An Effective Theory of Bias Amplification

Arjun Subramonian (they/them)
Meta FAIR
USC ISI NL Seminar

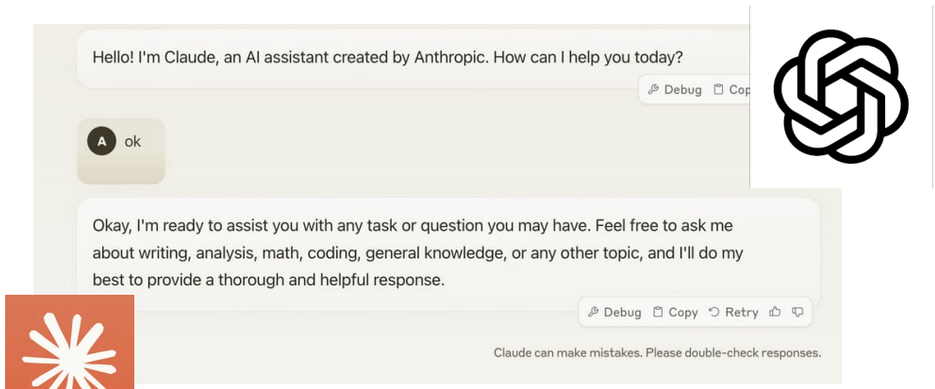
Overview

- **Motivation**
- An Effective Theory of Bias Amplification [ICLR 2025]
- Conclusion and Future Directions

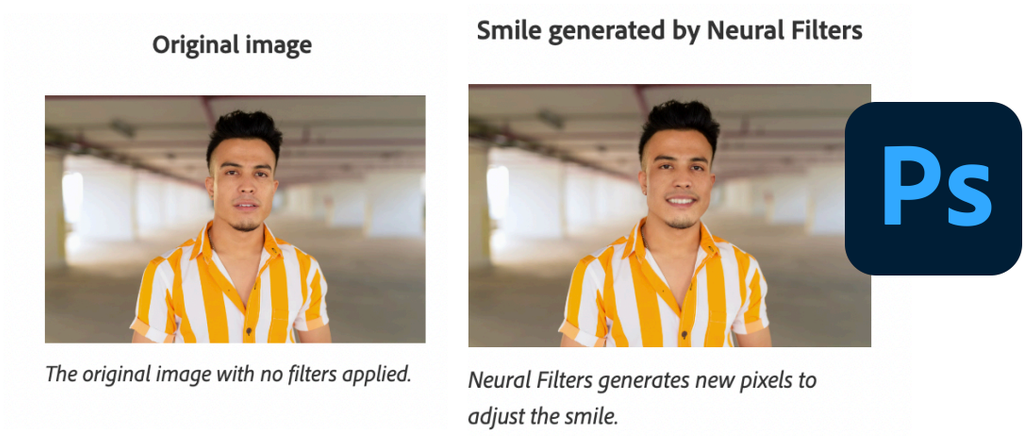
Motivation

Machine Learning

powerful user-facing products

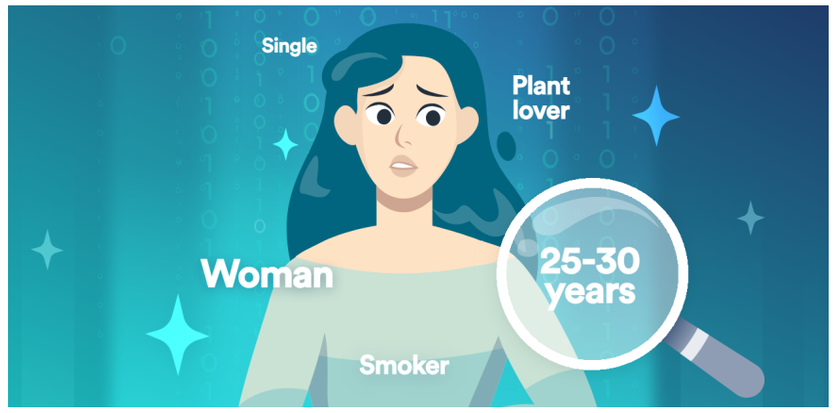


https://www.testingcatalog.com/claude-ai-teases-claudia-interface-with-a-sidebar-akin-to-chatgpt/

