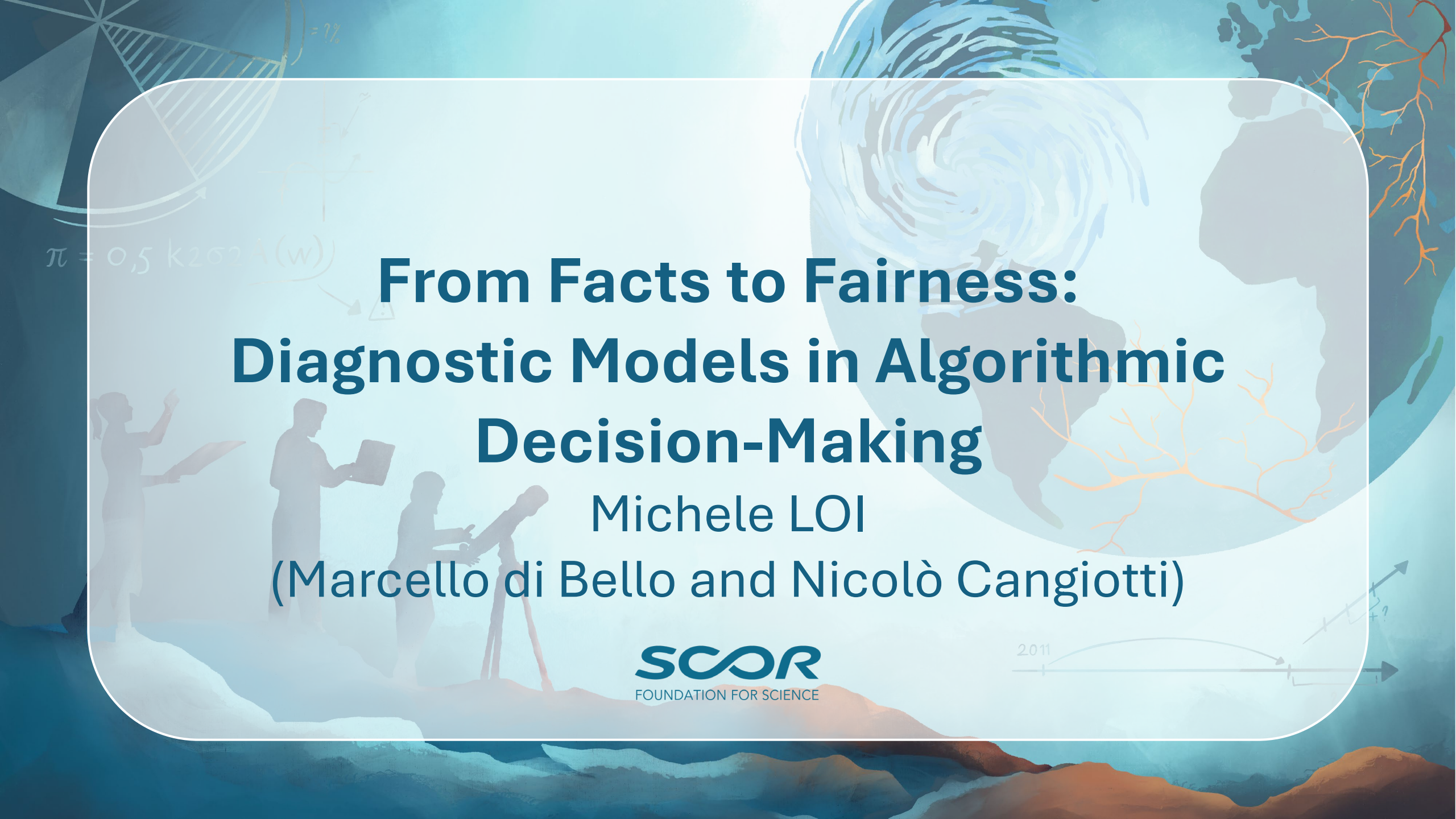
From Facts to Fairness: Diagnostic Models in Algorithmic Decision-Making

Michele LOI
(Marcello di Bello and Nicolò Cangiotti)



