University of Idaho

Department of Computer Science

CS 487/587 Adversarial Machine Learning

Dr. Alex Vakanski

Lecture 14

Bias and Fairness in Machine Learning

Lecture Outline

- Bias and fairness definitions
- Introduction to ML bias
  - Data bias
  - Algorithmic bias
- Bias amplification
- Fairness metrics
- Mitigation strategies

Introduction

Introduction

- Although ML models are increasingly used to influence decisions that impact society, from loan approvals to hiring processes, the models are not inherently neutral
- ML models reflect the biases present in their training data, potentially expanding and amplifying existing societal inequalities
- When deployed at scale, biased ML models can systematically disadvantage marginalized groups

Bias and Fairness Definitions

Introduction

- ***Bias in ML and AI***
  - **Bias** is a systematic error in the ML model that produces skewed results that favor or disadvantage certain demographic groups
  - Bias often originates from imbalanced or non-representative training data or biased assumptions in model development
- ***Fairness in ML and AI***
  - **Fairness** is the absence of prejudice or favoritism toward individuals or groups based on their protected characteristics
  - Fairness extends beyond equality of treatment, providing different levels of support to achieve just outcomes across diverse populations
  - Multiple mathematical definitions of fairness exist, each representing different perspectives on what constitutes "fair" outcomes

Introduction to ML Bias

Introduction to ML Bias

- Understanding bias and fairness in ML requires recognizing that these concepts are nuanced and context-dependent
  - What constitutes "fair" in one application may not be appropriate in another application: this requires ongoing evaluation and adjustment
- Sources of ML bias
  - Data bias
    - Data bias is one of the main reasons for fairness issues in ML systems
    - Recognizing the sources of data bias is important for building trustworthy AI, as ML models can only be as fair as the data used to train them
  - Algorithmic bias
    - Even with balanced training data, the choice of ML algorithm can introduce or intensify unfairness in model outcomes
    - Algorithmic bias can emerge at various stages of the ML pipeline, from data collection to deployment
      - Assumptions in the ML model design can amplify certain characteristics over others, leading to biased outputs

Types of Data Bias in ML

Data Bias

- Historical bias
  - Includes pre-existing historical societal inequalities reflected in data, even when the data is collected accurately
    - E.g., criminal justice data overrepresenting certain demographics
    - E.g., medical research data historically focused on male subjects
- Data aggregation bias
  - Combining data in ways that ignore disproportions between subgroups
    - E.g., averaging outcomes across different demographic groups
- Representation bias
  - Samples that fail to reflect the diversity of the population
    - E.g., facial recognition ML model trained primarily on light-skinned faces
    - E.g., voice recognition ML model struggling with certain accents
- Omitted variable bias
  - Arises when important features are excluded from the ML model
    - E.g., a credit scoring algorithm that considers only traditional credit history may disadvantage communities with limited access to banking services
- Measurement bias
  - Errors or inconsistencies in data collection or labeling
    - E.g., health sensors less accurate for darker skin tones
    - E.g., inconsistent labeling standards across different groups

Types of Algorithmic Bias in ML

Algorithmic Bias

- Algorithm bias
  - Originates from inherent limitations or design flaws in ML algorithms
  - Some ML model architectures may be more likely to amplify bias present in the training data, or may prioritize optimization methods or metrics that don't align with fairness considerations
- Regularization Bias
  - Regularization techniques can suppress minority group information and treat that information as noise
    - E.g., L_1/L_2 regularization might eliminate rare but important features
- Transfer Bias
  - Occurs when models trained in one context (domain, population) are reused elsewhere without proper adaptation
    - E.g., a healthcare model trained on U.S. patient data performs poorly on data from another country
- Evaluation bias
  - Occurs when the model is evaluated on a non-representative dataset
    - E.g., facial recognition evaluated on one demographic group lead to decreased performance when the system is deployed broadly across diverse populations