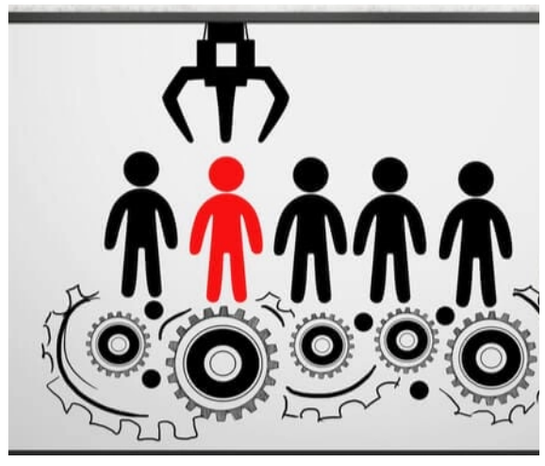


https://helpx.adobe.com/photoshop/using/neural-filters.html

predictions about individuals at scale



https://surfshark.com/blog/what-is-targeted-advertising



https://www.techfunnel.com/hr-tech/will-automated-hiring-systems-transform-recruitment/

Deep Learning, Deep Problems

- Unfairness issues cause real-world harm

- Why are ML models unfair?

  - Marginalized social groups underrepresented or misrepresented in data

  - Model design choices

Gender Classifier	Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap
Microsoft	94.0%	79.2%	100%	98.3%	20.8%
FACE++	99.3%	65.5%	99.2%	94.0%	33.8%
IBM	88.0%	65.3%	99.7%	92.9%	34.4%

Buolamwini, J., Gebru, T. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. FAccT 2018.

https://hai.stanford.edu/ai-index/2025-ai-index-report

3. Skepticism about the ethical conduct of AI companies is growing, while trust in the fairness of AI is declining. Globally, confidence that AI companies protect personal data fell from 50% in 2023 to 47% in 2024. Likewise, fewer people today believe that AI systems are unbiased and free from discrimination compared to last year.

As ML models are increasingly deployed to make predictions about humans and their relationships at scale…

…it is critical that models are *fair* and do not amplify social inequalities.

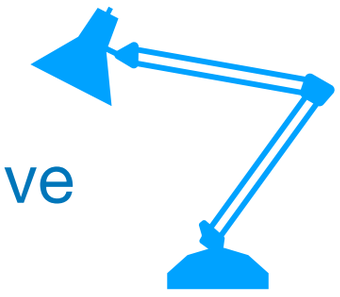
Ovalle, A., **Subramonian, A.**, Gautam, V., Gee, G., Chang, K.W. Factoring the Matrix of Domination: A Critical Review and Reimagination of Intersectionality in AI Fairness. AIES 2023.

Overview

- Motivation
- **An Effective Theory of Bias Amplification [ICLR 2025]**
  - Co-authors: Samuel J. Bell, Levent Sagun, Elvis Dohmatob
- Conclusion and Future Directions