From Facts to Fairness: Diagnostic Models in Algorithmic Decision-Making

Michele Loi

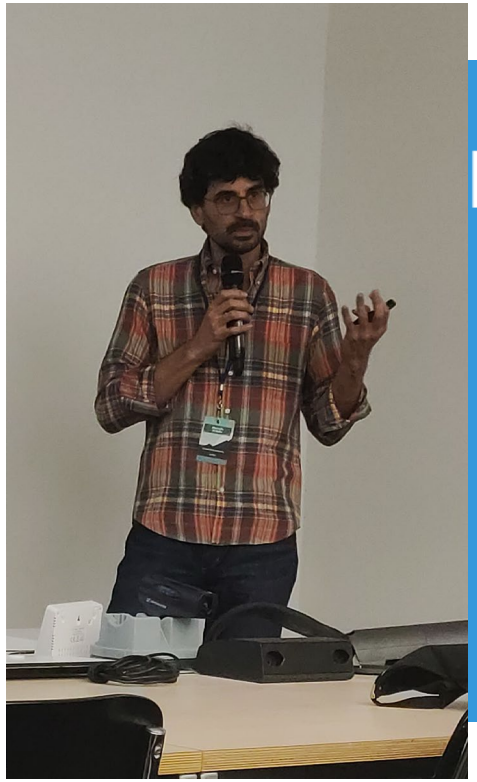
Independent Consultant/Algorithmwatch* (Berlin, Germany)

Confidence and Fairness: Scientific Foundations in AI and Risk

Paris May 15th, 2025

**Personal views, not AW's views*

Latest turn in my research, in collaboration with



Marcello Di Bello

Associate Professor of Philosophy
Arizona State University

Research | Teaching

mdibello@asu.edu | CV

www.nicolocangiotti.eu

Nicolò Cangiotti

Home page

Papers

Teaching

Download

Workshops and Confereces

Outreach

Academic Net

Paper #1

Computer Science > Computers and Society

[Submitted on 19 Feb 2024 (v1), last revised 5 Feb 2025 (this version, v4)]

Causal Equal Protection as Algorithmic Fairness

Marcello Di Bello, Nicolò Cangiotti, Michele Loi

By combining the philosophical literature on statistical evidence and the interdisciplinary literature on algorithmic fairness, we revisit recent objections against classification parity in light of causal analyses of algorithmic fairness and the distinction between predictive and diagnostic evidence. We focus on trial proceedings as a black-box classification algorithm in which defendants are sorted into two groups by convicting or acquitting them. We defend a novel principle, causal equal protection, that combines classification parity with the causal approach. In the do-calculus, causal equal protection requires that individuals should not be subject to uneven risks of classification error because of their protected or socially salient characteristics. The explicit use of protected characteristics, however, may be required if it equalizes these risks.

Comments: 18 pages, 7 figures
Subjects: **Computers and Society (cs.CY)**; Artificial Intelligence (cs.AI); Data Structures and Algorithms (cs.DS); Machine Learning (cs.LG)
Cite as: arXiv:2402.12062 **[cs.CY]**
(or arXiv:2402.12062v4 **[cs.CY]** for this version)
https://doi.org/10.48550/arXiv.2402.12062

Submission history

From: Nicolò Cangiotti [view email]
[v1] Mon, 19 Feb 2024 11:30:00 UTC (231 KB)
[v2] Tue, 20 Feb 2024 22:53:23 UTC (231 KB)
[v3] Thu, 8 Aug 2024 08:36:03 UTC (238 KB)
[v4] Wed, 5 Feb 2025 11:33:07 UTC (35 KB)

Paper #2

- Develops the idea of *spatio-temporal specificity*
- Argues for attention to spatio-temporal **specific** causal elements in prediction
- Explores the conditional probability of the **specific** event given the **generic** event

This talk will become paper #3

The Fairness Puzzle: Competing Frameworks

Three dominant approaches to algorithmic fairness:

- **Statistical parity**: Ensuring equal outcomes across demographic groups
- **Calibration**: Ensuring predictions mean the same thing across groups
- **Causal fairness**: Removing causal influence of protected attributes
- **Key tension**: These approaches often conflict with each other and with human intuitions about fairness

The Case-Specific Evidence Paradox

Consider two scenarios involving John, from a disadvantaged group:

1. **Generic profiling:** John is suspected because his demographic group statistically commits more crimes
2. **Case-specific evidence:** A man with a certain profile is witnessed near a crime scene

Both lead to the same outcome, but feel morally different. Why?

