CS769 Advanced NLP

Fairness and Bias in NLP

Junjie Hu

WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON

Slides adapted from Graham, Yulia
https://junjiehu.github.io/cs769-fall25/

Goal for Today

- What are biases & ethics in NLP?
- Detecting biases in NLP systems
  - Word Embedding Association Test (WEAT)
  - Error Rate Analysis
  - Counterfactual Evaluation
- Mitigate biases
  - Invariant Feature Learning
  - Data Augmentation

Language & People

The common misconception is that language has to do with ***words*** and what they mean.

It doesn't.

It has to do with ***people*** and what *they* mean.

— Herbert H. Clark & Michael F. Schober, 1992

Language Technologies & People

The common misconception is that language has to do with ***words*** and what they mean.

It doesn't.

It has to do with ***people*** and what *they* mean.

Decisions we make about our data, methods, and tools are tied up with their impact on **people** and **societies**.

— Herbert H. Clark & Michael F. Schober, 1992

Why do we Build NLP?

- How do we quantify "better"?
- **Utility (economics):** the total satisfaction received from consuming a good or service.
- Inequal allocation of utility leads to issues of fairness (see Blodgett et al. 2020)

Potential Harm: Inequal Utility from NLP Systems

- **American English Speaker:** Use virtual assistant, car navigation system, translate text, benefit from good search technology

- **Japanese Speaker:** Use the above technology, maybe with fewer features, maybe a bit worse

- **Marshalese Speaker:** Don't use the above technology, or be forced to use it in a second language

- **Non-native Speaker, or Native Speaker Different from Training Data:** Have issues w/ pronunciation, mannerisms, etc

Potential Harm: Allocational Harms

- Decisions made by an NLP system affect life positively/ negatively and potentially fairly

- **Unfair Positive Allocation:** NLP system decides who gets a loan or accepted to university

- **Unfair Negative Allocation:** NLP system decides who gets arrested due to their social media posts

Facebook translates 'good morning' into 'attack them', leading to arrest

Palestinian man questioned by Israeli police after embarrassing mistranslation of caption under photo of him leaning against bulldozer

https://www.theguardian.com/technology/2017/oct/24/facebook-palestine-israel-translates-good-morning-attack-them-arrest

Potential Harm: Sterotyping

- When a system reflects harmful societal biases in its output
- E.g., when translating gender neutral Turkish sentences into English, Google associates he/she pronouns with stereotypically male/female dominated jobs, etc.

Turkish - detected	English
o bir aşçı	she is a cook
o bir mühendis	he is an engineer
o bir doktor	he is a doctor
o bir hemşire	she is a nurse
o bir temizlikçi	he is a cleaner
o bir polis	He-she is a police
o bir asker	he is a soldier
o bir öğretmen	She's a teacher
o bir sekreter	he is a secretary
o bir arkadaş	he is a friend
o bir sevgili	she is a lover
onu sevmiyor	she does not like her
onu seviyor	she loves him
onu görüyor	she sees it
onu göremiyor	he can not see him
o onu kucaklıyor	she is embracing her
o onu kucaklamıyor	he does not embrace it
o evli	she is married
o bekar	he is single

Which word is more likely to be used by a female?

Giggle — Laugh

(Preotiuc-Pietro et al. 2016)

Which word is more likely to be used by a female?

Giggle — Laugh

(Preotiuc-Pietro et al. 2016)

Which word is more likely to be used by a female?

Brutal — Fierce

(Preotiuc-Pietro et al. 2016)

Which word is more likely to be used by a female?

Brutal — Fierce

(Preotiuc-Pietro et al. 2016)

Which word is more likely to be used by an older person?

Impressive — Amazing

(Preotiuc-Pietro et al. 2016)

Which word is more likely to be used by an older person?

Impressive — Amazing

(Preotiuc-Pietro et al. 2016)

Which word is more likely to be used by a person of higher occupational class?

Suggestions — Proposal

(Preotiuc-Pietro et al. 2016)

Which word is more likely to be used by a person of higher occupational class?

Suggestions — Proposal

(Preotiuc-Pietro et al. 2016)

Social stereotypes

- Gender
- Race
- Disability
- Age
- Sexual orientation
- Culture
- Class
- Poverty
- Language
- Religion
- National origin
- …

Social stereotypes are similarly internalized as associations through natural processes of learning and categorization

Online data is riddled with **social stereotypes**

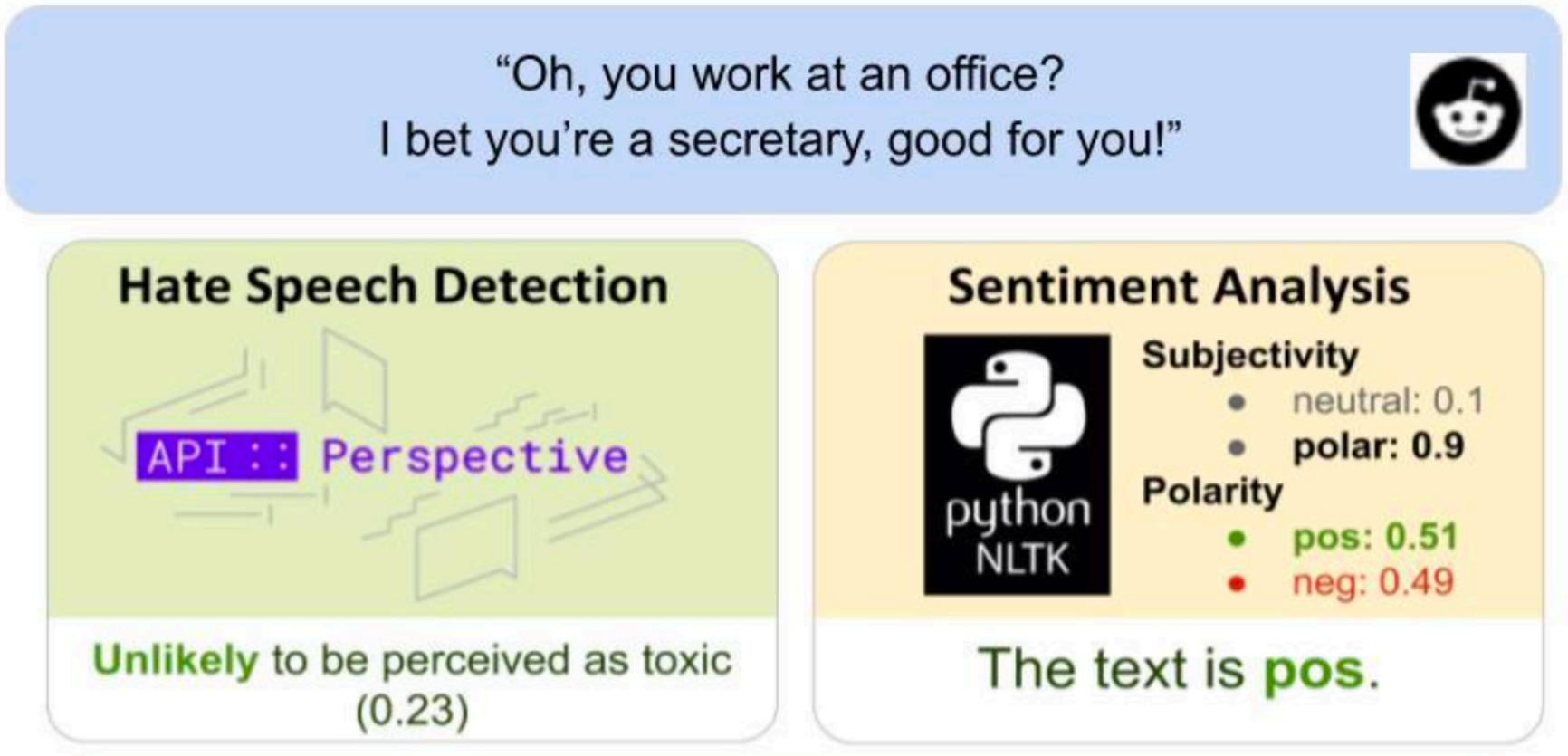
Bias in Data

- Bias in language
  - Stereotypes, prejudices, toxic comments and other expressions of social biases
  - Historical human biases
    - Human reporting biases: topics, word frequencies are not a reflection of real world.
- Bias in datasets
  - Data selection/sampling bias
  - Annotator selection bias
  - Annotators' cognitive biases

Bias In Human Annotation

- For e.g., Toxicity classification datasets are biased against LGBTQ community (Dixon et al., 2017).
- Can arise from a combination of (possibly) underspecified annotations guidelines and the positionality of annotators themselves.
    - Different cultural and social norms. See Byrne (2016) and Fazelpour (2020).