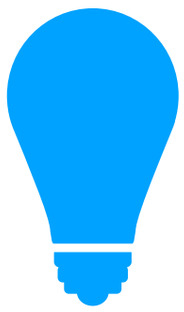


Proliferation of ML



Precise unifying theory of unfairness remains elusive

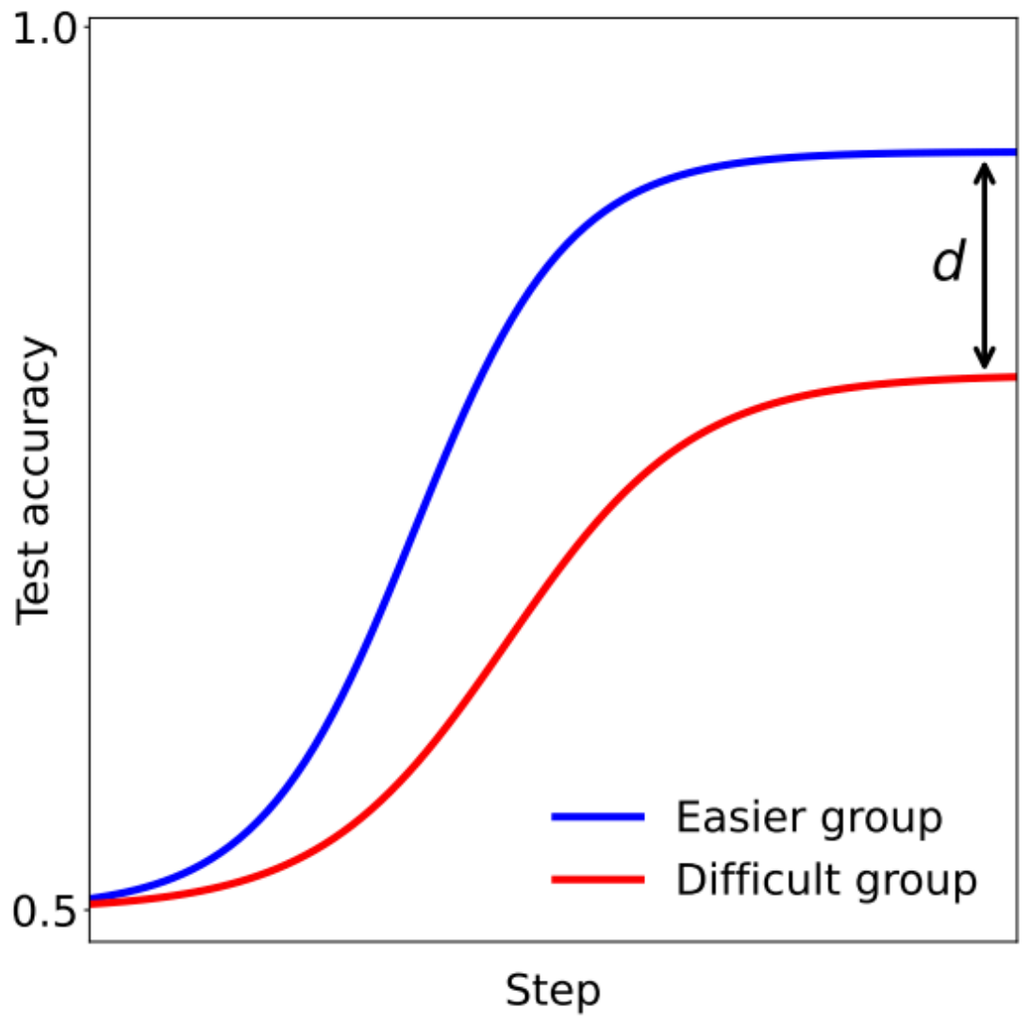
Greater interpretability



Design better unfairness evaluation and mitigation methods by identifying common sources of unfairness

Interpolate between sparse empirical findings

Precise analytical theory of how *model design choices* and *data distributional properties* interact to amplify bias.



Theoretical Setup

- Data: regression problem on dataset from multivariate Gaussian mixture with two groups $s = 1$ and $s = 2$

  - Groups could represent different demographic groups or protected categories

(Group ID) $\mathrm{Law}(s) = \mathrm{Bernoulli}(p),$

(Features) $\mathrm{Law}(x \mid s) = \mathcal{N}(0, \Sigma_s),$ more general covariance structure

(Ground-truth weights) $\mathrm{Law}(w_1^*) = \mathcal{N}(0, \Theta/d), \quad \mathrm{Law}(w_2^* - w_1^*) = \mathcal{N}(0, \Delta/d),$