Existing Approaches Fall Short

- **Lippert-Rasmussen:** Statistical discrimination is wrong when those who benefit created conditions making it useful [1]

- **Hellman:** Statistical discrimination is wrong when it compounds prior injustice [2]

[1] Lippert-Rasmussen, K. 2007. Nothing Personal: On Statistical Discrimination*. *Journal of Political Philosophy*. 15, 4 (Dec. 2007), 385–403. DOI:https://doi.org/10.1111/j.1467-9760.2007.00285.x.

[2] Hellman, D. 2023. Big Data and Compounding Injustice. *Journal of Moral Philosophy*. 21, 1–2 (2023), 62–83. DOI:https://doi.org/10.1163/17455243-20234373.

My/our explanatory pathway account

What distinguishes more unfair bias?

Key insight: Direct causal mechanisms transform explanatory relationships

More unfair bias occurs when:

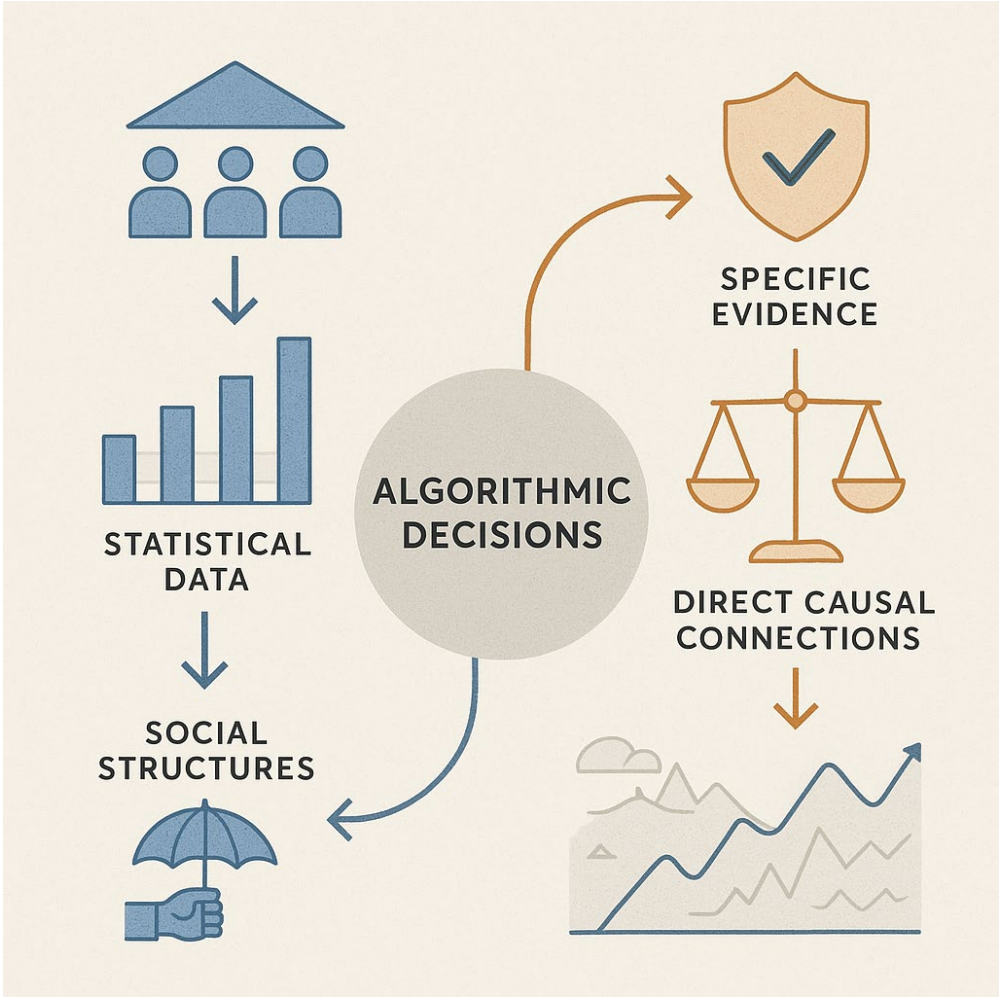
- The primary explanatory pathway runs through unjust social structures (shared with Lippert-Rasmussen/Hellman)
- This pathway is not “displaced” by direct causal mechanisms connecting the individual to the particular event

