Responsible Machine Learning*

Lecture 3: Bias Testing and Remediation

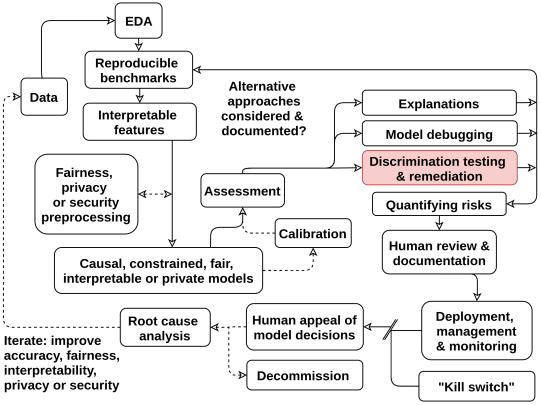
Patrick Hall

The George Washington University

March 31, 2025

Contents

A Responsible Machine Learning Workflow†



†*A Responsible Machine Learning Workflow*

Why Care About Bias in Machine Learning?

- **Responsible practice**: ML can affect millions of people! [7]
- **Legal risk**: Non-compliance fines and litigation costs.
- **Reputational risk**: Upon encountering a perceived unethical ML system, 34% of consumers are likely to, "stop interacting with the company."[a]

[a] See: Capgemeni, *Why addressing ethical questions in AI will benefit organizations* (Withdrawn).

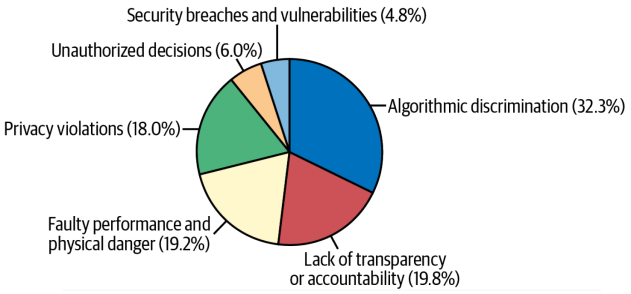


Figure: The frequency of different types of AI incidents based on a qualitative analysis of 169 publicly reported incidents between 1988 and February 1, 2021.

What Is Bias?

- ISO defines bias as, "the degree to which a reference value deviates from the truth."[‡]
- Because *all* data and models are simplified and approximate representations of reality, *all* data, statistical models, and ML models encode different types of *bias*, i.e., **systematic misrepresentations of reality**.
- Oftentimes, bias is *helpful*.
  - Shrunken and robust β_j coefficients in penalized linear models.
- Other types of bias can be unwanted, unhelpful, discriminatory, or illegal.

[‡]See: ISO 3534-1:2006.

What Is Bias in ML?

Recent authoritative work from NIST [8] established three main types of bias in ML:

- **Systemic bias**: Sociological biases (racism, sexism, ableism, etc.) encoded in training data, design choices, or deployment choices.
- **Human bias**: Cognitive mechanisms that alter human perception or cause humans to make wrong decisions.
- **Computational/statistical bias**: Systematically wrong model outcomes that often arise from poor data collection or model misspecification.

The next slide will address human biases, and the remainder of this lecture focuses on systemic biases. A great deal of the class topics, e.g., explainable models, post-hoc explanation, model debugging, are meant to decrease computational/statistical bias.

What are Human Biases in ML?

- **Anchoring**: When a particular reference point, or *anchor*, has an undue effect on people's decisions.
- **Availability heuristic**: We often confuse *easy to remember* with *correct*.
- **Confirmation bias**: People tend to prefer information that aligns with, or confirms, their existing beliefs.
- **Dunning-Kruger effect**: The tendency of people with low ability in a given area or task to overestimate their self-assessed ability.
- **Funding bias**: A bias toward highlighting or promoting results that support or satisfy the funding agency or financial supporter of a project.
- **Groupthink**: When people in a group tend to make nonoptimal decisions based on their desire to conform to the group or fear dissenting with the group.
- **McNamara fallacy**: The belief that decisions should be made based solely on quantitative information, at the expense of qualitative information or data points that aren't easily measured.
- **Techno-chauvinism**: The belief that technology is always the solution.

