Bias, Fairness, & Privacy

Nick Deas

COMS 4705 – Natural Language Processing

COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

How Do We Learn From Language?

How do we learn/model…

1. From language data
2. Token interactions/position
3. Specific applications and tasks
4. Human preferences
5. …

How Do We Learn From Language?

- From the pretraining lecture:

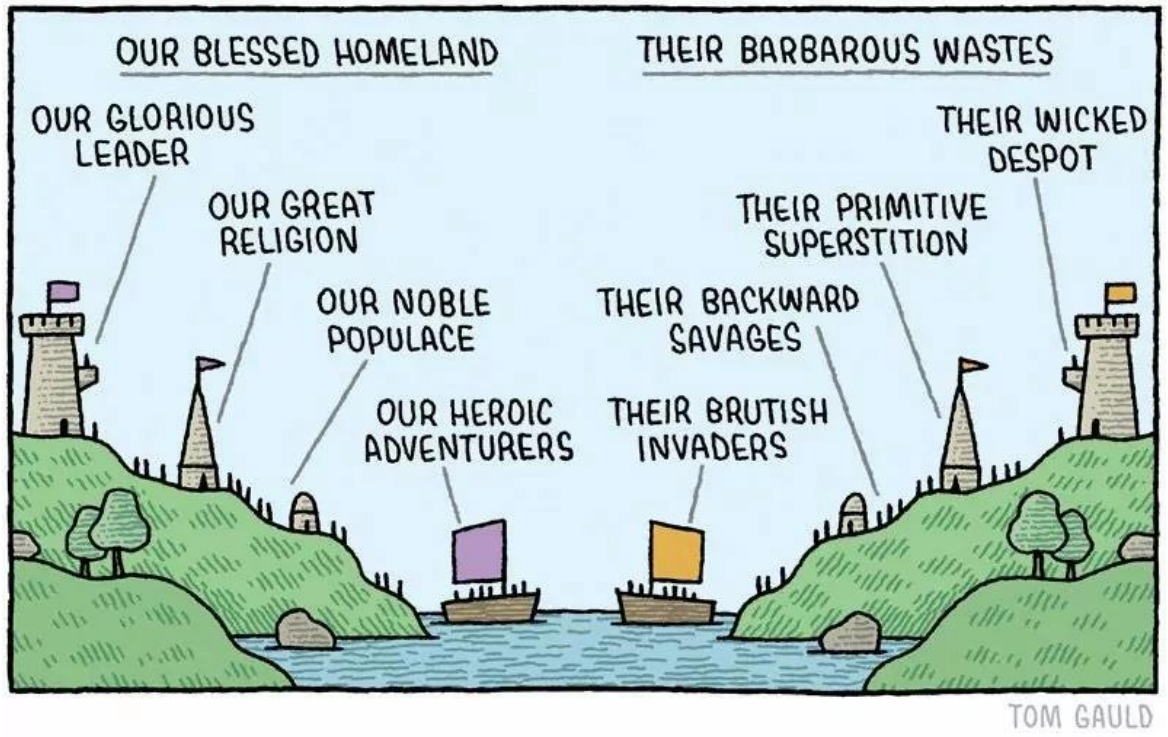
"You shall know a word by the company it keeps"
(J.R. Firth 1957: 11)

"…the complete meaning of a word is always contextual, and no study of meaning apart from a complete context can be taken seriously."
(J.R. Firth 1935)