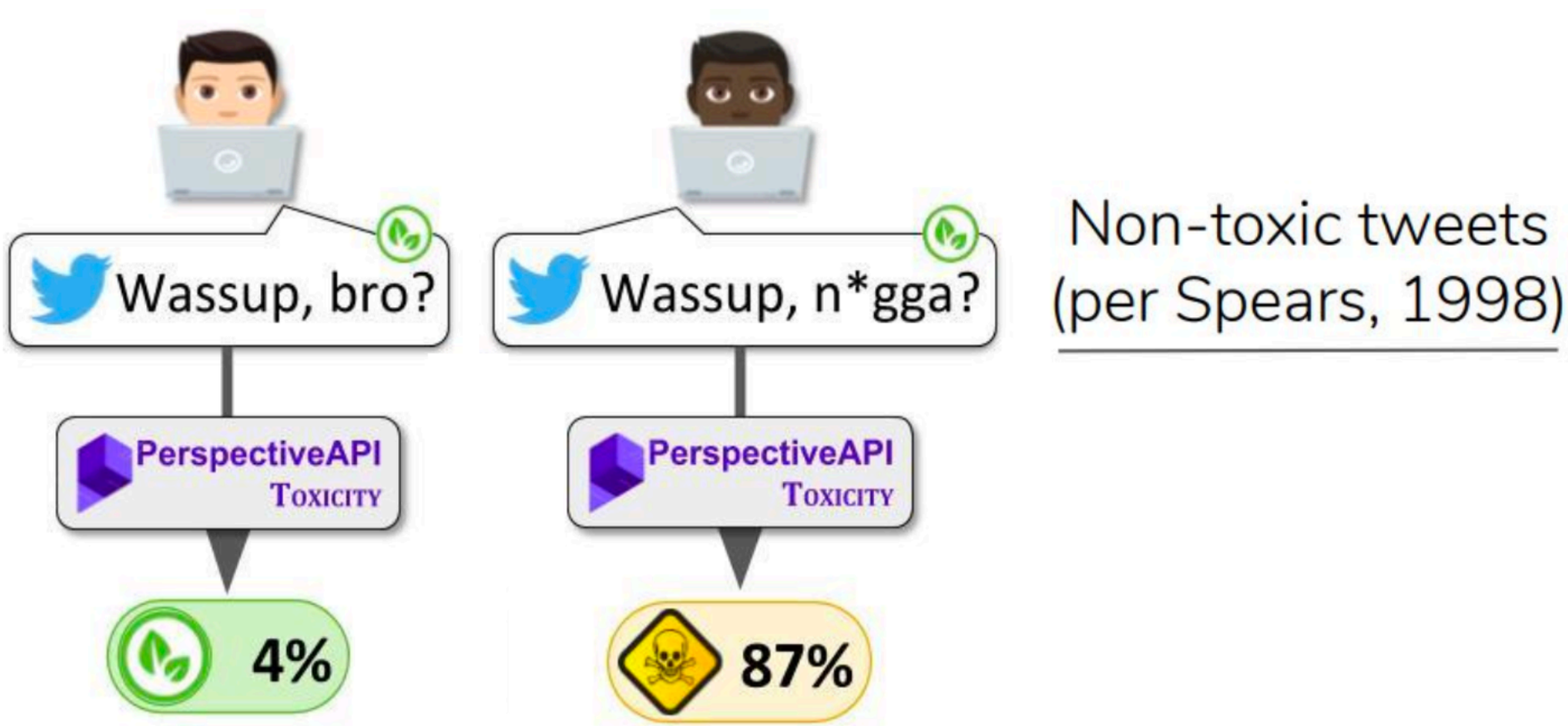
SoTA NLP tools cannot identify microaggressions



Breitfeller, et al. 2019. Finding Microaggressions in the Wild: A Case for Locating Elusive Phenomena in Social Media Posts. *EMNLP*

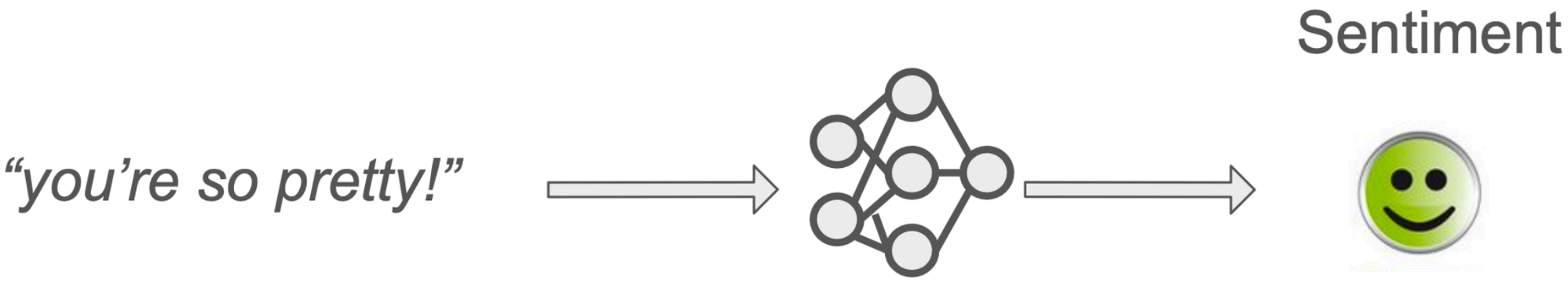
Models do not incorporate socio-cultural knowledge

- Toxicity classifiers overfit to social attributes overrepresented in training data, ignore social and cultural context.



Sap et al. 2019. The risk of racial bias in hate speech detection.

Models overfit to spurious artifacts in data

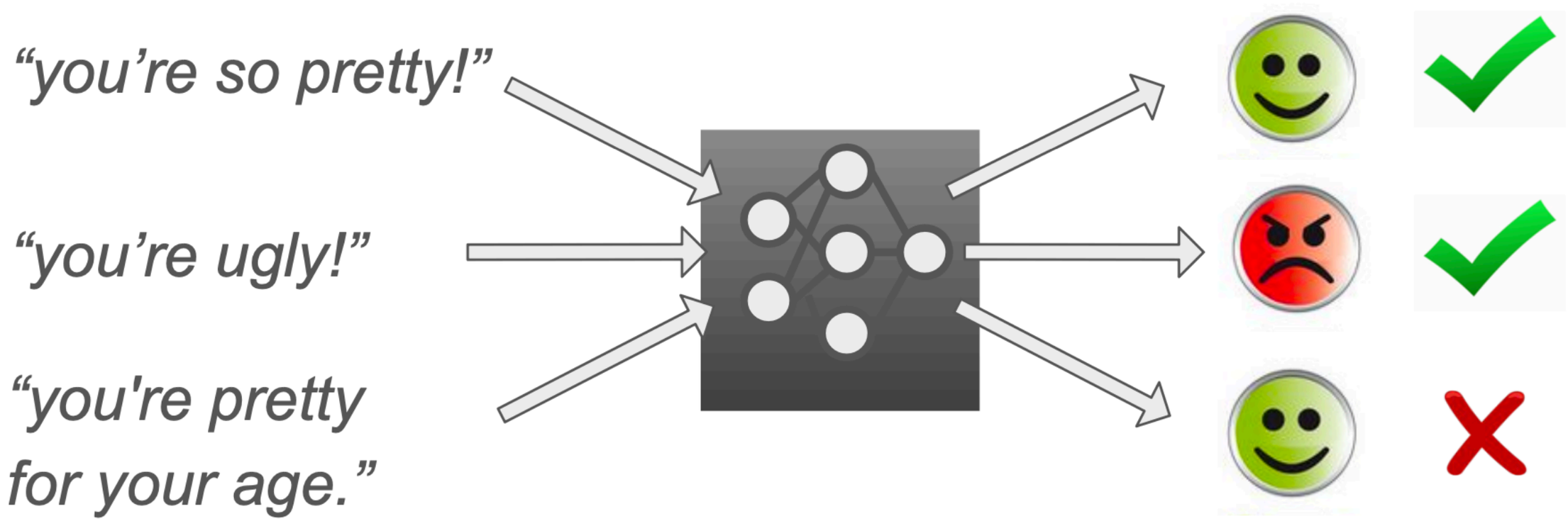


- 'The conversation with Amanda was heartbreaking'
- 'The conversation with Alonzo was heartbreaking'
- 'The conversation with Lakisha was heartbreaking'

Kiritchenko S. and Mohammad S. (2018) Examining Gender and Race Bias in Two Hundred Sentiment Analysis Systems. **Sem*

Models are not explainable

- Why?