Legal Notions of Systemic Bias

- **Protected groups**: In the US, many laws and regulations prohibit discrimination based on race, sex (or gender, in some cases), age, religious affiliation, national origin, and disability status, among other categories.
- **Disparate treatment**: A decision that treats an individual less favorably than similarly situated individuals because of a protected characteristic.[§]
- **Disparate impact**: The result of a seemingly neutral policy or practice that disproportionately harms a protected group.
- **Differential validity**: When an employment test is a better indicator of job performance for some groups than for others.
- **Screen out**: When a person with a disability is unable to interact with an employment assessment, and is screened out of a job or promotion by default.[¶]

[§]Concerns about disparate treatment, and more general systemic bias, are why we typically try not to use demographic markers as direct inputs to ML models.

[¶]See: *Algorithms, Artificial Intelligence, and Disability Discrimination in Hiring* (Withdrawn) for recent EEOC statements on screenout and AI.

What is Systemic Bias in ML?

- **Overt discrimination**: Using demographic information in models for the purpose of harming certain demographic groups. (Related to disparate treatment.)
- **Misallocation of resources**: When model *outputs* disproportionately favor certain demograhic groups over others. (Related to disparate impact, often discussed as a lack of "statistical parity.")
- **Unequal quality of service**: When model *performance* favors certain demograhic groups over others. (Related to differential validity.)
- **Individual or local bias**: When a model treats similar individuals differently based on demogrpahic group membership.

Who Tends to Experience Systemic Bias from ML Systems?

- **Age**:
  - **Older people**: Those 40 and above are more likely to experience discrimination in online content; age cutoff can be older in traditional applications–employment, housing, or consumer finance.
  - **Younger people**: Participation in Medicare or the accumulation of wealth may make older people the favored group in other scenarios.
- **Disability**: Those with physical, mental, or emotional disabilities.
- **Immigration status or national origin**: Immigrants with any immigration status, including naturalized citizens.
- **Language**: Those who use languages other than English or who write in non-Latin scripts.
- **Race and ethnicity**: Races and ethnicities other than white people, including those who identify as more than one race.
- **Sex and gender**: Sexes and genders other than cisgender men; in online content, women are often favored but in harmful ways ("male gaze bias").
- **Intersectional groups**: People who are in two or more of the preceding groups may experience bias or harms greater than the sum of the two broader groups.

What Harms Do People Experience?

- **Denigration**: Derogatory or offensive content.
- **Erasure**: Minization or deletion of content challenging dominant social paradigms or past harms suffered by marginalized groups.
- **Exnomination**: Treating notions like whiteness, maleness, or heterosexuality as central human norms.
- **Misrecognition**: Mistaking a person's identity or failing to recognize someone's humanity.
- **Stereotyping**: Assigning characteristics to all members of a group.
- **Underrepresentation**: The lack of fair or adequate representation of demographic groups in model output.
- **Economic harms**: Reduction of the economic opportunity or value for some groups.
- **Physical harms**: Hurting or killing someone.
- **Psychological harms**: Causing mental or emotional distress.

Considerations for Data

General issues:

- **Representativeness**: Does the number of rows of data for each group match real demographic distributions? If no, often results in differential validity/quality of service.
- **Distribution of outcomes**: Are favorable outcomes (y values) distributed equally across demographic groups? If no, often results in misallocation of resources/disparate impact.
- **Proxies**: Do seemingly neutral features actually encode systemic bias? If yes, often results in misallocation of resources/disparate impact.

Specific examples:

- Incomplete or inaccurate data, e.g., under-representation of minorities. See Gender Shades [2].
- Accurate but differing distributions of outcomes, correlation, or local dependencies between demographic groups and past outcomes, e.g., traditional FICO credit scores.
- Name or zip code acting as proxies for race.
- Explicit encoding of systemtic biases into training data, e.g., criminal records.

Common Test Metrics for Systemic Bias in ML

- **Misallocation/disparate impact, e.g.**:
  - Adverse impact ratio: $\frac{\%\ \text{accepted}_p}{\%\ \text{accepted}_c}$
  - Marginal effect: $\%\ \text{accepted}_p - \%\ \text{accepted}_c$
  - Standardized mean difference: $\frac{\bar{\hat{y}}_p - \bar{\hat{y}}_c}{\sigma_{\hat{y}}}$
- **Different quality of service/validity, e.g.**: $\frac{\text{quality}_p}{\text{quality}_c}$

where, $p \equiv$ protected group and $c \equiv$ control group (often white males).