What Do We Learn From Language *Use?*

- Language is **situated**

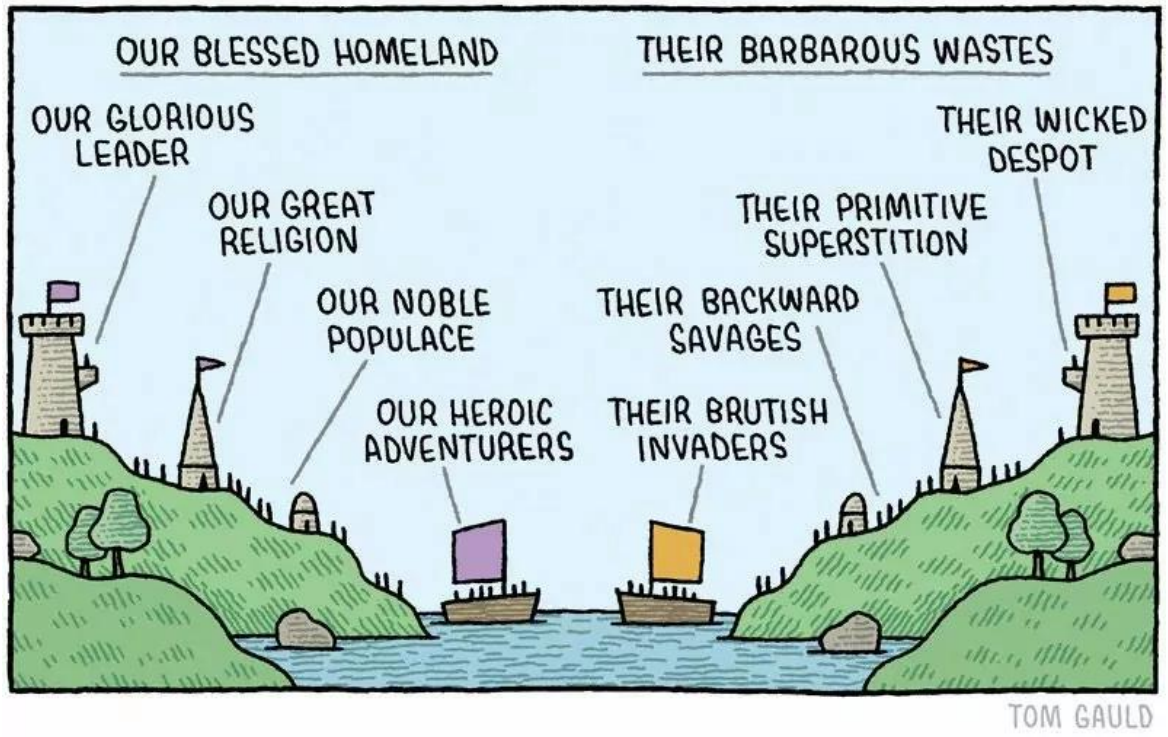
- Meaning in language is shaped by
  - Who says it
  - Who it is said to
  - Where/when it is said
  - Why it is said
  - …



What Do We Learn From Language *Use?*

- Language is **situated**

- Meaning in language is shaped by
  - Who says it
  - Who it is said to
  - Where/when it is said
  - Why it is said
  - …



What additional (and potentially undesirable) things do models learn from language?

Where We Are Heading

1. **How to study bias/fairness in language models**
   - *Useful taxonomies of bias/harm*
   - *Sources of bias*
2. **Historical Bias Research**
   - *Biases in language representations*
   - *Biases in language model applications/LLMs*
   - *Open problems in bias/fairness research*
3. **Privacy**
   - *Privacy considerations in NLP*
   - *Privacy risks in language technologies*

Where We Are Heading

1. **How to study bias/fairness in language models**
   - *Useful taxonomies of bias/harm*
   - *Sources of bias*
2. Historical Bias Research
   - *Biases in language representations*
   - *Biases in language model applications/LLMs*
   - *Open problems in bias/fairness research*
3. Privacy
   - *Privacy considerations in NLP*
   - *Privacy risks in language technologies*

AI Biases

- COMPAS system
  - Early finding of AI mimicking human biases

https://www.psychologytoday.com/us/blog/psyched/201801/law-enforcement-ai-is-no-more-or-less-biased-people

What Are Biases in NLP?