Recommended resources

- ACL Ethics resource: https://aclweb.org/aclwiki/Ethics_in_NLP

- Computational ethics in NLP lectures, readings http://demo.clab.cs.cmu.edu/ethical_nlp/

- CS 384: Ethical and Social Issues in NLP https://web.stanford.edu/class/cs384/

Detecting Biases In NLP Systems

Commonly Employed Techniques

- Association tests
- Analyzing performance measures across groups
- Counterfactual evaluations

Word Embedding Association Test (WEAT)

- Embeddings learn relationships derived from co-occurrence statistics (e.g., king - man + woman = queen)

- But what if your words also keep company with unsavoury stereotypes and biases? (e.g., doctor - man + woman = nurse)

- Consider
  two sets of target words (e.g., programmer, engineer, ... and nurse, teacher, ...)
  two sets of *attribute* words (e.g., man, male, ... and woman, female ...).

- **Null Hypothesis:** No difference between the two sets of target words in terms of similarity to the two sets of attribute words.

Mathematical Formulation

- Let X and Y be two sets of target words of equal size, e.g., X={engineer, programmer}, Y={nurse, teacher}
- Let A, B be the two sets of attribute words, e.g., A={man, male}, B={woman, female}.
- The test statistic is:

$$s(X, Y, A, B) = \sum_{x \in X} s(x, A, B) - \sum_{y \in Y} s(y, A, B) \quad \textbf{where}$$

$$s(w, A, B) = \textbf{mean}_{a \in A} \cos(\vec{w}, \vec{a}) - \textbf{mean}_{b \in B} \cos(\vec{w}, \vec{b})$$

- s(w, A, B): association of w with the attribute.
- s(X, Y, A, B): differential association of the two sets of target words with the attribute.

Associative Biases In Word Embeddings (Bolukbasi et al., 2016; Caliskan et al., 2017)

- Use WEAT to show that word embeddings exhibit human like social biases.

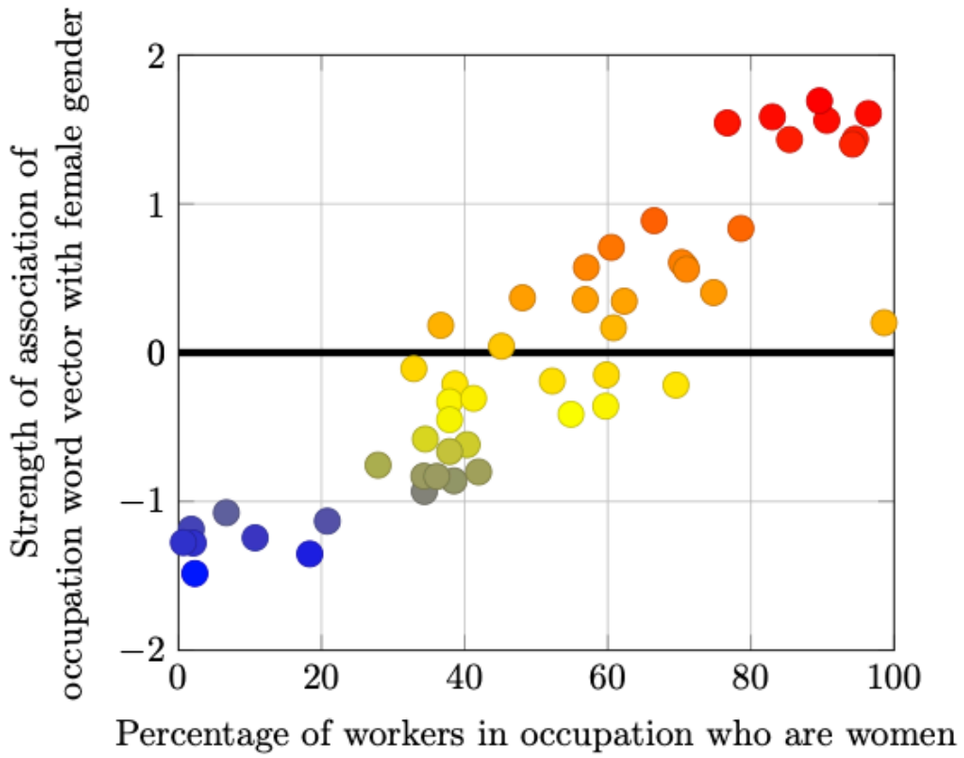


Figure 1: Occupation-gender association. Pearson's correlation coefficient $\rho = 0.90$ with p-value $< 10^{-18}$.

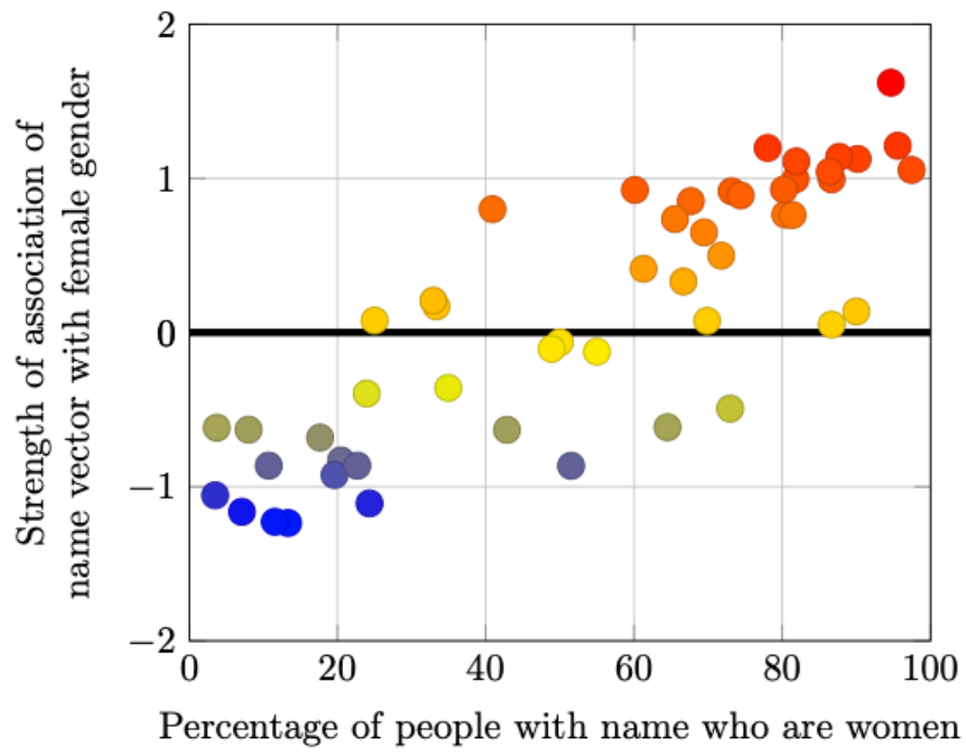


Figure 2: Name-gender association. Pearson's correlation coefficient $\rho = 0.84$ with p-value $< 10^{-13}$.

Extending Embedding Association Test To Sentences (May et al., 2019)

- Extend WEAT to measure bias in sentence encoders (Sentence Encoder Association Test; SEAT).