Sources of Bias in ML

Introduction to ML Bias

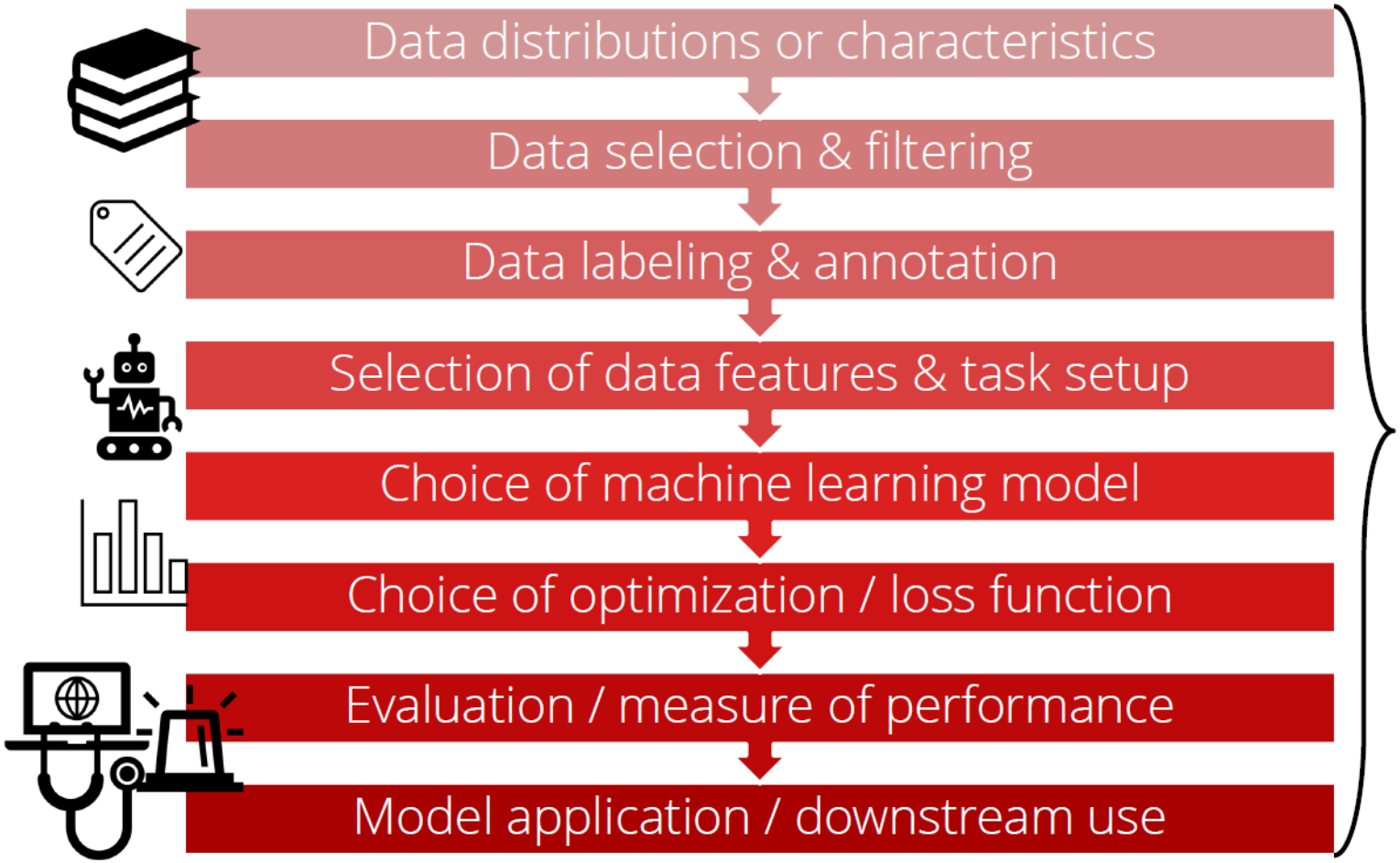


Figure from: Maarten Sap, "Safety, Ethics, and Biases in AI and NLP systems

Algorithmic Bias in ML Systems

Algorithmic Bias

- Safiya Umoja Noble (2018) Algorithms of Oppression
  - Examines how search engines often prioritize content that reinforces racial stereotypes, showing how algorithmic structures reflect societal biases
- Bender (2021) On the Dangers of Stochastic Parrots
  - Discusses risks in LLMs emphasizing that over-reliance on biased data leads to "parroting" harmful or inaccurate stereotypes
  - Highlights the responsibility of tech companies in managing and reducing bias in generative AI systems
- Humble (2024) War, Artificial Intelligence, and the Future of Conflict
  - Explores ethical concerns around AI deployment in military and conflict scenarios
  - Raises awareness of how biased AI in conflict situations could disproportionately target specific groups

