Theory, Not Practice: False Promises & False Premises of Fair Machine Learning

Nic Fishman

October 31, 2025

Outline

1. Introduction

2. A history of the formalism of fair machine learning

3. Causal problems for the canonical formalism

4. Why the formalism cannot change

5. Conclusions

Introduction

The Machine Learning Society

Members of technologically modern society will encounter (& be affected by) ML-based ADM/ADS systems on a daily basis

- Admissions and hiring
- Sentencing, Bail decisions, CPS decisions, and social welfare programs
- Health insurance and care decisions
- Lending decisions, credit card approvals, mortgages

Enter ML Fairness

- Integration of ML into society came with new ways that ML could do harm, but also with new opportunities to do good.
- ML fairness emerged within theoretical computer science, predicated on the idea that as ML swept across society it was actually an opportunity to make the world a better place.

What is the goal of fair ML?

- Introduce formal method/decision procedure that can substitute for critical thinking and domain expertise.
- Easily operationalized/adopted at the institutional level.

Notation

$$
\begin{aligned}
A_i &\equiv \text{ sensitive attribute (e.x. race, gender)}\\
X_i &\equiv \text{ other covariates (e.x. age, employment, income)}\\
Y_i &\equiv \text{ ground truth label to predict}\\
\hat{Y}_i &\equiv \text{ predicted label}
\end{aligned}
$$

A history of the formalism of fair machine learning

2012: "Fairness Through Awareness" (Dwork et al.)

This was the first major theoretical CS venture into mathematical fairness research. It opened the agenda by rejecting to poles: “unawareness” and statistical parity.

2012: "Fairness Through Awareness" (Dwork et al.)

- A prominent approach is to exclude race and gender from the input features, premised on legal notions of anti-discrimination.
- Dwork et al. argued that blindness to sensitive attributes perpetuates discrimination since natural data always contains proxies for sensitive attributes.

2012: "Fairness Through Awareness" (Dwork et al.)

Distinction between group fairness and individual fairness:

- Statistical parity: statistical parity across demographic groups

$$P(\hat{Y} = 1 | A = a) = P(\hat{Y} = 1 | A = a') \quad \forall a, a'$$

- Individual fairness: similar individuals should be treated similarly

$$d(f(x), f(x')) \leq D(x, x')$$

where D is a task-specific "similarity metric" which is left abstract.

2014: The Founding of FATML

- FATML (Fairness, Accountability, and Transparency in ML) established as a workshop at NeurIPS (the largest AI conference)
- Mission responded to Obama administration's Big Data Report
- Workshop scope statement framed fairness as a technical challenge:

  "How can we achieve high classification accuracy while eliminating discriminatory biases? What are meaningful formal fairness properties?"

FATML Workshop Evolution

- 2014: Initial meeting focused on talks, not papers
- 2015: Still small community—only 4 papers presented
- 2016: Major shift after ProPublica's COMPAS investigation
  - Attendance exploded from 7 speakers to 25
  - Added critical questions: "Are there any dangers in turning fairness into computational problems?"
- The workshop was connected to legal scholarship from the start
  - Brought legal concepts into computational framework
  - Sought computational translations of normative concepts

2016: ProPublica's COMPAS Investigation

- First major investigative journalism on algorithmic discrimination
- COMPAS: Risk assessment tool used in criminal justice
- ProPublica claimed the system was racially biased:
  - Black defendants were more likely to be falsely labeled high-risk
  - White defendants more likely to be falsely labeled low-risk
- Northpointe (creator) claimed system was fair by their metric
- Provided real data and concrete problem for researchers

2016: Inherent Trade-Offs in Fairness (Kleinberg et al.)

- Proved a fundamental mathematical impossibility result that 3 properties cannot be simultaneously satisfied (except in trivial cases):
  1. Calibration within groups:
  $$P(Y = 1|\hat{Y} = s, A = a) = s \quad \forall s \in [0, 1], a \in \mathcal{A}$$
  2. Balance for the positive class:
  $$E[\hat{Y}|Y = 1, A = a] = E[\hat{Y}|Y = 1, A = a'] \quad \forall a, a'$$
  3. Balance for the negative class:
  $$E[\hat{Y}|Y = 0, A = a] = E[\hat{Y}|Y = 0, A = a'] \quad \forall a, a'$$
- Showed that satisfying all three requires either:
  - Perfect prediction ($\hat{Y} = Y$ always)
  - Equal base rates across groups ($P(Y = 1|A = a) = P(Y = 1|A = a')$)

2016: Equality of Opportunity (Hardt et al.)