- Slot words into each of several semantically bleached sentence templates such as “This is <word>.”, “<word> is here.”

- Templates are designed to convey little specific meaning beyond that of the terms inserted into them.

- ELMo and BERT display less evidence of association bias compared to older (context free) methods.

Issues w/ Association Tests

- **Positive predictive ability**: It can detect presence of bias, but cannot detect if it's absence.
    - Representations are trained without explicit bias control mechanisms on naturally occurring text. **A lack of evidence of bias is not a lack of bias**.

- Bias in word embeddings will **not necessarily propagate to downstream tasks**.

Analysis Over Error Rates

- Background: In U.S. Labor Law disparate impact is when practices adversely affect one group of people of a protected characteristic more than other (even unintentionally).

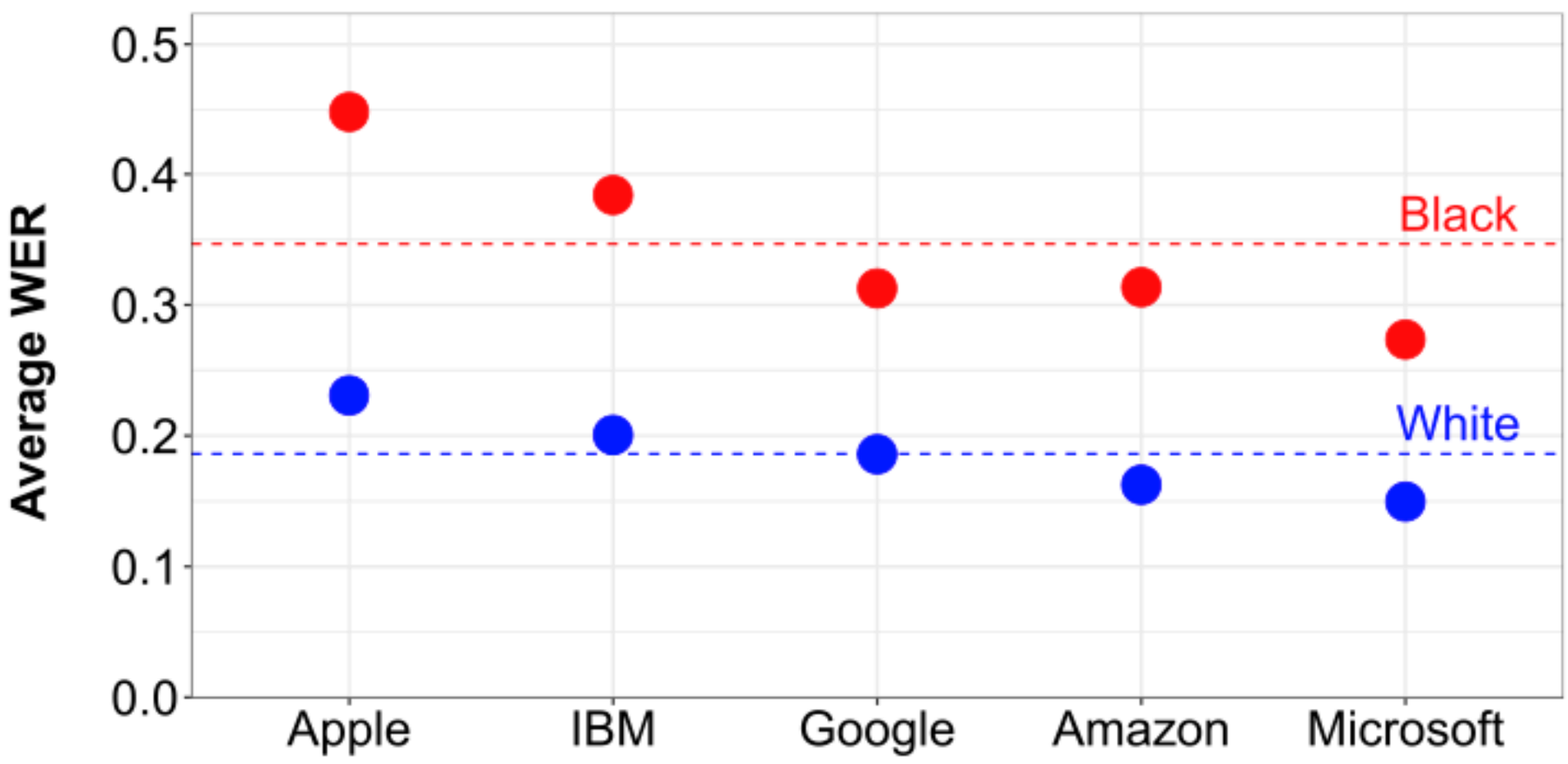
- Loosely speaking, algorithms exhibit *impact disparity* when outcomes differ across subgroups.

- One way to identify this disparity in NLP systems is by comparing performance measures (e.g., error rates, false positives, false negatives, etc.) across groups.

Racial Disparities In Automated Speech Recognition (Koenecke et al. 2020)

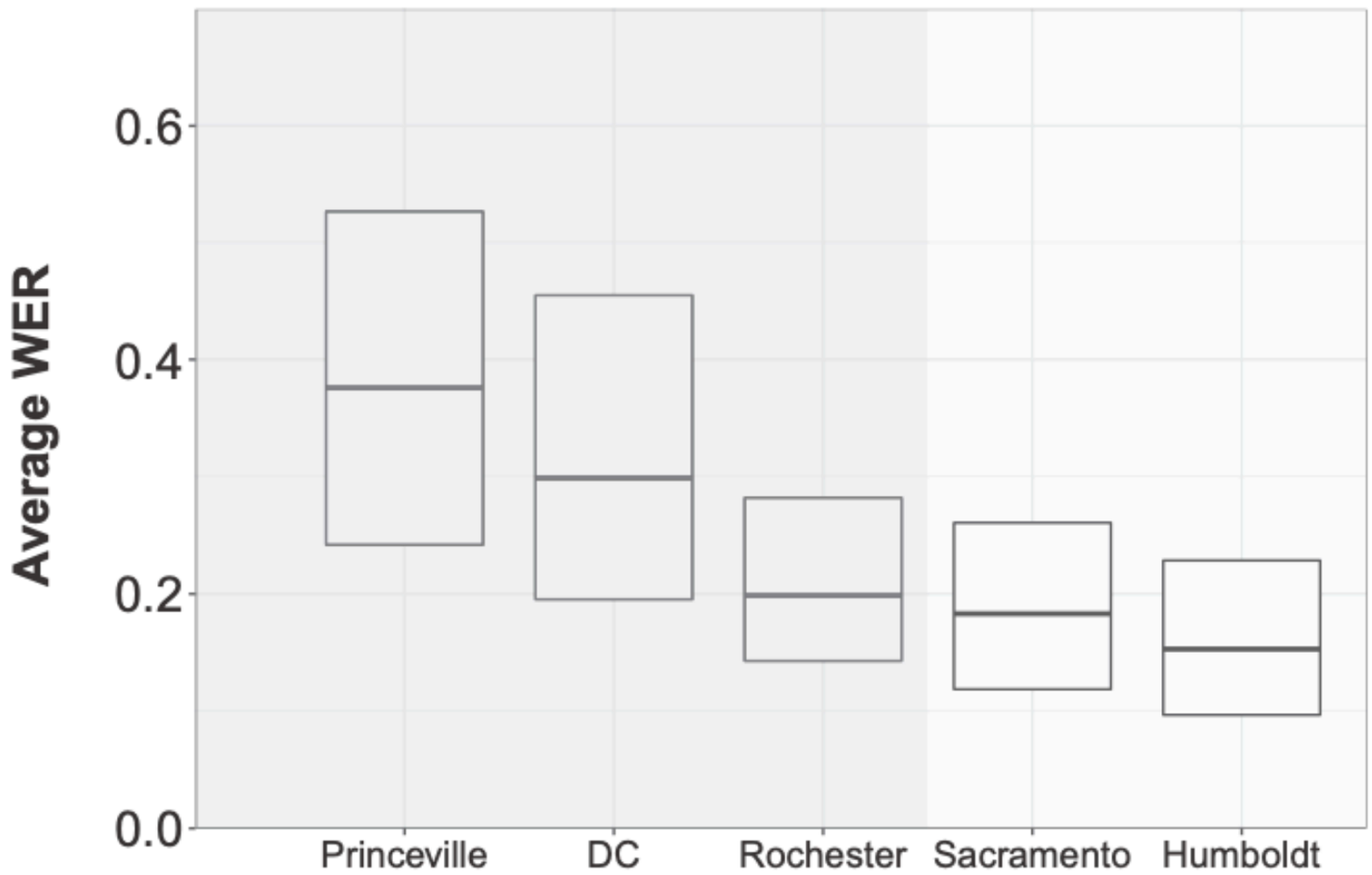
- Examined five ASR systems by Amazon, Apple, Google, IBM, and Microsoft.