Bias Amplification

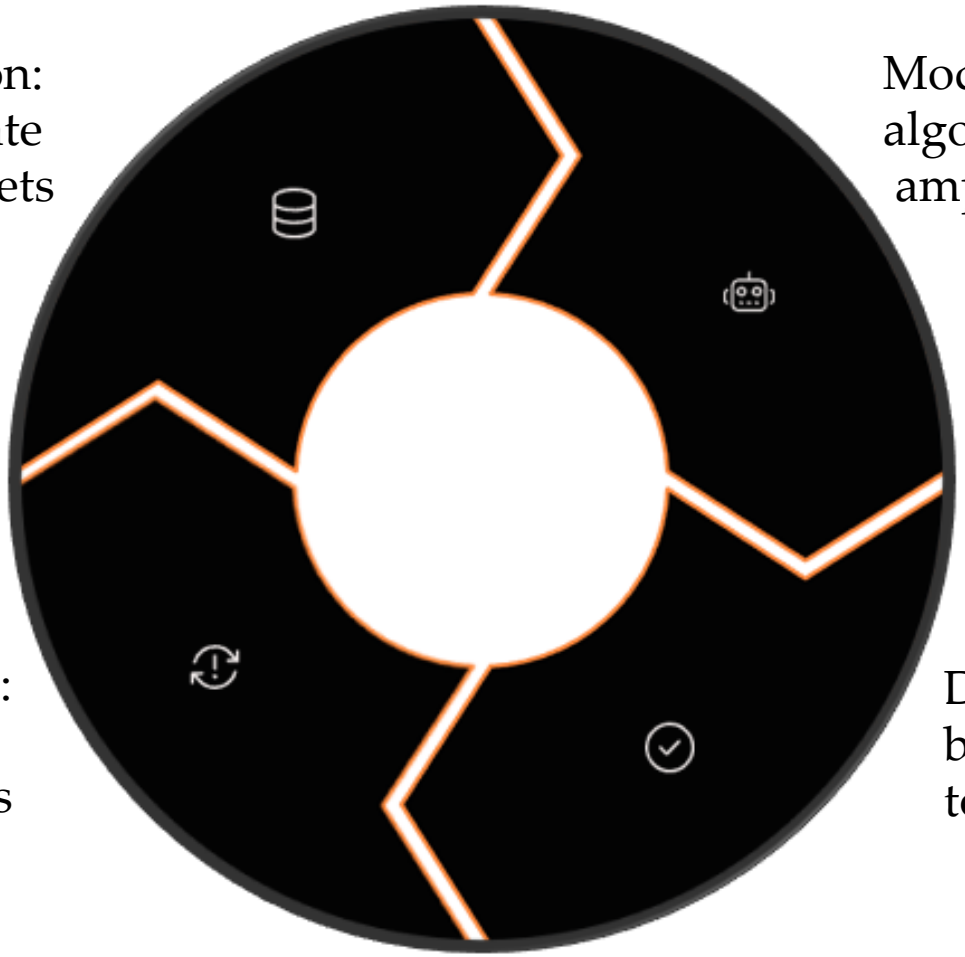
Bias Amplification

- One of the most concerning aspects of bias in ML is the potential for ***bias amplification*** effects
  - When ML models trained on biased data are deployed in real-world scenarios, they can create feedback loops that reinforce existing inequalities
- Example
  - ML algorithms direct more officers to neighborhoods with historically higher arrest rates
  - The increased police presence leads to more arrests
  - This generates more data points that reinforce the model's prediction that these areas have higher crime rates, regardless of the actual crime rate
- Similar feedback effects can occur in hiring, lending, healthcare, and other domains, creating a cycle that becomes difficult to break

Bias Amplification

Bias Amplification

Biased data collection: historical inequities create imbalanced datasets



Model training: algorithms learn and amplify patterns of bias

Feedback loop: new data reinforces existing patterns

Deployment decisions: biased predictions lead to unfair outcomes

Why ML Bias Matters

Introduction to ML Bias

- ML bias can have significant impact on society
  - Impact on discrimination: biased AI can lead to discriminatory decisions, affecting hiring, lending, policing, and healthcare
    - E.g., biased facial recognition algorithms may lead to wrongful arrests or surveillance of marginalized communities
    - E.g., predictive algorithms use historical crime data to predict future crimes: if past data overrepresents certain neighborhoods, ML models will produce biased outcomes
  - Impact on opportunities: AI bias can result in lost opportunities for demographic groups
    - E.g., biased hiring algorithms exclude qualified candidates based on gender, race, or socioeconomic background
- Importance of awareness of ML bias
  - For ML developers: being aware helps developers create fair ML systems, conduct bias testing, and promote the development of inclusive datasets for model training
  - For society: a better-informed public can hold AI developers and companies accountable