- Established the canoncial form of the fairness problem
- Framed fairness as a **constrained optimization problem**:
  - Optimize for accuracy/utility
  - Subject to fairness constraints defined by parity metrics
- **Oblivious post-processing**
  - Could take *any* black-box model (treating it as a score R)
  - Create a derived predictor: $\hat{Y}_a = h_a(R)$ for each group a
  - Adjust thresholds for each group to satisfy fairness constraints
  - Required only access to $\{A, Y, R\}$ - not model internals or features

The Canonical Mathematical Formalization

- Fairness as a constrained optimization problem:

  $$\min_{h \in \mathcal{H}} \mathbb{E}[\ell(h(X), Y)] \quad \text{subject to} \quad h \in \mathcal{G}$$

  where $\mathcal{G}$ is the set of functions satisfying fairness constraints
- The **oblivious** setting established as standard framework:
  - Focus on $(A, Y, \hat{Y})$ only, not internal model details or X
  - Post-processing as the preferred intervention method

Causal problems for the canonical formalism

The Causal Perspective on Fairness Evaluation

- Drawing on ideas from causal inference we can explore a number of failures of the canonical formalization:
  1. **Identification**: The population we observe represents the target population
  2. **Non-heterosketatic measurement error**: We observe the actual treatments and outcomes, or at least the measurement error is similar
  3. **Unconfoundedness**: All confounders between treatment and outcome are observed
  4. **SUTVA** (Stable Unit Treatment Value Assumption): Outcomes of one individual don't affect outcomes of others
- "Fairness" is not a property of a model, it is a property of a model along with our assumptions on the data.

Exclusion 1: Proxy Bias

- **Causal interpretation**: Differential measurement error
- Violates the perfect measurement assumption
- A_P (proxy) is a biased measure of A_T (true attribute)
- Y_P (proxy) is a biased measure of Y_T (true outcome)
- A affects the measurement process itself
- **Mathematical implication**:

$$P(A_P|A_T = a) \neq P(A_P|A_T = a')$$
$$P(Y_P|Y_T, A = a) \neq P(Y_P|Y_T, A = a')$$

Exclusion 2: Selection Bias

- **Causal interpretation**: Collider bias
- Violates the representativeness assumption
- Conditioning on S (selection) induces spurious correlations
- Creates a non-random sample that distorts population inferences
- **Mathematical implication**:

$$P(Y|A, X, S = 1) \neq P(Y|A, X)$$

Exclusion 3: Extra-Classificatory Policies

- **Causal interpretation**: Post-treatment variables
- Policy D is a mediator between protected attribute A and outcome Y
- Violates unconfoundedness by ignoring this causal pathway
- **Mathematical implication**:

$$P(Y|A = a) \neq P(Y|do(A = a))$$

Exclusion 4: Cohort Effects

- **Causal interpretation**: SUTVA violation
- Stable Unit Treatment Value Assumption is violated
- Individual i's outcomes depend on treatment of individual j
- Analogous to interference, spillover effects in social science
- Ignored in standard fairness formulations
- **Mathematical implication**:

$$Y_i = f(X_i, A_i, \{(X_j, A_j)\}_{j \neq i})$$

Loan Approval

An Automated Loan-Approval Algorithm:
A bank automates loan approval using data on loan repayment to fit a model predicting likelihood of repayment. The bank also sets interest rates using a separate model that considers factors including borrower's address, which correlates with race.

- **Selection Bias (Causal Identification)**:
  - Banks observe data for people they lend to making ML fairness difficult
  - But banks *decide* who to give loans to, creating the selection
- **Policy Bias (Post-treatment Variables)**:
  - Bank higher interest rates for minority neighborhoods
  - These rates causally affect repayment probability
  - Then uses resulting "base rate differences" to justify discrimination

Algorithmic Hiring

An Algorithmic Hiring System:
A company receives thousands of applications for every job they advertise. To handle this, they build an ML-based system to sift through candidates by predicting employee performance review scores from resumés. They check the model against standard fairness metrics on held-out data.