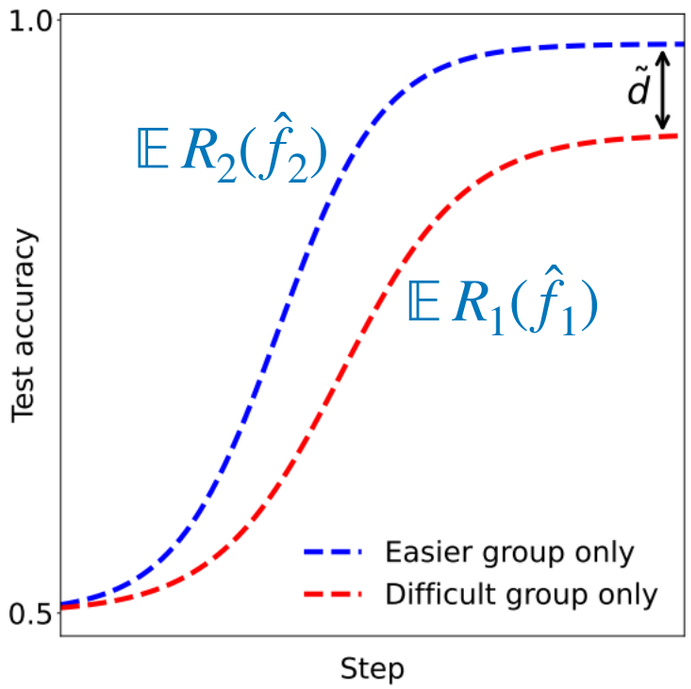
(Labels) $\mathrm{Law}(y \mid s, x) = \mathcal{N}(f_s^\star(x), \sigma_s^2),$ **with** $f_s^\star(x) := x^\top w_s^* .$

A Tale of Two Learning Paradigms

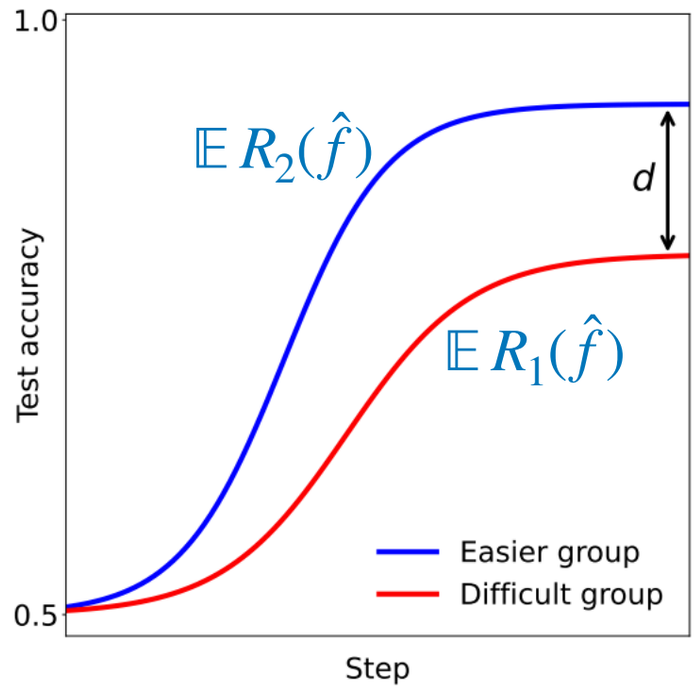
$$R_s(f) = \mathbb{E}\left[(f(x) - f_s^\star(x))^2 \mid s\right]$$

$$EDD = |\mathbb{E}\, R_2(\hat{f}_2) - \mathbb{E}\, R_1(\hat{f}_1)| \quad ODD = |\mathbb{E}\, R_2(\hat{f}) - \mathbb{E}\, R_1(\hat{f})|$$

$$ADD = \frac{ODD}{EDD} \square 1$$



(b) Train each group separately



(c) Train both groups together

Bell, S.J., Sagun, L. Simplicity Bias Leads to Amplified Performance Disparities. FAccT 2023.

Theoretical Setup

- Model: *linear* two-layer neural network in random features regime:

first-layer weights random and fixed

$$\widehat{w} = \boxed{S}\widehat{\eta} \in \mathbb{R}^d,$$

- $S \in \mathbb{R}^{d \times m}$ is random projection with entries that are IID sampled from $\mathcal{N}(0, 1/d)$