Case Study: COMPAS

- COMPAS is an ML system that has been used in New York, Wisconsin, California, Florida, and other states
  1. Person commits a crime, is arrested
  2. COMPAS software predicts the chance that the person will commit another crime in the future (recidivism)
  3. Recidivism scores impact criminal sentences: if a person is likely to commit another crime, they may get a longer sentence
- 2016 ProPublica article analyzed COMPAS scores for over 7,000 people arrested in Florida

Case Study: COMPAS

Case Study: COMPAS

- Error metrics

	Prediction: Low Risk	Prediction: High Risk
Outcome: No Recidivism	True Negative (TN)	False Positive (FP)
Outcome: Recidivated	False Negative (FN)	True Positive (TP)

- Error Rate $= \frac{FP+FN}{TN+FP+FN+TP}$
  - How often is the prediction wrong (opposite to accuracy)
- False Positive Rate $= \frac{FP}{FP+TN}$
  - **False alarm**: How often were non-offenders predicted to reoffend?
  - A true negative (non-offender) is incorrectly classified as positive (offender)
- False Negative Rate $= \frac{FN}{FN+TP}$
  - **Miss**: How often were offenders predicted not to reoffend?
  - A true positive (offender) is incorrectly classified as negative (non-offender)

Case Study: COMPAS

Case Study: COMPAS

- Metrics by the COMPAS model

	Prediction: Low Risk	Prediction: High Risk
Outcome: No Recidivism	2681 (TN)	1282 (FP)
Outcome: Recidivated	1216 (FN)	2035 (TP)

- Error Rate = $\frac{FP+FN}{TN+FP+FN+TP} \approx 34.6\%$
- False Positive Rate = $\frac{FP}{FP+TN} \approx 32.4\%$
- False Negative Rate = $\frac{FN}{FN+TP} \approx 37.4\%$

Case Study: COMPAS

Case Study: COMPAS

- Metrics by the COMPAS model for black defendants and white defendants

Black Defendants	Prediction: Low Risk	Prediction: High Risk
Outcome: No Recidivism	990 (TN)	805 (FP)
Outcome: Recidivated	532 (FN)	1369 (TP)

Error Rate ≈ 36.2%
False Positive Rate ≈ 44.9%
False Negative Rate ≈ 28.0%

White Defendants	Prediction: Low Risk	Prediction: High Risk
Outcome: No Recidivism	1139 (TN)	349 (FP)
Outcome: Recidivated	461 (FN)	505 (TP)

Error Rate ≈ 33.0%
False Positive Rate ≈ 23.5%
False Negative Rate ≈ 47.7%

- Similar error rates between white and black defendants
- Black defendants have almost 2 times higher False Positive Rate (false alarm)
- White defendants have almost 2 times higher False Negative Rate (miss)