Explanatory Pathways Make the Difference

- **Generic profiling pathway:**
  Social structures → Prediction

- **Case-specific evidence pathway:**
  Direct causal link

  Specific facts/individuals/spatio-temporal events → Prediction

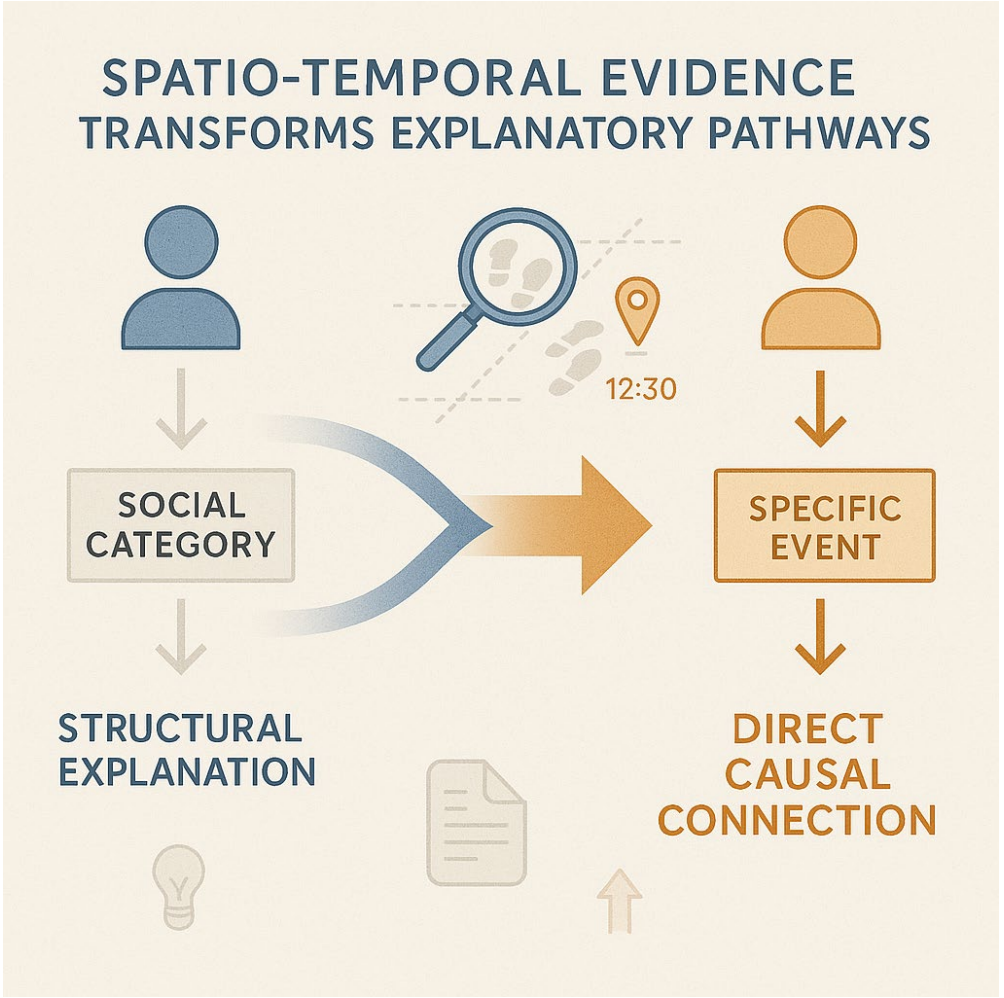


The Spatio-Temporal Specificity Solution

- **Key insight:** Evidence becomes *less unfair* when its explanatory relevance is *not* primarily mediated through unjust social structures, but instead, through a direct causal mechanisms specific to the case.

When e.g. “a black man” is witnessed at the crime scene at a specific time of the day:

- Explanatory pathway runs through direct causal connections
- Not primarily explained by structural factors