Bias/Fairness is Difficult to Define

- Bias/Fairness are typically broad terms

Language (Technology) is Power: A Critical Survey of “Bias” in NLP (Blodgett et al., 2020)

Bias/Fairness is Difficult to Define

- Bias/Fairness are typically broad terms
- Different people have different conceptions of *bias* or *fairness*

Language (Technology) is Power: A Critical Survey of “Bias” in NLP (Blodgett et al., 2020)

Bias/Fairness is Difficult to Define

- Bias/Fairness are typically broad terms
- Different people have different conceptions of *bias* or *fairness*
- What is considered fair/unfair is a *normative* concern
  - i.e., describes how something *ought* to be, not how it is

Language (Technology) is Power: A Critical Survey of “Bias” in NLP (Blodgett et al., 2020)

Bias/Fairness is Difficult to Define

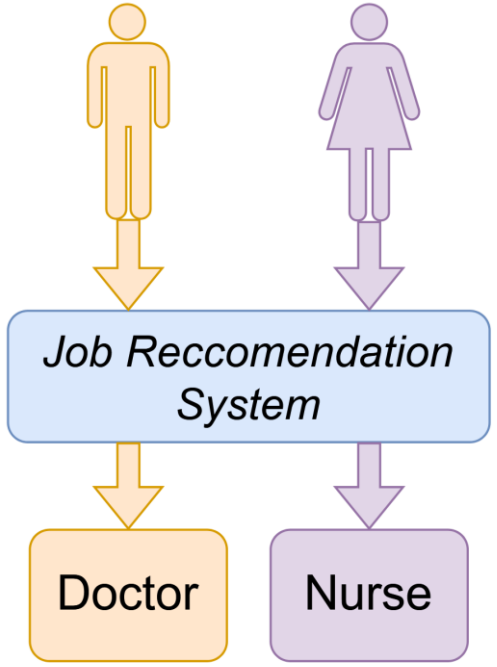
- Bias/Fairness are typically broad terms
- Different people have different conceptions of *bias* or *fairness*
- What is considered fair/unfair is a *normative* concern
  - i.e., describes how something *ought* to be, not how it is
- Need to consider bias in terms of **harm**
  - *What* is harmful, in *what ways*, *to whom*, and *why*

Language (Technology) is Power: A Critical Survey of "Bias" in NLP (Blodgett et al., 2020)

Taxonomy of Harm

Allocational Harms

Inequitable allocation of resources/opportunities

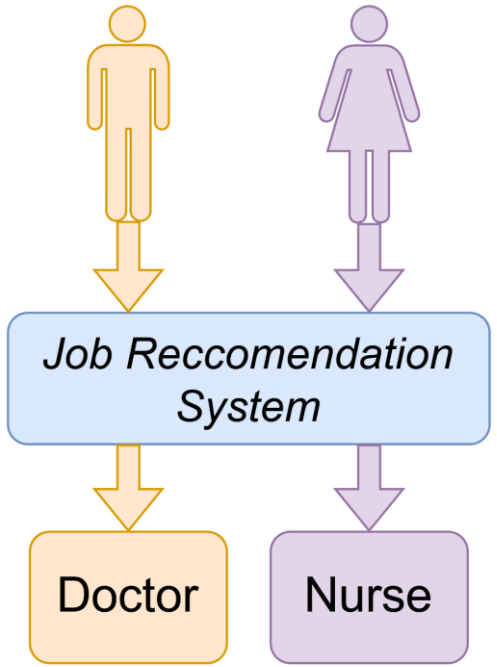


Language (Technology) is Power: A Critical Survey of "Bias" in NLP (Blodgett et al., 2020)
"The Problem With Bias: Allocative Versus Representational Harms in Machine Learning" (Barocas et al., 2017)