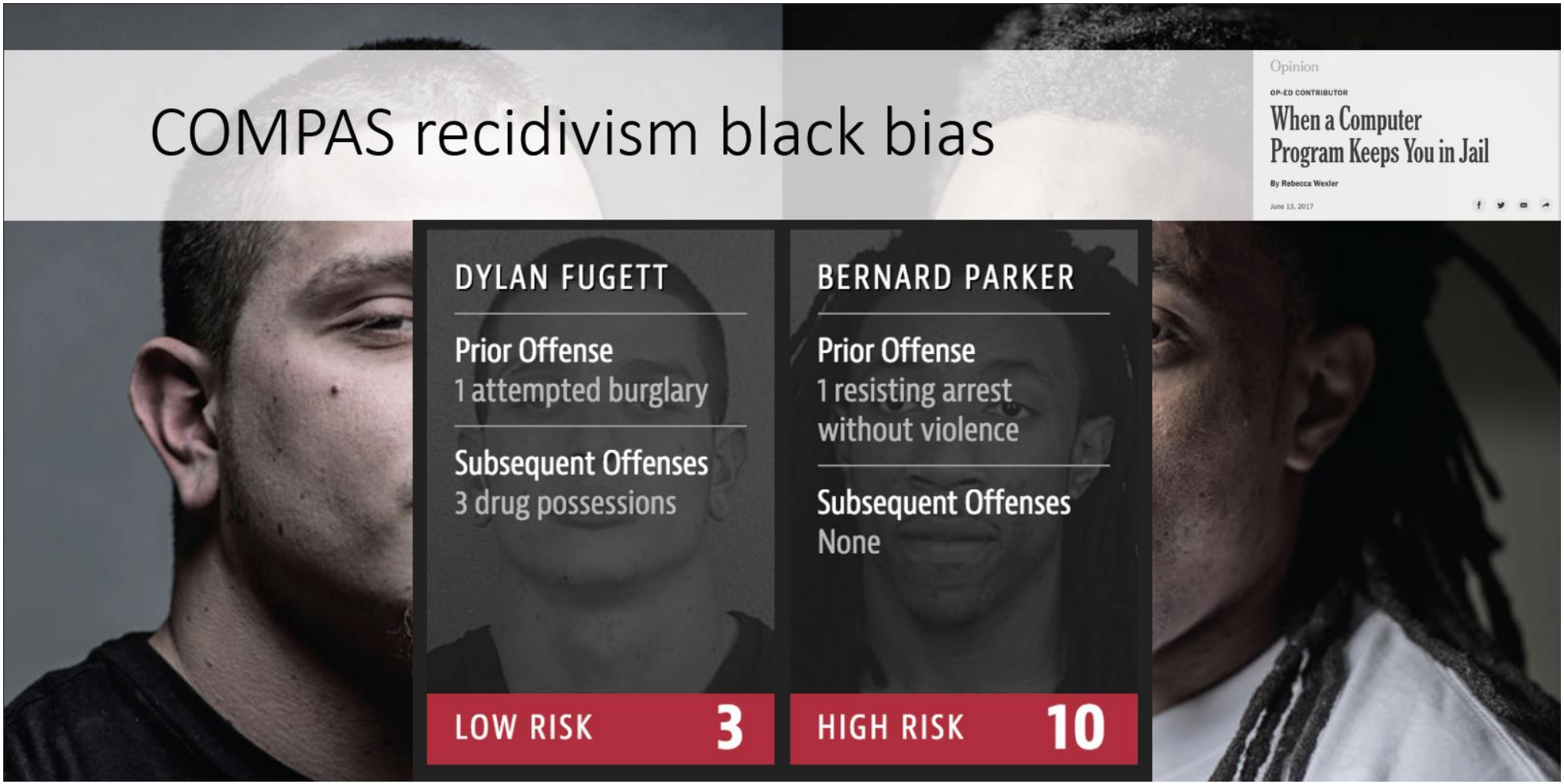
Case Study: COMPAS

Case Study: COMPAS

- COMPAS gives very different outcomes for white vs black defendants, but it does not use race as an input to the algorithm
- Even if a sensitive feature (e.g. race) is not an input to the algorithm, other features (e.g. zip code) may correlate with the sensitive feature

Case Study: COMPAS

Case Study: COMPAS



Fairness Metrics: Statistical Parity

Fairness Metrics

- ***Statistical Parity*** (a.k.a. Demographic Parity): requires the probability of receiving a positive outcome to be equal across all groups
  - $P(\hat{Y} = 1 \mid A = a) = P(\hat{Y} = 1 \mid A = b)$ for all values a, b of the protected attribute A
  - Notation
    - $\hat{Y}$ – prediction by the model
    - A is a protected attribute like race, gender, or age
    - a and b are two different groups
- E.g., in a loan approval system, the approval rate must be the same for all demographic groups, regardless of other factors
  - Let's assume the protected attribute A is gender, with groups $a = female$ and $b = male$
  - $\hat{Y} = 1$ means the model outputs a positive decision
  - $P(\hat{Y} = 1 \mid A = a)$ is the probability the loan is approved for a female applicants
  - $P(\hat{Y} = 1 \mid A = b)$ is the probability the loan is approved for a male applicants
    - If 200 women applied for the loan, and 120 were approved, then: $P(\hat{Y} = 1 \mid A = a) = 0.6$
    - If 250 men applied for the loan, and 200 were approved, then: $P(\hat{Y} = 1 \mid A = b) = 0.8$
    - Fair model should have approximately the same approval rate for men and women

