Finding Privacy and Discrimination Bugs in ML Systems

Matt Fredrikson
Computer Science Department &
Institute for Software Research
ML systems are opaque

User data → Online Advertising System → Decisions
Opacity and privacy

2014: Canada’s Privacy Commissioner reports violation of the PIPEDA regulations

- **Key finding:** Google delivered behavior-targeted ads for CPAP devices to a man with sleep apnea

- Individual had issued searches and visited sites seeking information about his condition

Multiple explanations for violation

- First claimed it was due to a bug in the advertising system
- Later revised: remarketing issue, advertiser didn’t follow Google’s policy
Opacity and discrimination

Amazon scraps secret AI recruiting tool that showed bias against women

Jeffrey Dastin
Harmful information use

Both can be seen as *inappropriate/harmful use of protected information*

• Privacy
  • Use of *health information* for *marketing*
  • Exceptions derive from contextual information norms

• Fairness/discrimination
  • Use of *gender* for *employment decisions*
  • Business necessity exceptions

**Opacity enhances the threat**
• Developers don’t know how information is being used
• Auditors can’t easily check compliance, verify claimed violations
Explicit information use

How much **causal influence** do various inputs have on a classifier’s decision about individuals or groups?

Replace feature with random values from the population, and examine distribution over outcomes.
Challenge: Proxy use

Example: targeted advertisements

Visit: webmd.com/apnea

Visit: discountcpap.com

Ad Targeting Model

Ad display

Need to determine that health status is *inferred*, then used
Proxy use: a closer look

Intuition: “proxy” is a random artifact that is associated with protected information

1. Explicit use is also proxy use
   - I.e., random variable is associated with itself
2. “Inferred use” is proxy use
   - I.e., intermediate model state is associated with protected variable
   - State must still be causally influential on model’s output
   - Associations must be two-sided
Checking proxy use [Sen et al. CCS’18, Leino et al. ITC’18, Yeom et al. NeurIPS’18]

Goal: find intermediate states that are
  • associated with protected variable
  • influential on final outcome

Approach:
  • Brute force: search all possible states
  • Stochastic search: sample paths, estimate association and influence
  • Probabilistic model checking: leverage abstractions that save work in common cases

Outcome:
  • If no violation, guarantee or bound on likelihood of proxy use
  • Otherwise, a witness identifying a program component in violation
New kind of bug: *bias amplification*

In training data, 66% of “cooking” images depict women.

In predictions, 84% of “agent” roles in cooking images are labeled “woman”.

Image source: [Zhao et al., EMNLP 2018]
Feature-wise bias amplification [Leino et al., ICLR’19]

Intuition: “kitchen features” are weak predictors (proxies) of female label
- Model over-reliant on weak features
- Prevalent weak features for class → biased predictions
- Common outcome with gradient descent

Intriguing connection to privacy: this can also leak information about training data
Summary

• ML systems are starting to be widely-deployed

• Their opacity introduces risks
  • Privacy and discrimination issues follow from inappropriate information use
  • Developers don’t know how information is being used
  • Auditors can’t easily check compliance, verify claimed violations

• Whitebox analysis lets us peer inside an opaque model
  • Identify specific components and examples that witness violations
  • Guide model repair to mitigate harmful behavior
  • Future: incorporate methods into training pipeline for safe by design systems

Grateful for support from: NSF, AFOSR, Bosch, NVIDIA