Taxonomy of Harm

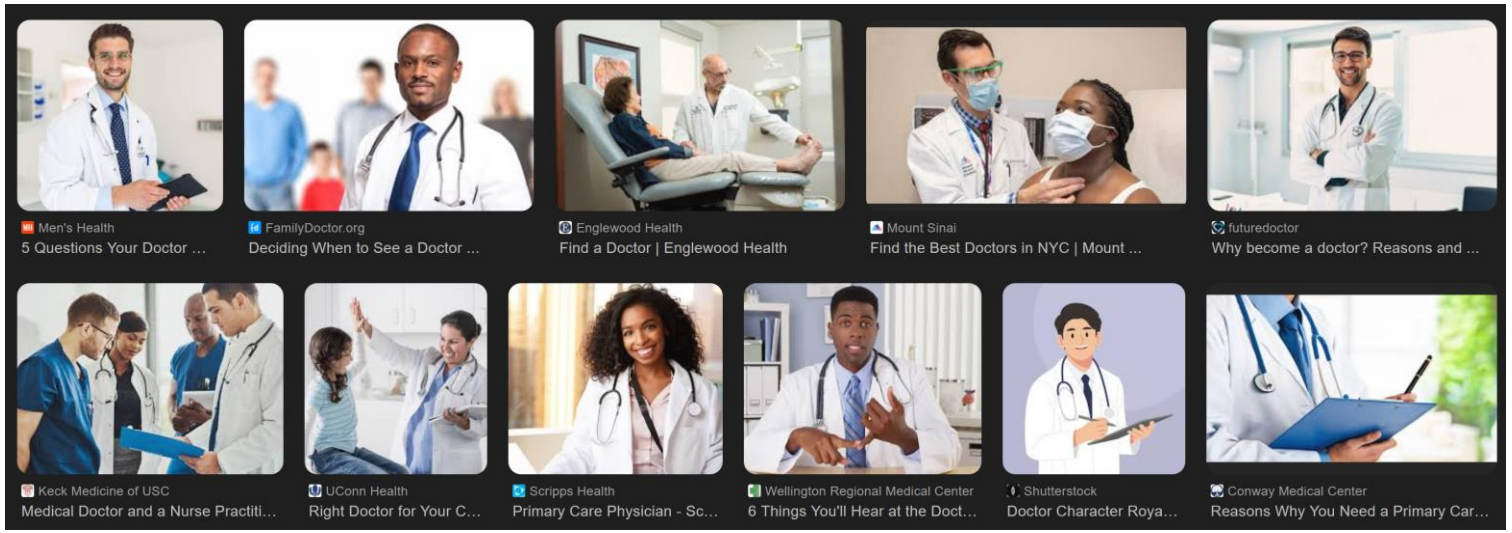
Allocational Harms

Inequitable allocation of resources/opportunities



Representational Harms

Representation of certain groups in less favorable light than others



Language (Technology) is Power: A Critical Survey of "Bias" in NLP (Blodgett et al., 2020)

"The Problem With Bias: Allocative Versus Representational Harms in Machine Learning" (Barocas et al., 2017)

Allocational Harms

- Usually concerns behavior/performance disparities between groups

$$d_M(M(x_a), M(x_b)) \leq L d_x(x_a, x_b)$$

Language (Technology) is Power: A Critical Survey of “Bias” in NLP (Blodgett et al., 2020)

Allocational Harms

- Usually concerns behavior/performance disparities between groups

Difference in model outputs on x_a and x_b

Scaled difference between x_a and x_b

$$d_M(M(x_a), M(x_b)) \leq L d_x(x_a, x_b)$$

Language (Technology) is Power: A Critical Survey of "Bias" in NLP (Blodgett et al., 2020)