- Captures effect of model size on bias amplification

- ℓ_2-regularization penalty λ

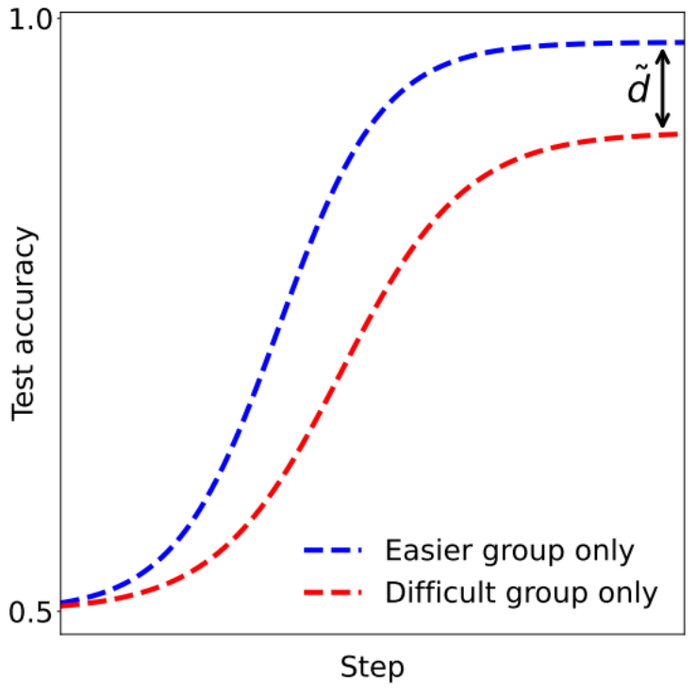
Theoretical Setup

- High-dimensional limit: proportionate scaling captures effect of different feature and parametrization regimes:

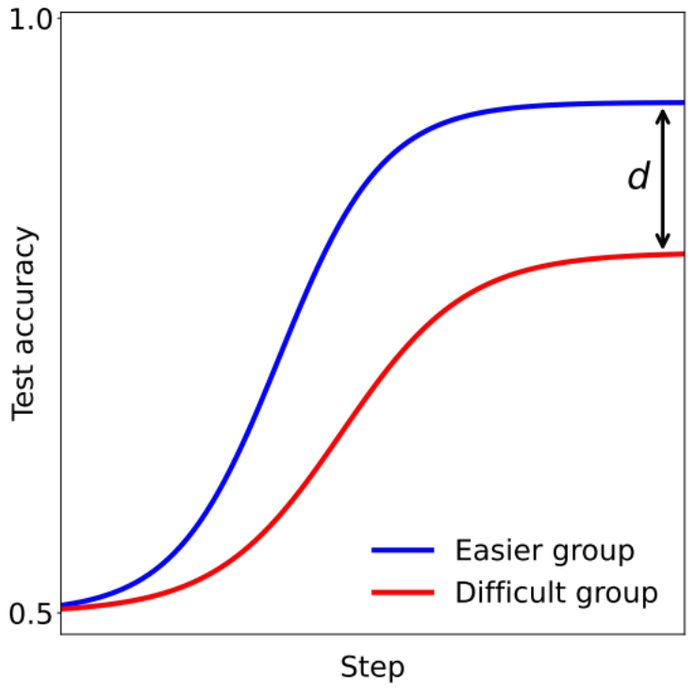
$$n, n_1, n_2, d, m \to \infty, \quad d/n \to \phi, m/n \to \psi, m/d \to \gamma$$

- ψ captures rate of parameters to samples
  - $\psi > 1$: overparameterized regime
- ϕ captures rate of features to samples

$$EDD = |\mathbb{E}\,R_2(\hat{f}_2) - \mathbb{E}\,R_1(\hat{f}_1)| \quad ODD = |\mathbb{E}\,R_2(\hat{f}) - \mathbb{E}\,R_1(\hat{f})|$$



(b) Train each group separately



(c) Train both groups together

Bell, S.J. and Sagun, L. Simplicity Bias Leads to Amplified Performance Disparities. FAccT 2023.