- **Proxy Bias (Measurement Error)**:
  - Performance reviews reflect discriminatory culture within the firm
- **Policy Bias (Post-treatment Variables)**:
  - Organizational policies set salaries, mentoring, advancement opportunities, and resources; imbalances create disparate outcomes
- **Cohort Bias (Interference)**:
  - A firm hiring majority men and a few women that causally effects the outcomes of the women

Predictive Policing

A Predictive Policing System:
A police department implements a system to predict crime hotspots based on historical arrest and incident data. The system allocates patrols to neighborhoods based on predicted crime rates. The department validates the system by comparing predictions to actual arrest rates.

- **Proxy Bias (Measurement Error)**:
  - Department chooses to use arrests as proxy for crime
  - Could use alternative, less biased crime measures but doesn't
- **Selection Bias (Causal Identification)**:
  - Department *controls* where police are deployed
  - Reinforces the proxy bias in the arrest data
- **Cohort Bias (Interference)**:
  - Incarceration is criminogenic and disproportionately arresting one group reinforces disparate outcomes

Why the formalism cannot change

These forms of bias are ubiquitous

Proxy	Selection	ECP	Any Bias	Any Two	All Three
69%	85%	85%	100%	92%	61%

- Analysis of 14 commonly used datasets from various domains (finance, healthcare, employment, criminal justice)
- These are not edge cases—they represent the norm in real-world data

Many of these issues have been clear almost from the start

- Assumes access to unbiased outcomes (Y) despite known proxy biases [Bao et al., 2021; Fogliato et al., 2020]
- Assumes access to the relevant population despite selection bias [Lakkaraju et al., 2017; Kallus & Zhou, 2018]

Disciplinary Constraints

- Addressing these concerns is not possible while maintaining theoretical elegance. They require careful, context-specific attention.
- TCS is a mathematical discipline, so any setting that precludes straightforward mathematical formalism cannot be incorporated.
- The field has now begun to fragment, with the more critical work being published at FAccT and theory moving to the FORC conference.

A failed political economy

Core premise: firms want to be fair all things equal but

1. do not know exactly what that would look like
2. do not want to invest significant resources

This elides differing incentive structures across differing contexts

- Adversarial settings (Lending/Hiring/Sentencing): the classifier wants to predict whether the classified will repay the loan/be a good worker/commited a crime; the classified wants the "good" outcome
- Cooperative settings (Healthcare/Fraud detection): both the classified and the classifier want to determine whether the treatment will be effective/the transaction is fraudulent

Adversarial settings and Classifier Agency

- In adversarial settings, fairness is costly unless base rates are similar.
- Firms actually control many of the forms of bias discussed above.
- The canonical form of ML fairness enables firms to claim fairness by causing or widening differences in base rates.

Conclusions

The Fragility of Canonical Fairness Formulations

- The canonical mathematical formalization of fairness:
  - Emerged from theoretical computer science
  - Focused narrowly on parity metrics across demographic groups
  - Became the dominant paradigm in both research and practice
- But canonical formulations ignore crucial causal insights:
  - Assume causal assumptions that are systematically violated in practice
  - Ignore how organizations create and maintain these violations
  - Treat "unfairness" as a property of algorithms rather than sociotechnical systems
  - Position fairness as a technical problem rather than an institutional one

Demographic Parity is all we need?

- Where do we go from here? Complex metrics that depend on observed outcomes are very hard to evaluate.
- We should return to outcome free metrics like demographic parity as a core “gut check” for whether a model is fair.

Collaborators

These ideas emerged from deep collaborations with: Mel Andrews, Jake Fawkes, Andrew Smart, Stephen Pfohl, Zachary Lipton