Allocational Harms

- Usually concerns behavior/performance disparities between groups

Difference in model outputs/performance on x_a and x_b

Scaled difference between x_a and x_b

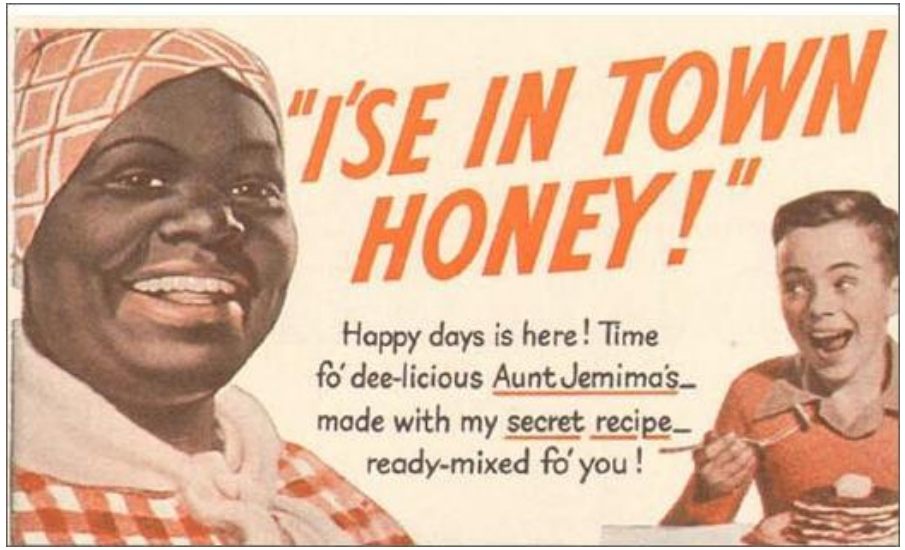
$$d_M(M(x_a), M(x_b)) \leq L d_x(x_a, x_b)$$

$$|f(M(x_a), y_a) - f(M(x_b), y_b)| \leq L d_x(x_a, x_b)$$

Language (Technology) is Power: A Critical Survey of "Bias" in NLP (Blodgett et al., 2020)