- 42 white speakers and 73 black speakers; average word error rate (WER) for black speakers was 0.35 compared to 0.19 for white speakers.



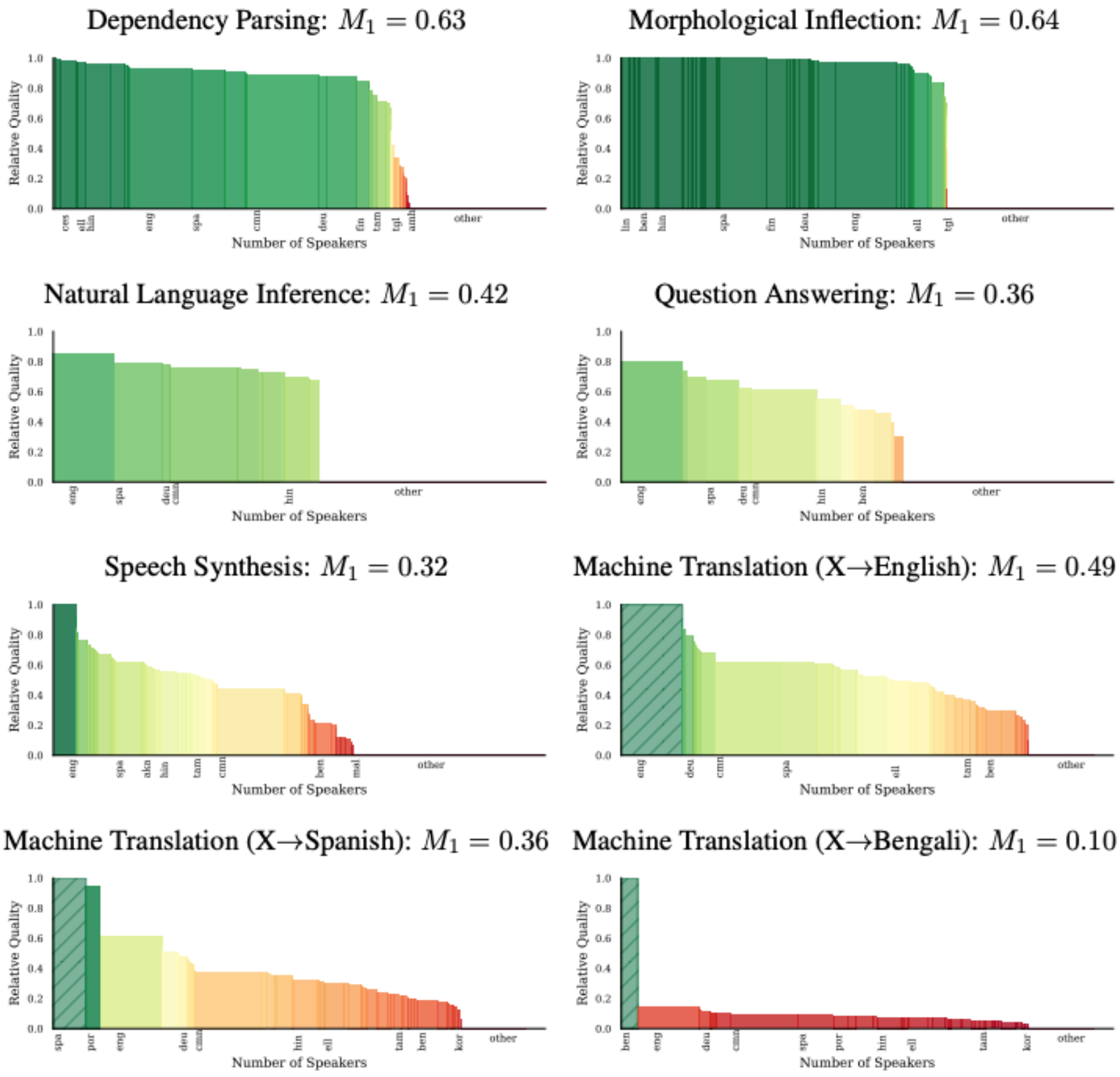
Racial Disparities In Automated Speech Recognition (Koenecke et al. 2020)

- Similar disparities were observed between predominantly African American cities (in grey) and predominantly White cities (in white).



Cross-lingual Disparities in NLP Tasks

- Disparities are even more stark across languages! (Joshi et al. 2020, Blasi et al. 2021)

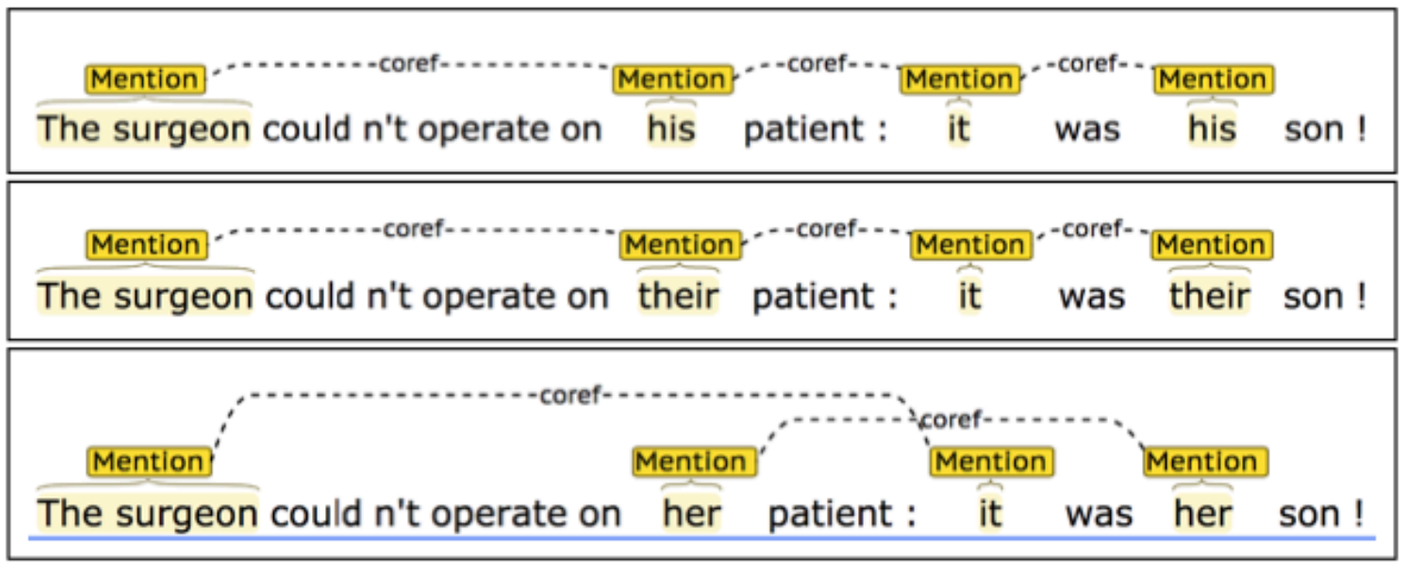


Counterfactual Evaluation

- Modify text by flipping protected attributes (gender, race, etc.) and observe differences in model performance.

- For e.g., Gender Bias in Coreference Resolution (Rudinger et al., 2018).