Theory of Bias Amplification

group proportions

$$R_s(\hat{f}) \simeq B_s(\hat{f}) + V_s(\hat{f})$$

parameter and feature rates

$$V_s(\hat{f}) = V(s, p_1, p_2, \sigma_1^2, \sigma_2^2, \phi, \psi, v_{fo}, v_{so}),$$

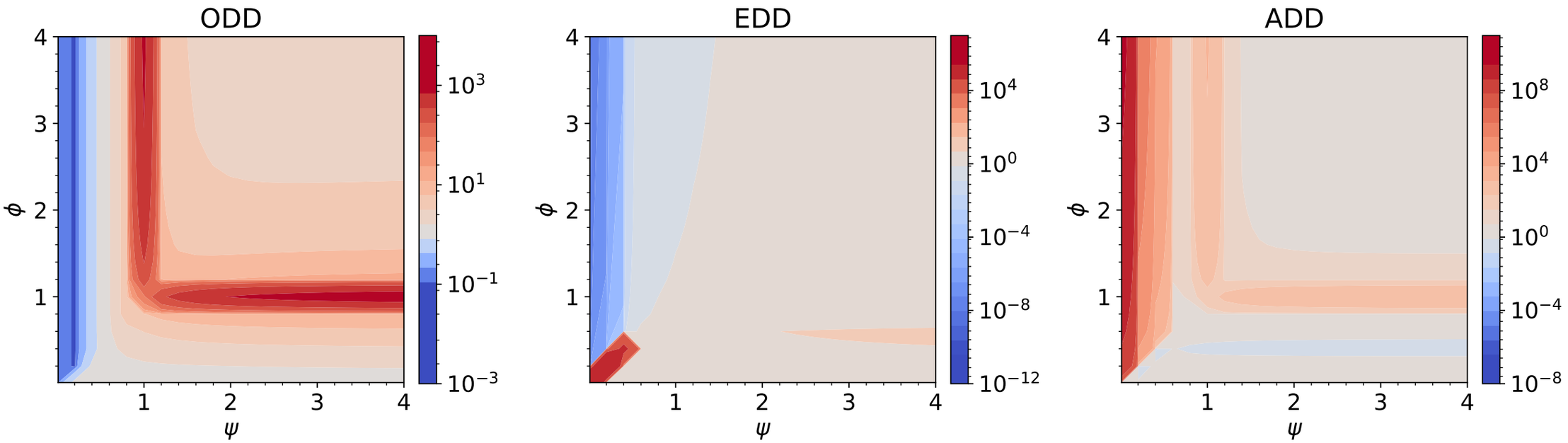
$$B_s(\hat{f}) = B(s, p_1, p_2, \sigma_1^2, \sigma_2^2, \phi, \psi, v_{fo}, v_{so})$$

label noises

- $v_{fo} = (e_1, e_2, \tau)$, $v_{so} = (u_1, u_2, \rho)$ satisfy fixed-point equations:
  - Involving λ and 1st and 2nd-order degrees of freedom of covariance matrices $\Sigma_1, \Sigma_2, \Theta, \Delta$
  - Given by **operator-valued free probability theory**
- Compute $R_s(\hat{f}_s)$ by taking limit $p_s \to 1$