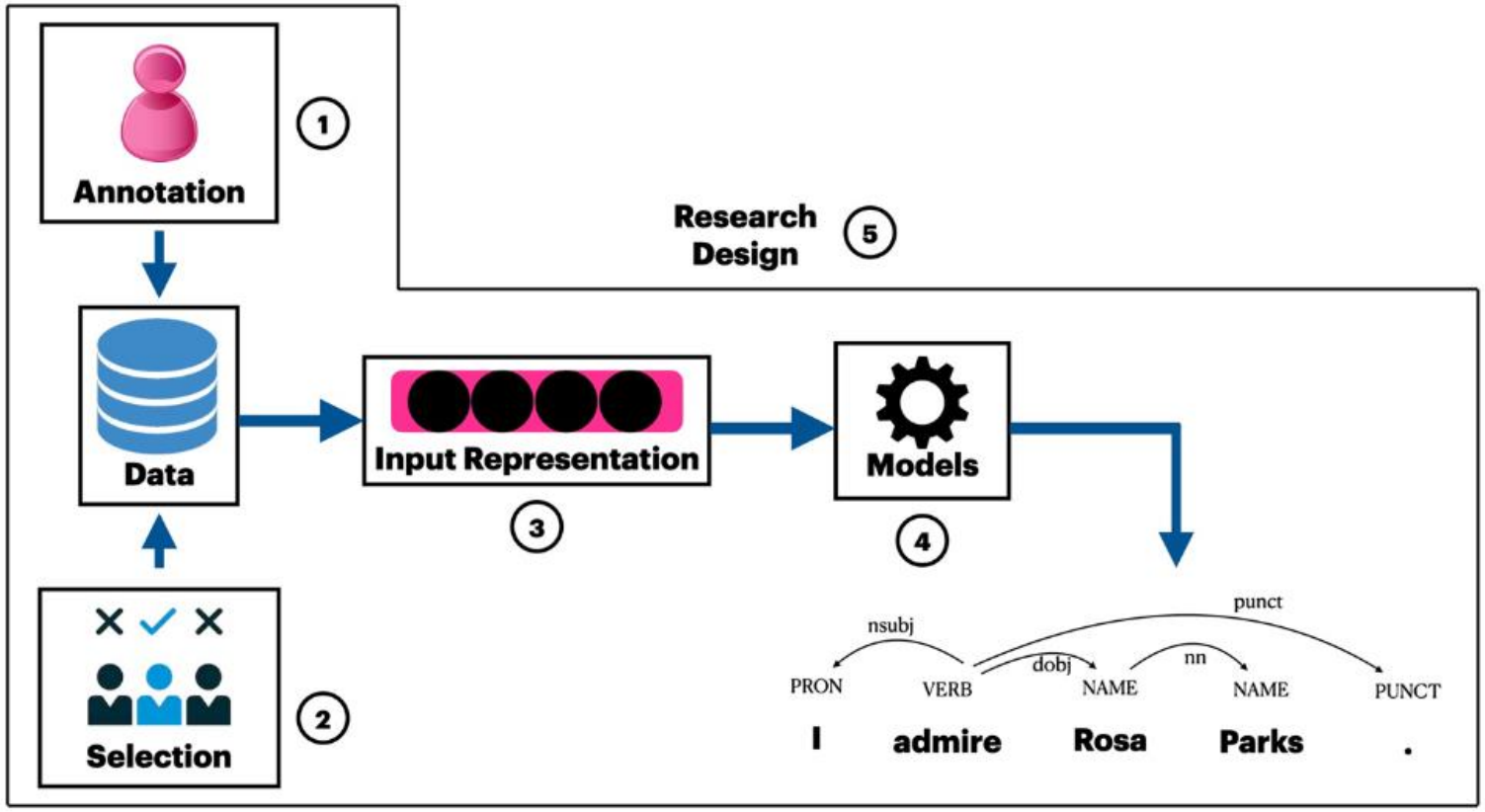
Representational Harms

- Difficult to formalize
- Typically takes the form of *stereotypes*
  - Association between groups and concepts
  - Qualitative observations about representation



Language (Technology) is Power: A Critical Survey of "Bias" in NLP (Blodgett et al., 2020)

5 Sources of Bias in NLP



Five sources of bias in natural language processing (Hovy & Prabhumoye, 2020)

Where We Are Heading

1. How to study bias/fairness in language models
   - *Useful taxonomies of bias/harm*
   - *Sources of bias*
2. **Historical Bias Research**
   - *Biases in language representations*
   - *Biases in language model applications/LLMs*
   - *Open problems in bias/fairness research*
3. Privacy
   - *Privacy considerations in NLP*
   - *Privacy risks in language technologies*

How Do We Research Biases/Fairness Issues?

1. **What evaluation methods should we use?**
   - *Apply human-centered frameworks to embeddings and language models*

2. **How do we consider specific social groups?**
   - *Use proxies of groups (e.g., dialect, names, identity mentions)*

3. **How do we mitigate these biases?**
   - *Rely on the evaluation approach*

Historical Research on Bias/Fairness

- Bias was a major topic only relatively recently (~2016)

NLP task	Papers
Embeddings (type-level or contextualized)	54
Coreference resolution	20
Language modeling or dialogue generation	17
Hate-speech detection	17
Sentiment analysis	15
Machine translation	8
Tagging or parsing	5
Surveys, frameworks, and meta-analyses	20
Other	22

Language (Technology) is Power: A Critical Survey of "Bias" in NLP (Blodgett et al., 2020)