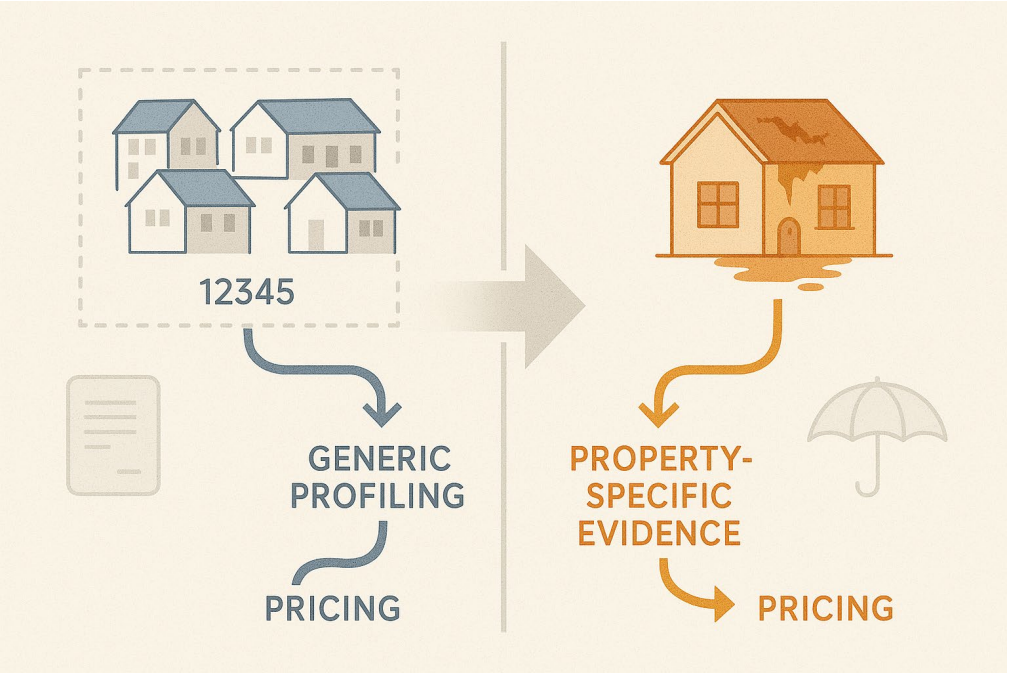


Insurance Applications: When Is Risk Classification Fair?

Consider two insurance scenarios:

1. **Generic profiling:** Charging higher premiums to a neighborhood with predominantly disadvantaged residents
2. **Case-specific evidence:** Charging higher premiums based on documented water damage to a specific property

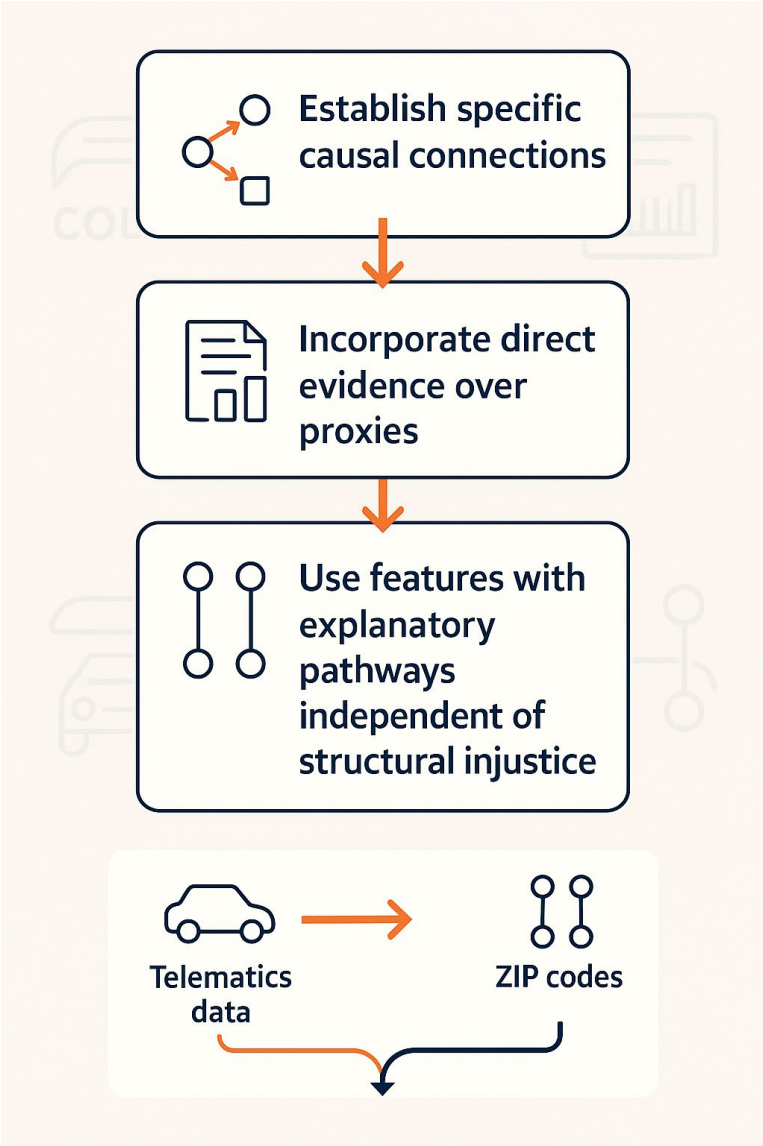
The difference: Explanatory pathways, not just outcomes