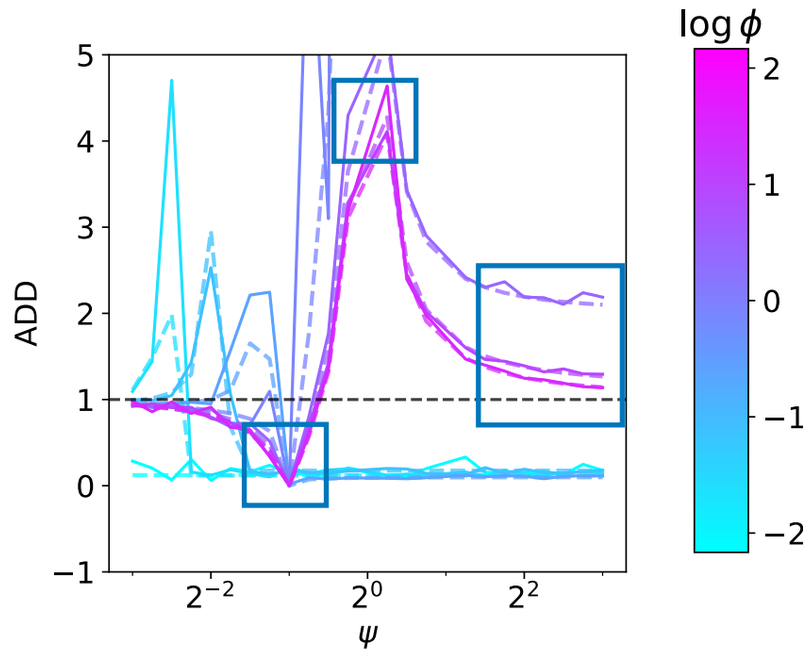
Theory of Bias Amplification

- Phase diagrams reveal interpolation thresholds at $\phi = 1, \psi = 1, \phi = \psi$



Empirical Validation

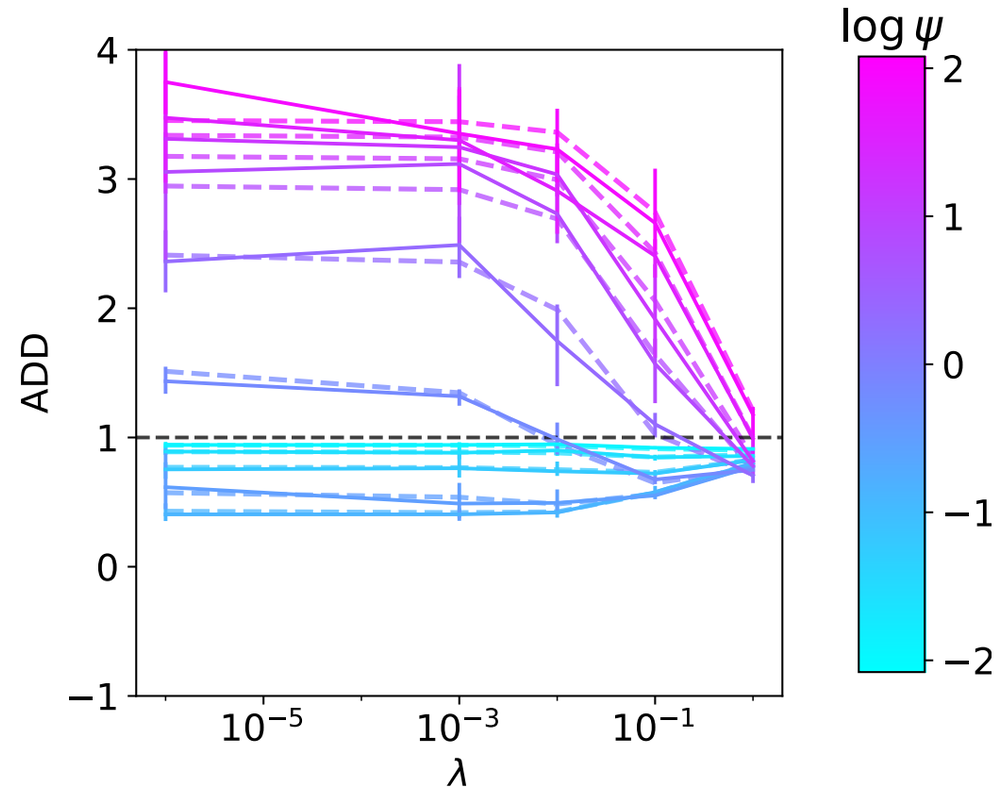
- Balanced data ($p_1 = p_2 = 1/2$)
- No spurious correlations ($\Sigma_1 = 0.5 I_d, \Sigma_2 = I_d$)



Models can amplify bias even with balanced groups and without spurious correlations.

Empirical Validation

- Use calibration $\lambda \sim 1/t$



May be an optimal regularization penalty or training time to deamplify bias.

Ali, A., Kolter, J.Z., Tibshirani, R.J. A Continuous-Time View of Early Stopping for Least Squares. AISTATS 2019.