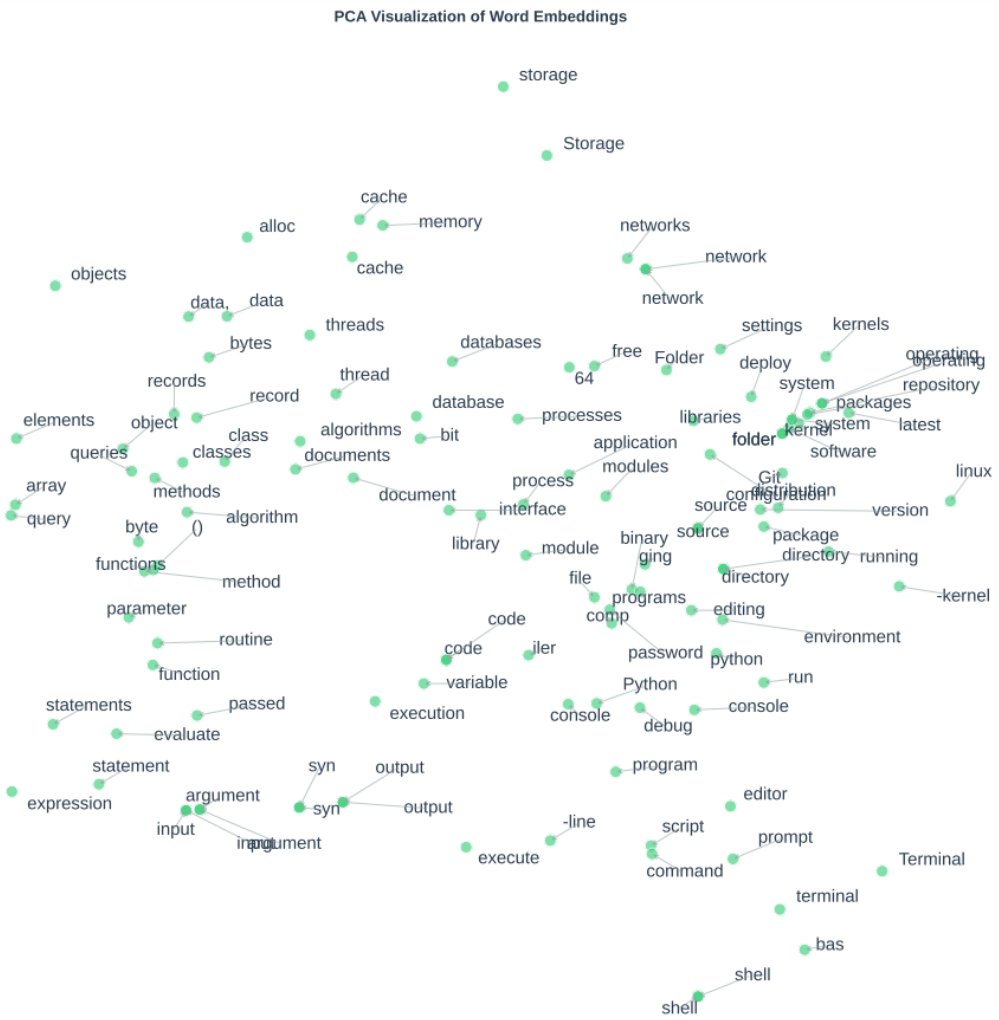
Historical Research on Bias/Fairness

- Bias was a major topic only relatively recently (~2016)

NLP task	Papers
Embeddings (type-level or contextualized)	54
Coreference resolution	20
Language modeling or dialogue generation	17
Hate-speech detection	17
Sentiment analysis	15
Machine translation	8
Tagging or parsing	5
Surveys, frameworks, and meta-analyses	20
Other	22

Language (Technology) is Power: A Critical Survey of "Bias" in NLP (Blodgett et al., 2020)

Early Biases in Embeddings

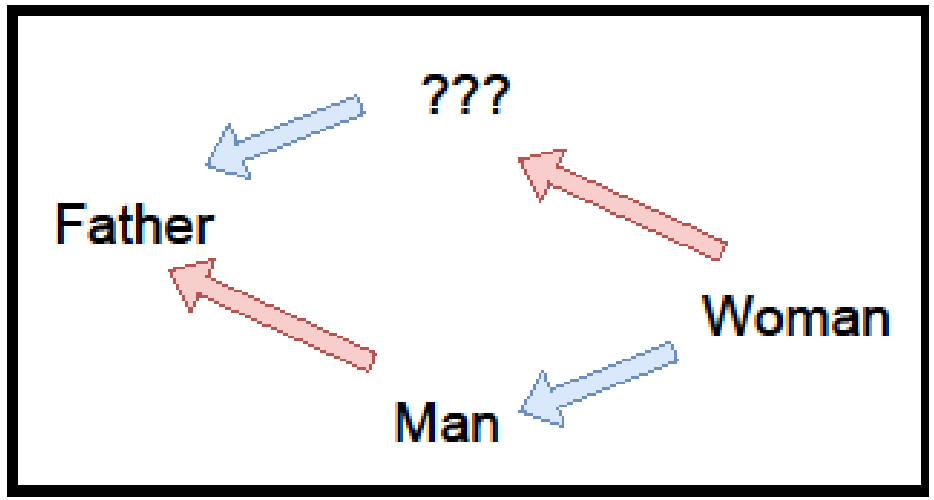


Early Biases in Embeddings - Analogies

- Word embeddings encode analogy-like relationships

$$(\overrightarrow{woman} - \overrightarrow{man}) + \overrightarrow{father} \approx \; ???$$

Man : Father :: Woman : ???