Policy Implications for Predictive Models

Models are more likely to be fair (less unfair) when they:

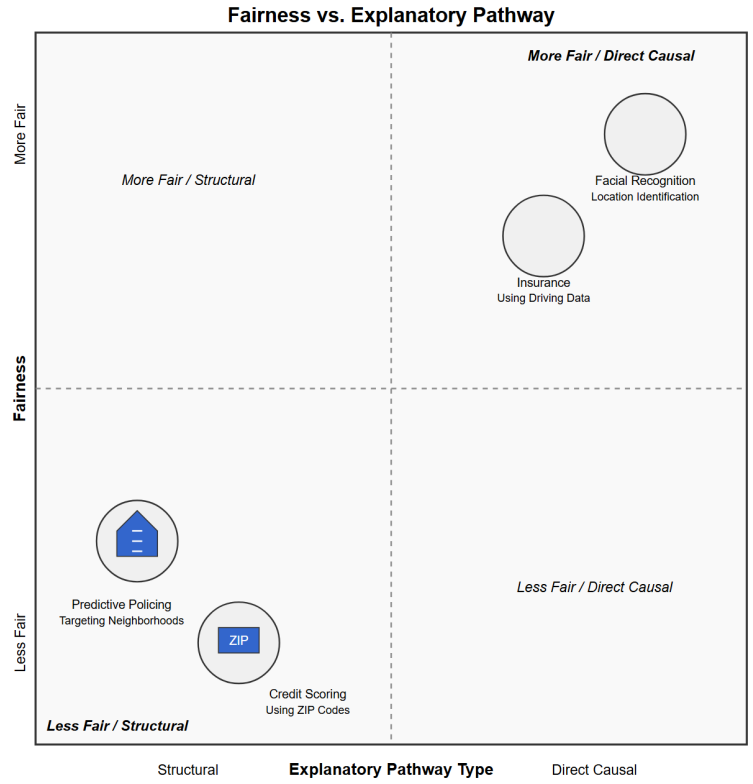
1. Establish **specific causal connections** rather than generic correlations
2. Incorporate **direct evidence** over demographic proxies
3. Use features with **explanatory pathways independent** of structural injustice



Fairness in Algorithmic Systems

Implications for algorithmic systems:

- Facial recognition identifying someone at specific location (more fair)
- Predictive policing targeting neighborhoods based on demographics (less fair)
- Insurance pricing using individual driving data (more fair)
- Credit scoring using ZIP codes that are proxy for race (less fair)



Why Socially Salient Traits Matter More

Socially salient traits (race, gender, etc.) matter more for fairness because:

- Their predictive power stems from social processes (typically) implicated in injustice
- These categories become predictive precisely because of structural patterns
- Naturally occurring variations lack this problematic explanatory connection

Distinguishing Our Approach

The key theoretical contribution is:

- Not just prioritizing disadvantaged groups (Lazar & Stone)
- Not just identifying compounding injustice (Hellman)
- But recognizing the moral significance of **explanatory pathways**

Interdisciplinary Connections

Our framework connects to:

- Insurance risk classification ethics
- Legal evidence theory
- Algorithmic fairness metrics (Chiappa et al, “path-specific counterfactual fairness”
- Epistemology of statistical evidence

Practical Guidelines for Fairer Predictive Models

1. **Evaluate explanatory pathways**, not just statistical performance
2. **Prioritize features with direct causal connections** to the predicted variable
3. **Be skeptical of features that derive predictive power** from structural injustice
4. **Consider the nature of evidence**, not just its statistical strength

Challenges & Future Work

- **Empirical identification:** How to trace explanatory pathways in practice
- **Quantitative implementation:** Translating the theory into algorithmic constraints
    - **Theoretical mathematical framework:** necessary for algorithmic exploration/implementation

Conclusion: Beyond Statistical Fairness

Key takeaway: Understanding fairness requires looking at how predictions are explained, not just what they predict

- Shifts focus from outcomes to explanatory mechanisms
- Provides principled distinction between less unfair and more unfair evidence
- Offers practical guidance for algorithm design

Thank You

Michele Loi

Algorithmwatch

m.loi@icloud.com

Questions & Discussion