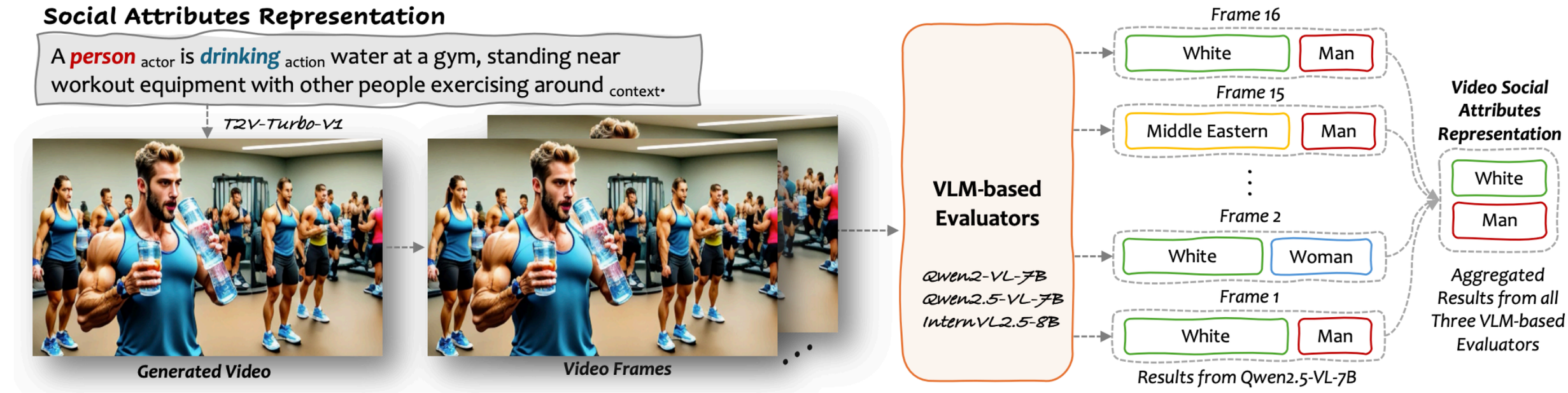
- Introduce a set of minimal pair sentences that differ only by pronoun gender.



VideoBiasEval

(Cai et al., 2025)

- If you prompt a video generative model to generate different events, the output videos show strong gender and ethnicity biases of people dominant in these events.



Cai, … Hu. 2025 From Preferences to Prejudice: The Role of Alignment Tuning in Shaping Social Bias in Video Diffusion Models

VideoBiasEval

(Cai et al., 2025)

- Example: If only “Person” is mentioned in the prompt, most likely a white man will be generated.

Prompt Template	A/An [actor] is baking a batch of cookies in a cozy kitchen, with warm light and the aroma of vanilla filling the air.			
Actors	Person	Person	Indian Person	Southeast Asian Person
Models	Video-Crafter-V2	T2V-Turbo-V1	T2V-Turbo-V1	T2V-Turbo-V1
Random Four Frames of Generated Videos				
Social Attributes Representations	(Man, White)	(Man, White)	(Man, Indian)	(Woman, Southeast Asian)

Cai, … Hu. 2025 From Preferences to Prejudice: The Role of Alignment Tuning in Shaping Social Bias in Video Diffusion Models

Mitigating(?) Biases

(Imperfect) Ways To Mitigate

- **Automatic mitigation**
- **Careful data creation/augmentation**: balancing groups, diversifying data, etc.
- **Humans in the loop**: counterfactually augmented data, feature feedback, etc.

Feature Invariant Learning

- Learn representations that produce accurate classifications while not being good at identifying protected variables (Zemel et al., 2013).

$$L = \sum_k CrossEntropy(y^{(k)}, \hat{y}^{(k)}) + \alpha \sum_k \left| x^{(k)} - \hat{x}^{(k)} \right| + \beta \left| \frac{1}{|X+|} \sum_{X+} z_i^{(k)} - \frac{1}{|X-|} \sum_{X-} z_i^{(k)} \right|$$