There are many other, sometimes conflicting, mathematical definitions of discrimination. See 21 Definitions of Fairness and Their Politics.

Example AIR Calculation

protected acceptance rate: (tp + fp) / (tp + fp + tn + fn)

control acceptance rate: (tp + fp) / (tp + fp + tn + fn)

Overall	Actual: 1	Actual: 0
Predicted: 1	4610	2933
Predicted: 0	6747	17157

control

Men	Actual: 1	Actual: 0
Predicted: 1	2954	1021
Predicted: 0	2809	8316

Women	Actual: 1	Actual: 0
Predicted: 1	1656	1912
Predicted: 0	3938	8841

protected

Example AIR Calculation

protected acceptance rate: (tp + fp) / (tp + fp + tn + fn)

control acceptance rate: (tp + fp) / (tp + fp + tn + fn)

control

Men	Actual: 1	Actual: 0
Predicted: 1	2954	1021
Predicted: 0	2809	8316

Women	Actual: 1	Actual: 0
Predicted: 1	1656	1912
Predicted: 0	3938	8841

protected

(1656 + 1912)/(1656 + 1912 + 3938 + 8841) = 0.218...

(2954 + 1021)/(2954 + 1021 + 2809 + 8316) = 0.263 ...

0.829 > 0.8

Common Test Metrics for Systemic Bias in ML

Test type	Discrete outcome/Classification tests	Continuous outcome/Regression tests
Statistical significance	Logistic regression coefficient	Linear regression coefficient
Statistical significance	χ^2 test	*t*-test
Statistical significance	Fisher's exact test	
Statistical significance	Binomial-*z*	
Practical significance	Comparison of group means	Comparison of group means
Practical significance	Percentage point difference between group means/marginal effect	Percentage point difference between group means
Practical significance	Adverse impact ratio (AIR) (acceptable: `0.8 – 1.25`)	Standardized mean difference (SMD, Cohen's *d*) (small difference: `0.2`, medium difference: `0.5`, large difference: `0.8`)
Practical significance	Odds ratios	
Practical significance	Shortfall to parity	
Differential validity	Accuracy or AUC ratios (acceptable: `0.8 – 1.25`)	R^2 ratio (acceptable: `0.8 – 1.25`)
Differential validity	TPR, TNR, FPR, FNR ratios (acceptable: `0.8 – 1.25`)	MSE, RMSE ratios (acceptable: `0.8 – 1.25`)
Differential validity	Equality of odds ([control TPR ≈ protected TPR | `y = 1`] and [control FPR ≈ protected FPR | `y = 0`])	
Differential validity	Equality of opportunity ([control TPR ≈ protected TPR | y = 1])	

Additional Considerations for Bias Testing

- Understand past similar incidents, via the AI Incident Database.
- **Adversarial models**: Use another ML model ($h_{adversary}$) to predict demographic group membership from predictions ($\hat{y}$) or model inputs (X_j).
- **Local bias**: Search around probability thresholds, apply adversarial models, apply counterfactual explanations.
- Understand drivers of bias with post-hoc explanation:
  - To be conducted after bias is confirmed by standard tests.
  - Be aware: lack of demographic features, fairwashing [1], and scaffolding [9].
- **Multinomial classication**: χ^2 or equality of opportunity tests; dimension reduction on $\hat{Y}$ matrix, apply continous outcome tests.
- **Unsupervised learning**: Adversarial models on cluster labels or extracted features, discrete outcome tests for cluster labels, and continuous outcome tests for extracted features.

How to Mitigate Systemic Bias in ML?

Fix the data:

- Collect demographically representative training data.
- Label and annotate data carefully.
- Select features judiciously.
- Sample and reweigh training data to minimize bias (consider group size and outcomes).[4]

How to Mitigate Systemic Bias in ML?

Fix the model:

- Consider bias measures when selecting hyperparameters and cutoff thresholds.
- Train less biased models directly:
  - Learning fair representations (LFR) and adversarial de-biasing.[10], [11]
  - Use dual objective functions that consider both accuracy and fairness metrics.
- Edit model mechanisms to ensure less biased predictions, e.g., with GA2M/EBM models.

How to Mitigate Systemic Bias in ML?