Other Fairness Metrics

Fairness Metrics

- ***Equalized Odds:*** requires that both true positive and false positive rates are equal across groups
  - E.g., the COMPAS example
- ***Equal Opportunity:*** focuses on fairness for qualified individuals, ensuring they have the same chance of favorable outcomes regardless of their protected characteristics
  - $P(\hat{Y} = 1 \mid Y = 1,\ A = a) = P(\hat{Y} = 1 \mid Y = 1,\ A = b)$
  - $\hat{Y} = 1$ means the model predicts a positive outcome, and $Y = 1$ means the person is qualified
  - Qualified individuals have equal chance of positive prediction regardless of group
    - E.g., qualified candidates from all demographics have equal chance of receiving job interview

Other Fairness Metrics

Fairness Metrics

- ***Predictive Parity:*** focuses on fairness in positive predictions, ensuring they are equally reliable across all demographic groups regardless of protected characteristics
  - $P(Y = 1 \mid \hat{Y} = 1,\ A = a) = P(Y = 1 \mid \hat{Y} = 1,\ A = b)$
  - When the model predicts a positive outcome, that prediction is equally likely to be correct for all demographic groups
    - E.g., loan approvals have same default rate regardless of borrower demographics

- Different fairness metrics capture different aspects of what it means for a system to be "fair"
- Unfortunately, mathematical impossibility results show that these different fairness criteria cannot all be satisfied simultaneously when groups have different base rates, forcing practitioners to make principled trade-offs

Testing for Bias in Language Models

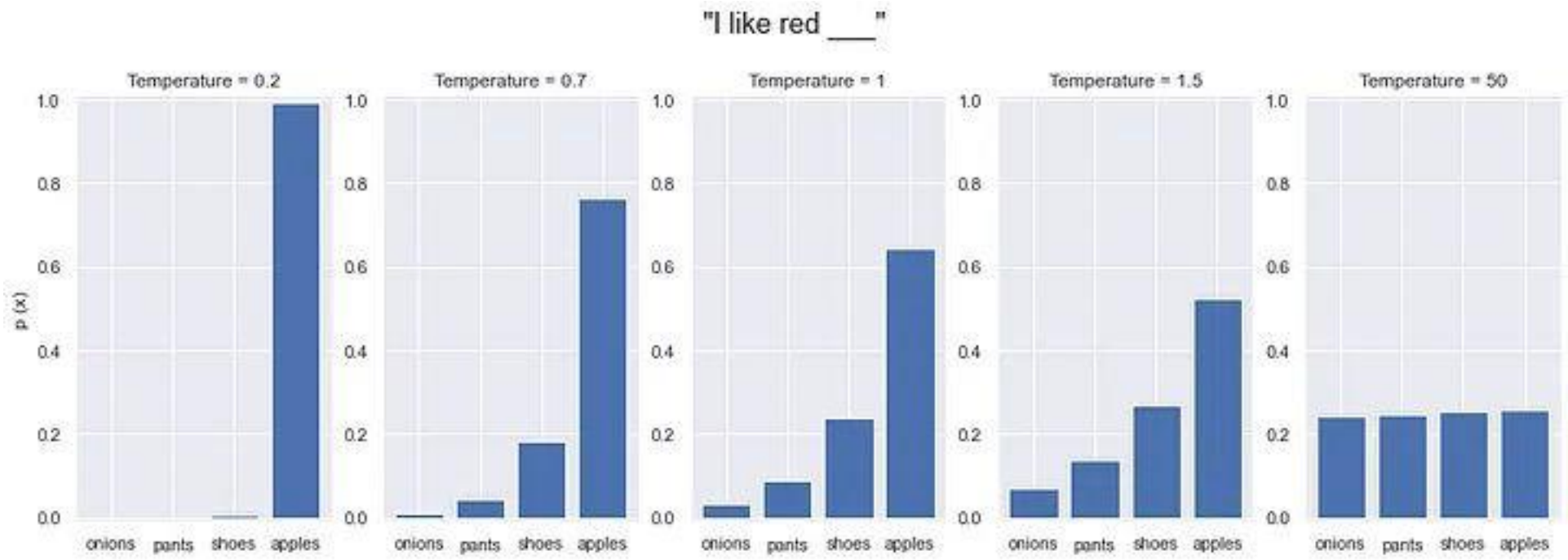
Bias in LLMs

- Kotek (2024) "Gender bias and stereotypes in Large Language Models"
  - Researchers tested prompts like "The doctor phoned the nurse because she was late for the morning shift."
  - The AI often inferred that the nurse was female and the doctor male
- Findings
  - On average, models were 6.8 times more likely to assign stereotypical female occupations to female pronouns and male occupations to male pronouns
  - Reinforces gender stereotypes, potentially affecting how language models are applied in industries like hiring and social media
- Impact
  - This kind of bias, if unchecked, may influence societal expectations and perpetuate harmful stereotypes
  - Biases affect users' trust in AI and the decisions influenced by AI