References

- Bao, M., et al. (2021). It's COMPASlicated: The Messy Relationship between RAI Datasets and Algorithmic Fairness Benchmarks. arXiv preprint arXiv:2106.05498.
- Coston, A., et al. (2020). Counterfactual Risk Assessments, Evaluation, and Fairness. In Proceedings of FAccT.
- Dai, J. and Singh, S. (2021). Fair Machine Learning under Partial Compliance. In Proceedings of AIES.
- Fogliato, R., et al. (2020). Fairness Evaluation in Presence of Biased Noisy Labels. In International Conference on Artificial Intelligence and Statistics.
- Goel, N., et al. (2021). The Importance of Modeling Data Missingness in Algorithmic Fairness: A Causal Perspective. In Proceedings of AAAI.
- Jacobs, A. and Wallach, H. (2021). Measurement and Fairness. In Proceedings of FAccT.
- Kallus, N. and Zhou, A. (2018). Residual Unfairness in Fair Machine Learning from Prejudiced Data. In International Conference on Machine Learning.
- Lakkaraju, H., et al. (2017). The Selective Labels Problem: Evaluating Algorithmic Predictions in the Presence of Unobservables. In Proceedings of KDD.

Thank you for your attention!

Questions?

Technical Appendix

Technical Appendix: Key Fairness Metrics

This appendix provides formal definitions of common fairness metrics used in ML fairness research.

Notation:

- A_i: Sensitive attribute (e.g., race, gender)
- X_i: Other covariates (e.g., age, income)
- Y_i: Ground truth label
- $\hat{Y}_i$: Predicted label

Group Fairness Metrics I

Statistical/Demographic Parity:

$$P(\hat{Y} = 1 | A = a) = P(\hat{Y} = 1 | A = a') \quad \forall a, a'$$

Ensures that the likelihood of a positive prediction is the same across all demographic groups.

Equality of Opportunity:

$$P(\hat{Y} = 1 | Y = 1, A = a) = P(\hat{Y} = 1 | Y = 1, A = a') \quad \forall a, a'$$

Ensures that the true positive rate is the same across all demographic groups.

Group Fairness Metrics II

Equalized Odds:

$$P(\hat{Y} = 1 | Y = y, A = a) = P(\hat{Y} = 1 | Y = y, A = a') \quad \forall y \in \{0, 1\}, \forall a, a'$$

Ensures both true positive and false positive rates are the same across all demographic groups.

Predictive Parity:

$$P(Y = 1 | \hat{Y} = 1, A = a) = P(Y = 1 | \hat{Y} = 1, A = a') \quad \forall a, a'$$

Ensures that the precision (positive predictive value) is the same across all demographic groups.

Group Fairness Metrics III

Calibration within Groups:

$$P(Y = 1|\hat{Y} = s, A = a) = s \quad \forall s \in [0,1], \forall a \in \mathcal{A}$$

Ensures that the probability estimates produced by the model are well-calibrated within each group.

Balance for the Positive/Negative Class:

$$E[\hat{Y}|Y = y, A = a] = E[\hat{Y}|Y = y, A = a'] \quad \forall a, a', y \in \{0,1\}$$

Ensures that the average predicted score for each true outcome class is the same across demographic groups.

Individual Fairness Metrics

Individual Fairness (Dwork et al.):

$$d(f(x), f(x')) \leq D(x, x')$$

Ensures that similar individuals receive similar predictions, where:

- d is a distance metric in the prediction space
- D is a task-specific similarity metric in the feature space
- f is the prediction function

Counterfactual Fairness:

$$P(\hat{Y}_{A\leftarrow a}(U) = y|X = x, A = a) = P(\hat{Y}_{A\leftarrow a'}(U) = y|X = x, A = a)$$

Ensures that predictions would be the same in the counterfactual world where an individual's protected attribute were different.

Impossibility Results

Kleinberg et al. Impossibility Theorem:
It is impossible to simultaneously satisfy the following three fairness criteria except in trivial cases:

1. Calibration within groups
2. Balance for the positive class
3. Balance for the negative class

These three properties can only be simultaneously satisfied if either:

- The predictor is perfect ($\hat{Y} = Y$ always), or
- Base rates are equal across groups ($P(Y = 1|A = a) = P(Y = 1|A = a')$)

Fairness Metrics: Tradeoffs

Metric	**Key Tradeoffs**
Demographic Parity	May reduce accuracy when base rates differ; ignores potential justified disparities
Equality of Opportunity	Focuses only on true positives; may permit disparate impact for negative cases
Equalized Odds	Strongest constraint; often comes with greatest accuracy cost
Calibration	Compatible with optimal prediction; appropriate for how
Individual Fairness	Moves the whole problem a stage deeper, to the definition of the similarity metric