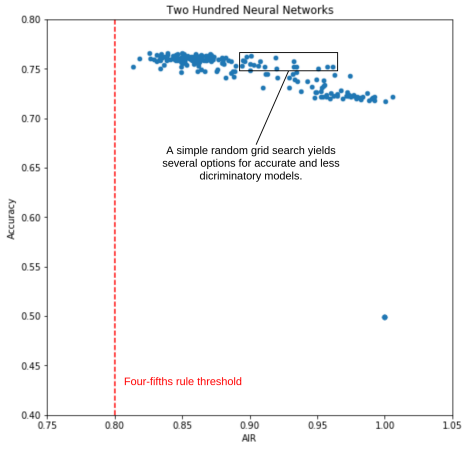


Figure: An example of considering bias measures in model selection.

How to Mitigate Systemic Bias in ML?

Fix the predictions:

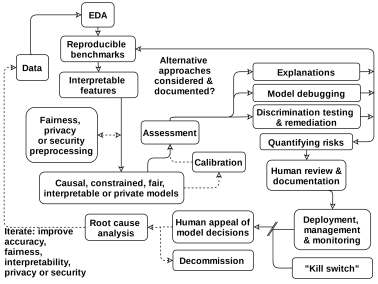
- Balance model predictions, e.g., reject-option classification.[5][‖]
- Correct or override predictions with model assertions or appeal mechansims.[3], [6]

[‖] Re-balancing predictons based on protected class information may give rise to risks relating to disparate treatment, reverse discrimination, or unapproved affirmative action.

How to Mitigate Systemic Bias in ML?

- Fix organizational processes/apply governance: Lecture 6.
- Apply the scientific method and experimental design to ML systems.
- Increase demographic and professional diversity in ML teams.

As part of a responsible ML workflow.



Acknowledgements

This presentation borrows heavily from the expertise of Nicholas Schmidt of BLDS, LLC, a leading fair lending compliance firm.

Thanks to Lisa Song for her continued assistance in developing these course materials.

References

Ulrich Aïvodji et al. "Fairwashing: the Risk of Rationalization." In: *arXiv preprint arXiv:1901.09749* (2019). URL: https://arxiv.org/pdf/1901.09749.pdf.

Joy Buolamwini and Timnit Gebru. "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification." In: *Conference on Fairness, Accountability and Transparency*. URL: http://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf. 2018, pp. 77–91.

Patrick Hall, Navdeep Gill, and Nicholas Schmidt. "Guidelines for the Responsible Use of Explainable Machine Learning." In: *arXiv preprint arXiv:1906.03533* (2019). URL: https://arxiv.org/pdf/1906.03533.pdf.

Faisal Kamiran and Toon Calders. "Data Preprocessing Techniques for Classification Without Discrimination." In: *Knowledge and Information Systems* 33.1 (2012). URL: https://bit.ly/2lH95lQ, pp. 1–33.

Faisal Kamiran, Asim Karim, and Xiangliang Zhang. "Decision Theory for Discrimination-aware Classification." In: *2012 IEEE 12th International Conference on Data Mining*. URL: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.722.3030&rep=rep1&type=pdf. IEEE. 2012, pp. 924–929.

References

Daniel Kang et al. *Debugging Machine Learning Models via Model Assertions*. URL: https://www-cs.stanford.edu/~matei/papers/2018/mlsys_model_assertions.pdf. 2019.

Ziad Obermeyer et al. "Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations." In: *Science* 366.6464 (2019). URL: https://bit.ly/36XK6yk, pp. 447–453.

Reva Schwartz et al. "Towards a Standard for Identifying and Managing Bias in Artificial Intelligence." In: *NIST Special Publication* 1270 (2022). URL: https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1270.pdf, pp. 1–77.

Dylan Slack et al. "How Can We Fool LIME and SHAP? Adversarial Attacks on Post-hoc Explanation Methods." In: *arXiv preprint arXiv:1911.02508* (2019). URL: https://arxiv.org/pdf/1911.02508.pdf.

Rich Zemel et al. "Learning Fair Representations." In: *International Conference on Machine Learning*. URL: http://proceedings.mlr.press/v28/zemel13.pdf. 2013, pp. 325–333.

Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. "Mitigating Unwanted Biases with Adversarial Learning." In: *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. URL: https://arxiv.org/pdf/1801.07593.pdf. ACM. 2018, pp. 335–340.