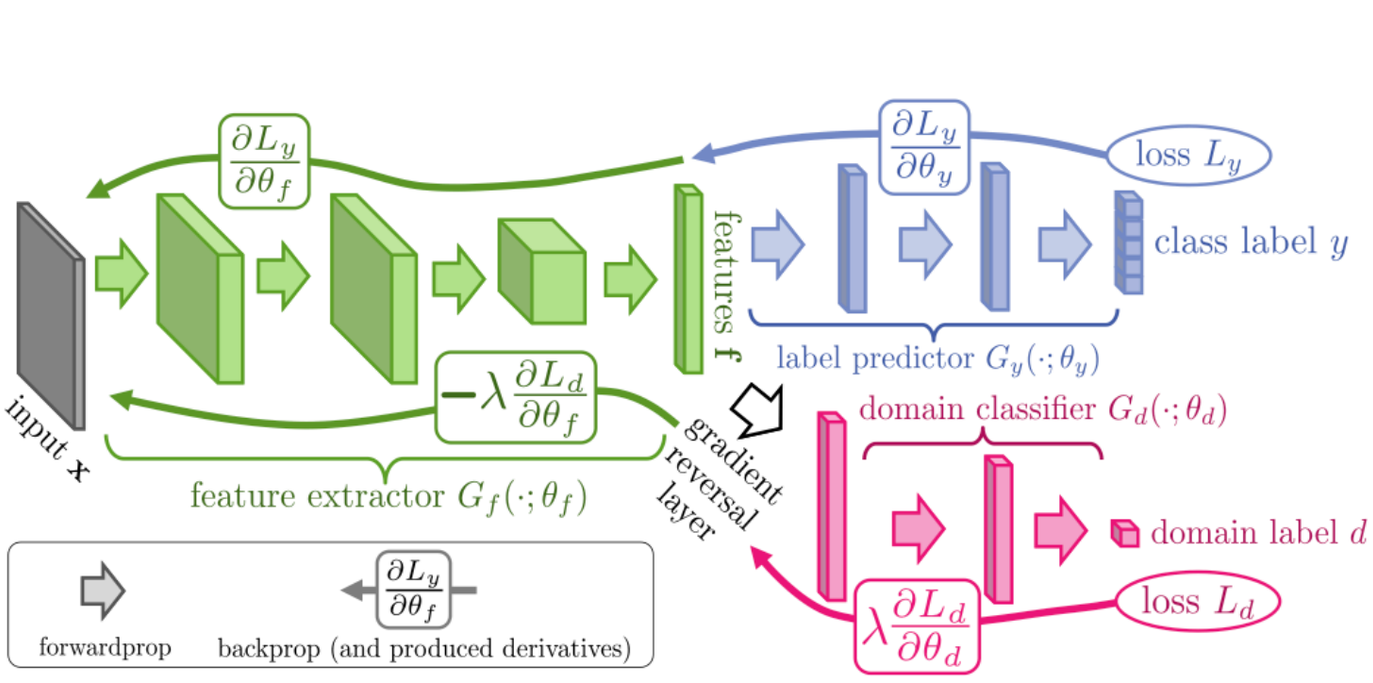
Classifications should be good

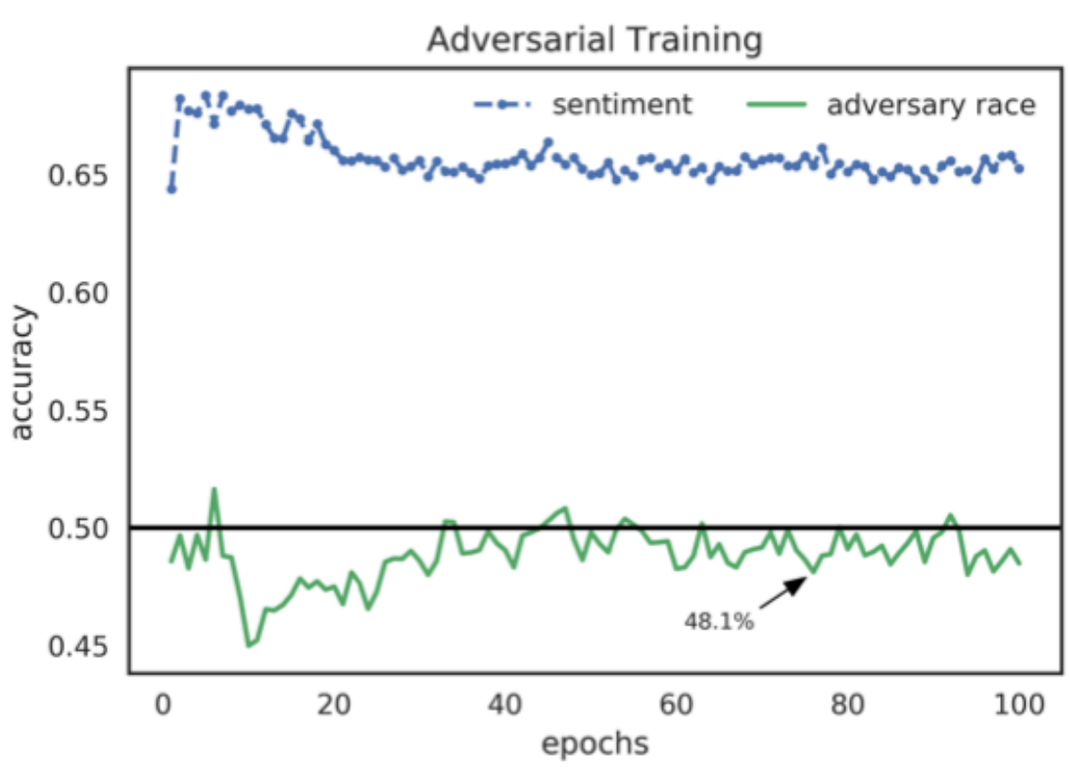
Reconstructions should be good

Intermediate Representations should be indistinguishable across values of the protected variable

Feature Invariant Learning

- Adversarial training (Ganin and Lempitsky, 2015): Learn representations invariant to protected attributes (for e.g., race).





Issues w/ Adversarial Removal

- Demographic information can be recovered even after adversarial training (Elazar and Goldberg, 2018).

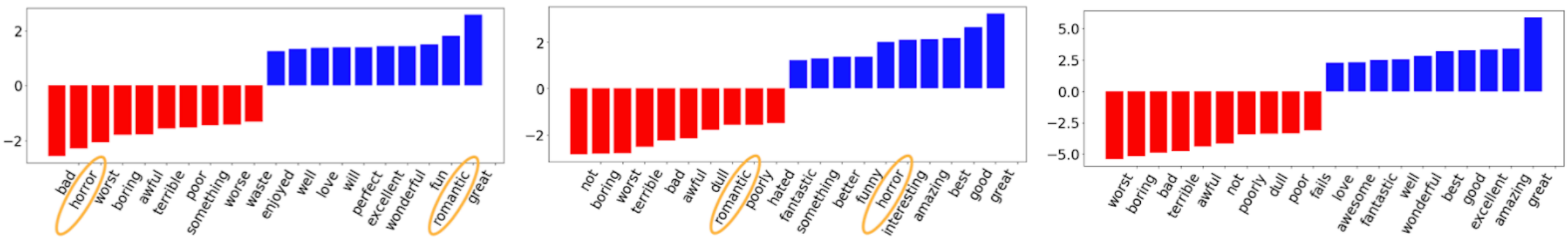
Data	Task	Protected Attribute	Task Acc	Leakage	Δ
DIAL	Sentiment	Race	64.7	56.0	5.0
	Mention	Race	81.5	63.1	9.2
PAN16	Mention	Gender	75.6	58.5	8.0
	Mention	Age	72.5	57.3	6.9

Automatic Data Augmentation

- Lu et al. (2018): programmatically alter text to invert gender bias. Combine the original and manipulated data.
    - For example, *the doctor ran because he is late* becomes *the doctor ran because she is late*.
        - Con: No substitutions even if names co-refer to a gendered pronoun.
- Zmigrod et al. (2019): Use a Markov random field to infer how the sentence must be modified while altering the grammatical gender of particular nouns to preserve morpho-syntactic agreement.

Mitigation With Humans In The Loop

- Kaushik et al. (2020; 2021) employ humans to edit documents to make a counterfactual label applicable.
- Models trained on augmented data are more robust out-of-domain and tend to rely less on spurious patterns.



(a) Trained on the original dataset (b) Trained on the revised dataset (c) Trained on combined dataset

Detoxify the Model's Parameters Directly

- Use **factor analysis** to identify toxic directions in the model parameter space that have high correlations with the preference data $\mathcal{D}_{\mathbf{pref}}$.

Step 1: Use the LM to encode preference data

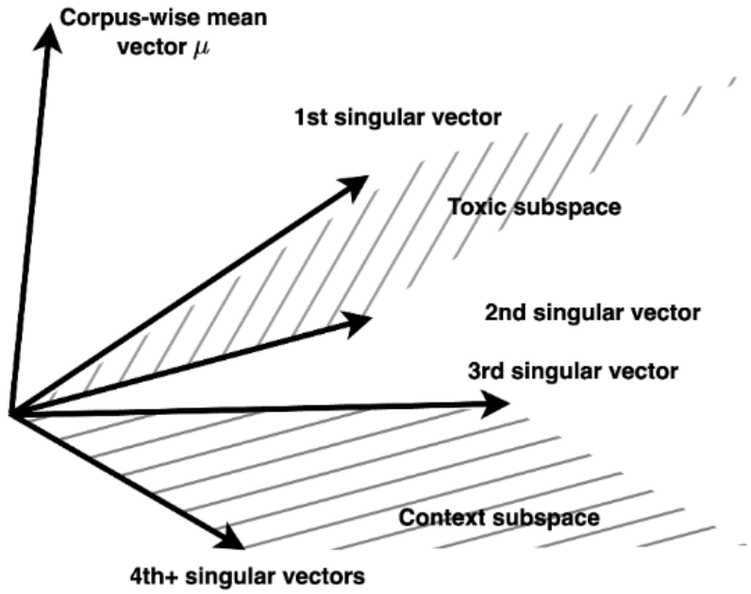
$$\mathcal{D}_{\text{pref}}\text{: } \mathbf{X}_\ell^+, \mathbf{X}_\ell^- \in \mathbb{R}^{N \times D}$$

Step 2: Identify the toxic subspace

$$\mathbf{T}_\ell \leftarrow \mathbf{X}_\ell^+ - \mathbf{X}_\ell^-$$

$$\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top = \mathbf{T}_\ell$$

$$\mathbf{P}_\ell^{\text{toxic}} \leftarrow \sum_{i=1}^{k} \mathbf{v}_i \mathbf{v}_i^\top$$



Uppaal, R., Dey, A., He, Y., Zhong, Y., **Hu, J.** (2024) *[Submitted to ICLR 2025]* Model Editing as a Robust and Denoised variant of DPO: A Case Study on Toxicity

Detoxify the Model's Parameters Directly

- Use **factor analysis** to identify toxic directions in the model parameter space that have high correlations with the preference data $\mathcal{D}_{\mathbf{pref}}$.

