Proxy Bias in AI: How Should We Solve "Equal Opportunity" Laws Clashing With Equitable Outcomes?

Avazjon Yusufjonov, 1980230

UNIVERSITÄT BAYREUTH

Background

Apple Card Scandal: In 2019, Apple's co-branded credit card (Goldman Sachs) systematically assigned women lower credit limits than men with identical financial profiles triggering widespread accusations of employing sexist practices and public backlash that resulted in an New York Department of Financial Services inquiry under U.S. fair-lending laws. Apple and Goldman Sachs said that they did not use "gender" as an input for training the algorithm nor for reviewing the loan application thus the algorithm cannot discriminate against specific gender[1].

DHH ✓
@dhh

The @AppleCard is such a f king sexist program. My wife and I filed joint tax returns, live in a community-property state, and have been married for a long time. Yet Apple's black box algorithm thinks I deserve 20x the credit limit she does. No appeals work.

Steve Wozniak ✓
@stevewoz

@AppleCard The same thing happened to us. I got 10x the credit limit. We have no separate bank or credit card accounts or any separate assets. Hard to get to a human for a correction though. It's big tech in 2019.

Regulatory Blindness: Indeed, under the U.S. laws (ECOA, FHA) prohibit explicit use of race/gender; GDPR likewise restricts "special category" data in the EU. Thus, no company (including Apple and Goldman Sachs) can use such protected characteristics in their business decision making processes nor to even improve the products they offer. These laws assume that attribute-blind algorithms guarantee unbiased results and provide equal opportunities.

Blind Does Not Mean Unaware: Even when the algorithm cannot see some characteristics explicitly, it does not mean that it cannot infer that implicitly based on other variables. This is why the premise that the unbiasedness of attribute-blind algorithms is shattered by the problem of proxy variables.

Problem of Proxy Variables

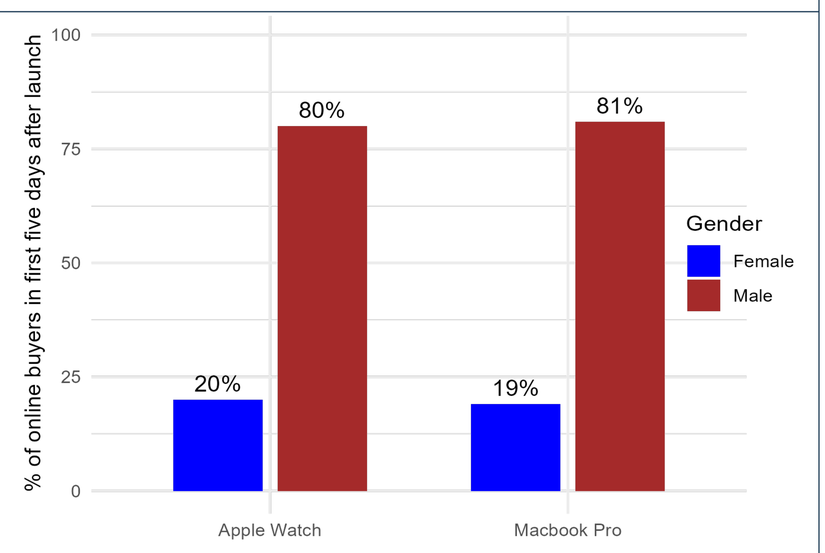
What is Proxy Variable? A proxy variable is a measured variable used instead of a desired input variable that is usually unobservable.

Proxy Resurgence: When protected attributes are omitted, Artificial Intelligence algorithms leverages correlated features (e.g. ZIP code, education, purchase patterns) as stand-ins, creating discriminative impacts under regulators' radar. Thus, even "blind" models can create biased outcomes.

Example

Finding Proxy Variable: In 2015, Slice Intelligence conducted an analysis regarding gender distribution of purchases of newly released apple products i.e. Apple Watch and Macbook Pro. The analysis revealed that 80% of purchases made by male consumers.

Empirical Results: The research shows that Macbook users have the lowest default rate on loans they take[2].



Outcome: Now, if we feed the empirical findings into the algorithm, it will give more loans to the people who have recently purchased Mac products (i.e. mainly men), creating disparate outcomes.

Solutions[3]

Regulatory

Bias-Impact Statements: Mandate disclosures of demographic breakdowns (collected or imputed), pre/post fairness metrics, and quantitative mitigation summaries .

Sandboxes & Safe Harbors: Allow temporary exemptions to collect sensitive data solely for audit and mitigation, insulated from enforcement actions .

Certification & Standards: Develop ISO-style fairness certifications and interoperable GDPR/anti-discrimination APIs to standardize encrypted attribute handling and third-party audits .

Technical

Pre-processing

- **Targeted Feature Suppression**: Identify top 5–10 proxy features (via mutual information) and removing them cuts disparities > 40 % with < 1 % accuracy loss .

In-processing

- **Causal Deconfounding**: Project out sensitive-attribute subspace from input embeddings via adversarial training, preventing reconstruction of gender proxies.
- **Fairness Regularizers**: Add error-rate-parity penalties to loss functions which can trace accuracy–fairness balance.

Post-processing

- **Equalized Odds Adjustment**: Flip minimal predictions for advantaged groups to align false rates—restores parity at modest utility cost .
- **Explainability Enhancements**: Augment local explanations with proxy-correlation flags by highlighting features whose high importance signals strong protected-group links.

Social

Cross-functional Teams & Inclusive Design: Combine engineers, ethicists, lawyers, and affected-community representatives early to surface context-specific proxies and ethical trade-offs .

Algorithmic Literacy & Feedback: Publish public guides on proxy bias, maintain consumer hotlines for reporting suspicious outcomes, and host transparency portals with live fairness dashboards

Conclusion

- **Equal-Opportunity Blindness Is Insufficient.** Regulations that ban explicit use of protected attributes can't prevent AI from finding them via proxies, resulting in hidden discriminatory actions.
- **Integrated Solutions.** Combining regulatory reforms (bias-impact statements, safe harbors), technical safeguards (proxy suppression, deconfounding, post-processing), and social practices (inclusive design, algorithmic literacy) can close the equity gap.
- **Call to Action.** Policymakers, practitioners, and researchers must collaborate to pilot cross-jurisdictional sandboxes, develop robust similarity metrics, and scale explainability and disclosure interventions, ensuring AI decisions delivers not just equal opportunity, but equitable outcomes for all.

References

1. Knight, Will. "The Apple Card Didn't 'See' Gender—and That's the Problem." *WIRED*, 19 Nov. 2019, www.wired.com/story/the-apple-card-didnt-see-genderand-thats-the-problem.
2. Berg, Tobias, et al. *On The Rise of FinTechs – Credit Scoring Using Digital Footprints*. 1 Apr. 2018, https://doi.org/10.3386/w24551.
3. Barton, Genie, et al. "Algorithmic Bias Detection and Mitigation: Best Practices and Policies to Reduce Consumer Harms." *Brookings*, 22 May 2019, www.brookings.edu/articles/algorithmic-bias-detection-and-mitigation-best-practices-and-policies-to-reduce-consumer-harms.