Randomness and Bias in LLMs

Bias in LLMs

- To mitigate the potential harms of biased training data, LLM developers often implement safeguards such as fine-tuning and alignment techniques to reduce biased outputs
  - However, increasing the model's temperature introduces more randomness into its responses, which can reduce the consistency and reliability of the safeguards
  - This raises concerns about the effectiveness of bias mitigation under high-temperature settings



Mitigation Strategies: Pre-processing

Mitigation Strategies

- ***Pre-processing techniques*** focus on addressing bias in the training data
  - By modifying the dataset before ML model training, these approaches can help prevent biases from being encoded into the model
  - Pre-processing strategies are model-agnostic and can be applied regardless of the specific algorithm being used
- Pre-processing techniques include:
- Data reweighing
  - Assign different weights to training examples to counteract representation imbalances
  - Examples from underrepresented groups receive higher weights, increasing their influence during model training
    - Reduces impact of majority group dominance
    - Preserves all original data points
    - Helps model learn patterns from minority groups

Mitigation Strategies: Pre-processing

Mitigation Strategies

- Strategic resampling
  - Modify the training dataset by either duplicating instances from underrepresented groups (over-sampling) or removing instances from overrepresented groups (under-sampling)
    - Creates more balanced class distribution
    - Relatively simple to implement
    - Can be combined with other techniques
- Synthetic data generation
  - Create artificial data points to increase representation of underrepresented groups
  - E.g., using techniques like SMOTE (Synthetic Minority Over-sampling Technique) or generative models
    - Addresses data imbalance
    - Creates novel training examples

Mitigation Strategies: In-processing

Mitigation Strategies

- ***In-processing techniques*** modify the ML algorithm to incorporate fairness considerations during training
  - This can be achieved via constrained optimization methods that balance prediction accuracy with fairness metrics
- Advantages
  - Incorporate fairness constraints directly into the learning algorithm
  - Add regularization terms that penalize unfair solutions during training
  - Optimize for both accuracy and fairness as dual objectives

Mitigation Strategies: Post-processing

Mitigation Strategies

- ***Post-processing techniques*** adjust the ML predictions after training is complete, such as applying different thresholds for different groups to match error rates
  - These approaches offer flexibility but may sacrifice some predictive performance to achieve fairness goals
- Techniques
  - Adjust model outputs after training to satisfy fairness criteria
  - Apply different decision thresholds for different groups
  - Calibrate probability estimates across protected groups
- Example: identify uncertain predictions near decision boundary and apply fairness-oriented rules
- While pre-processing strategies focus on the data, in-processing and post-processing strategies address bias within the algorithm itself or its outputs

Towards Ethical AI

Towards Ethical AI

- Building fair and unbiased ML systems requires continuous effort
  - It requires vigilance at every stage of the AI lifecycle, from problem formulation to deployment and monitoring
- Transparency by ML developers is crucial, both in documenting design choices and in communicating limitations to users and stakeholders
  - This includes acknowledging trade-offs between different fairness criteria and being honest about the imperfect nature of all fairness solutions

Steps Towards Ethical AI

Towards Ethical AI

- Education and awareness
  - Building fairness literacy among practitioners
- Community engagement
  - Including affected communities in design and evaluation
- Transparency and explainability
  - Making AI systems understandable to stakeholders
- Multidisciplinary collaboration
  - Bringing together technical, ethical, legal, and domain experts
- Continuous improvement
  - Ongoing monitoring, auditing, and refinement of ML systems

References

- Aditi Rajesh Shah, Nick Szydlowski, "AI Bias: A Hands-On Workshop and Discussion," available at https://tiny.sjsu.edu/ai-bias
- Justin Joshnson, David Fouhey, "Lecture: Fairness in AI,", EECS 442, University of Michigan, 2021