Technical Difficulty

- Non-trivial generalization and not special case of existing results in high-dimensional ML literature

  - Requires machinery of free probability theory (FPT)

  - Released tool called *auto-fpt* to automate FPT calculations

$$V_s(\hat{f}) = V_s^{(1)}(\hat{f}) + V_s^{(2)}(\hat{f}),$$

$$\textbf{with } V_s^{(j)}(\hat{f}) = \sigma_j^2 \phi \mathbb{E}\,\mathrm{tr}\, X_j^\top X_j S(S^\top X^\top X S + \lambda I_m)^{-1} S^\top \Sigma_s S(S^\top X^\top X S + \lambda I_m) S^\top / d,$$

$$B_s(\hat{f}) = \mathbb{E}\| S(S^\top X^\top X S + \lambda I_m) S^\top (X_1^\top X_1 w_1^* + X_2^\top X_2 w_2^*) - w_s^* \|_{\Sigma_s}^2.$$

Subramonian, A., Dohmatob, E. auto-fpt: Automating Free Probability Theory Calculations for Machine Learning Theory.

Actionable Insights

- Form intuition about when models may amplify bias
  - Disparities in variance of features and labels
  - Overparamaterization and feature-to-sample rate ≈ 1
- Inform hyperparameter selection

Extending our Theory

- More than two groups
- Features and label noise sampled from other distributions
- Different proportionate scaling limits, e.g., d^2/n has finite limit instead of d/n
- Missing features and unknown group information
- Wide fully-trained networks in NTK and lazy regimes

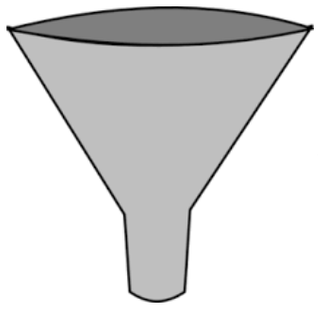
Overview

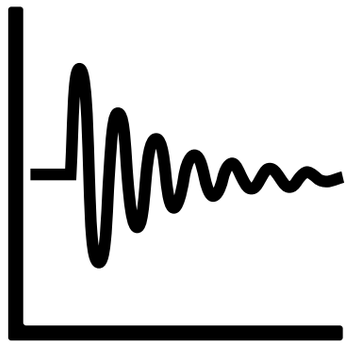
- Motivation
- An Effective Theory of Bias Amplification [ICLR 2025]
- **Conclusion and Future Directions**

ML Development Lifecycle

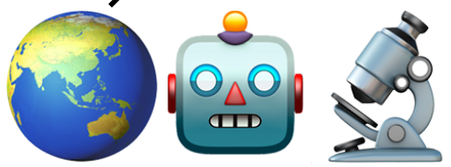
[ICLR 2025]

Data curation:
How do practitioner choices about data sources and curation mechanisms affect representation of different social groups?

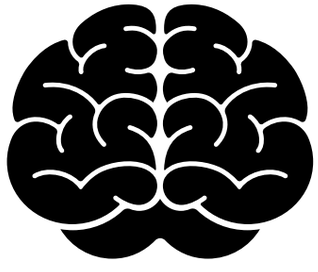




Training dynamics:
How do data order and hyperparameter choices impact fairness?



Finetuning and alignment:
Whose context and preferences are marginalized? How can we reconcile different preferences across social groups?



Evaluations:
How do automatic evaluation methods fail to capture the severity and nuances of unfairness?

Social Dimensions of Fairness

- Work focuses on *technical* dimensions of fairness
- Fair ML requires participation of marginalized communities throughout model lifecycle
- Fair ML should combat social inequality and advance justice

QueerInAI, O., Ovalle, A., **Subramonian, A.**, … Queer In AI: A Case Study in Community-Led Participatory AI. FAccT 2023.

QueerInAI, O., Dennler, N., Ovalle, A., Singh, A., Soldaini, L., **Subramonian, A.**, … Bound by the Bounty: Collaboratively Shaping Evaluation Processes for Queer AI Harms. AIES 2023.

Ovalle, A., **Subramonian, A.**, Gautam, V., Gee, G., Chang, K.W. Factoring the Matrix of Domination: A Critical Review and Reimagination of Intersectionality in AI Fairness. AIES 2023.

Thank you! Questions?

- Motivation
- An Effective Theory of Bias Amplification [ICLR 2025]
- Conclusion and Future Directions