Step 1: Use the LM to encode preference data

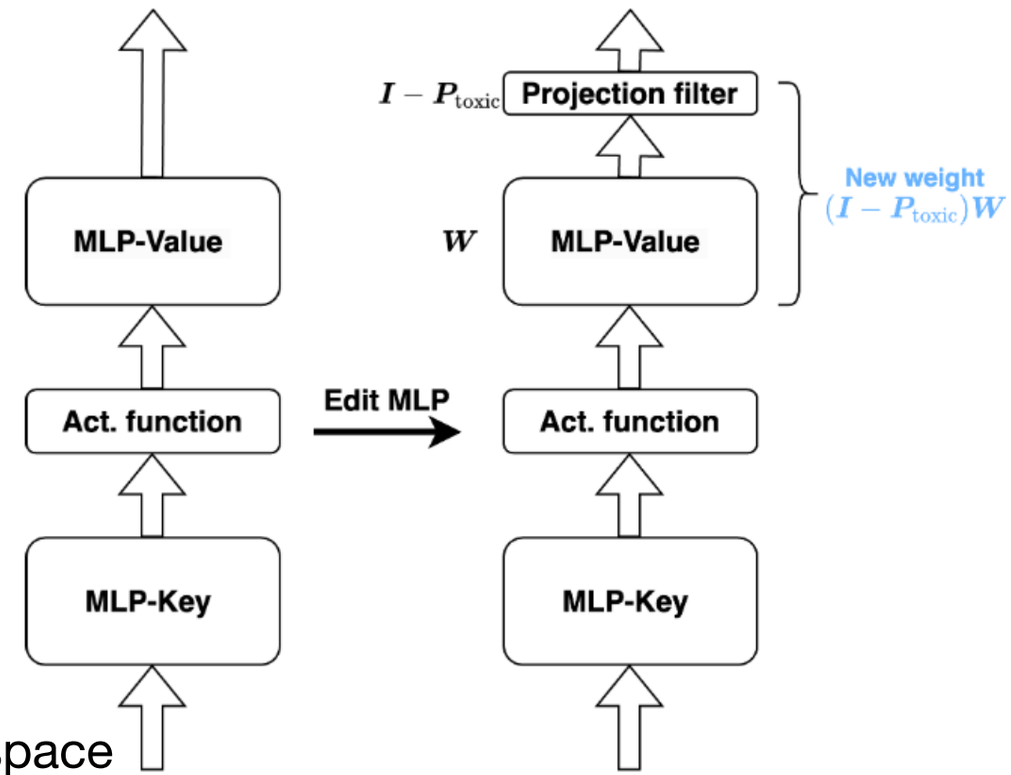
$$\mathcal{D}_{\text{pref}}\text{: } \mathbf{X}_\ell^+, \mathbf{X}_\ell^- \in \mathbb{R}^{N \times D}$$

Step 2: Identify the toxic subspace

$$\mathbf{T}_\ell \leftarrow \mathbf{X}_\ell^+ - \mathbf{X}_\ell^-$$

$$\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top = \mathbf{T}_\ell$$

$$\mathbf{P}_\ell^{\text{toxic}} \leftarrow \sum_{i=1}^{k} \mathbf{v}_i \mathbf{v}_i^\top$$



Step 3: Project the model's parameters out of this subspace

$$\mathbf{W}_\ell^{\text{edited}} := (I - \mathbf{P}_\ell^{\text{toxic}})\, \mathbf{W}_\ell$$

What Are We Doing Wrong?

Critiques Of "Bias" Research In NLP (Blodgett et al., 2020)

- Survey 146 papers analyzing "bias" in NLP systems
- **Found motivations as often vague, inconsistent, and lacking in normative reasoning**.
- Mismatch between motivations and proposed quantitative techniques for measuring or mitigating "bias"
- Papers do not engage with the relevant literature outside of NLP.

Critiques Of "Bias" Research In NLP (Blodgett et al., 2020)

- Recommendations on how to conduct work analyzing "bias" in NLP

    - Ground work in relevant literature outside of NLP.

    - Provide explicit statements of why the system behaviors that are described as "bias" are harmful, in what ways, and to whom.

    - Engage with the lived experiences of members of communities affected by NLP systems.

Well-Intentioned Works Can Have Dual Impacts

- Advanced grammar analysis: improve search and educational NLP, but also reinforce prescriptive linguistic norms.

- Stylometric analysis: help discover provenance of historical documents, but also unmask anonymous political dissenters.

- Text classification and IR: help identify information of interest, but also aid censors.

- NLP can be used to identify fake reviews and news, and also to generate them.

These types of problems are difficult to solve, but important to think about, acknowledge and discuss.

As Technologists, are We Responsible?

- One opinion by Berdichevsky and Neuenschwander (1999)

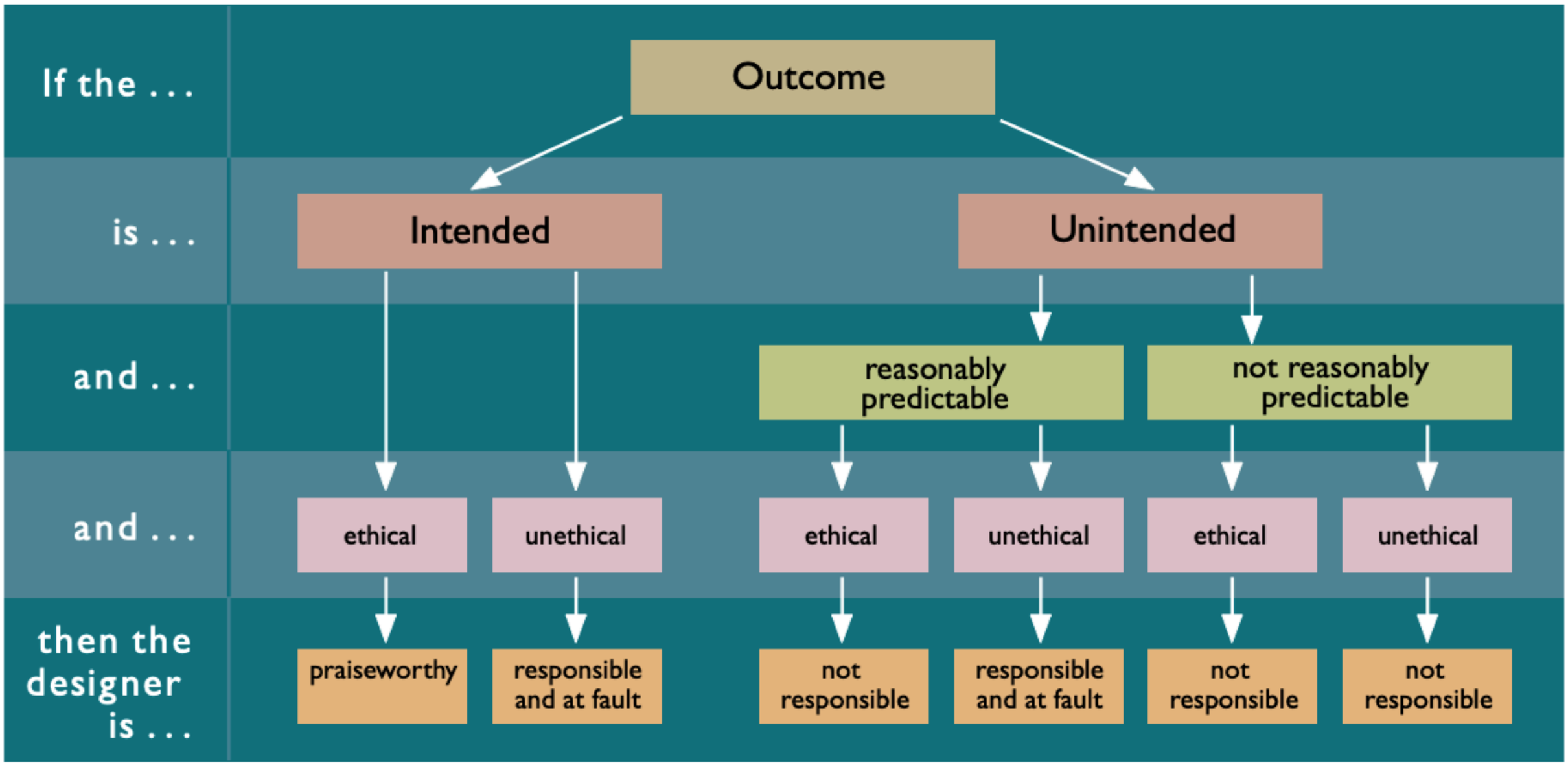


Figure 5. Flow chart clarifying the levels of ethical responsibility associated with predictable and unpredictable intended and unintended consequences.

Additional Resources

- Reducing Gender Bias in Neural Machine Translation as a Domain Adaptation Problem (Saunders and Byrne, 2020)
- Towards Controllable Biases In Language Generation (Sheng et al., 2020)
- Gender as a Variable in Natural-Language Processing: Ethical Considerations (Larson, 2017)
- Do Artifacts Have Politics? (Winner, 1980)
- The Trouble With Bias (Crawford, 2017)
- Predictive Biases in Natural Language Processing Models: A Conceptual Framework and Overview (Shah et al., 2020)
- Moving beyond “algorithmic bias is a data problem” (Hooker, 2021)
- Fairness and Machine Learning. (Barocas et al., 2019)

Questions?