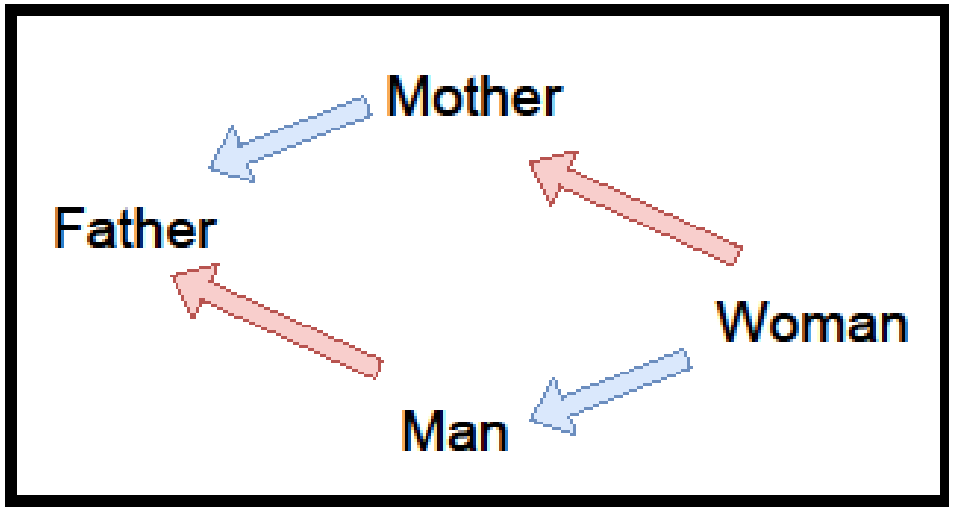
"Man is to computer programmer as woman is to homemaker? debiasing word embeddings" (Bolukbasi et al., 2016)

Early Biases in Embeddings – Analogies

- Word embeddings encode analogy-like relationships

$$(\overrightarrow{woman} - \overrightarrow{man}) + \overrightarrow{father} \approx \overrightarrow{mother}$$

Man : Father :: Woman : Mother



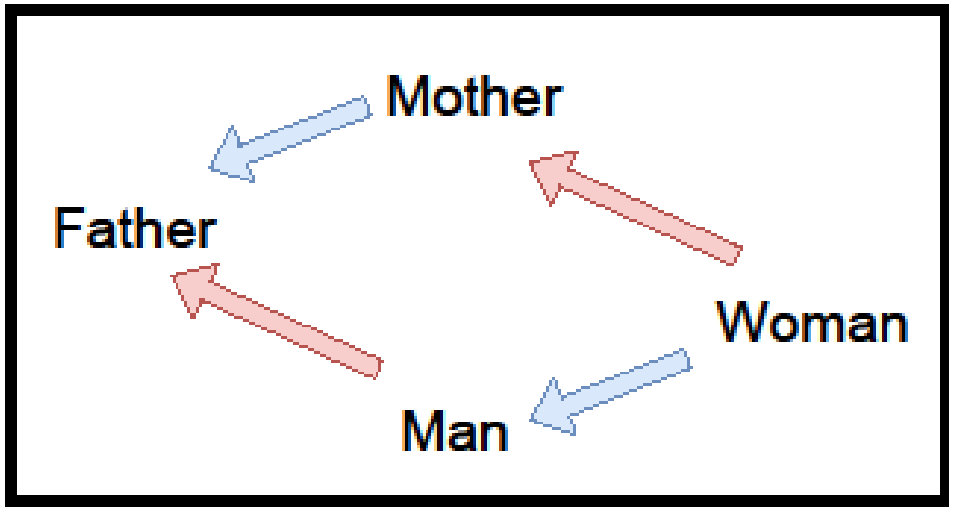
"Man is to computer programmer as woman is to homemaker? debiasing word embeddings" (Bolukbasi et al., 2016)

Early Biases in Embeddings – Analogies

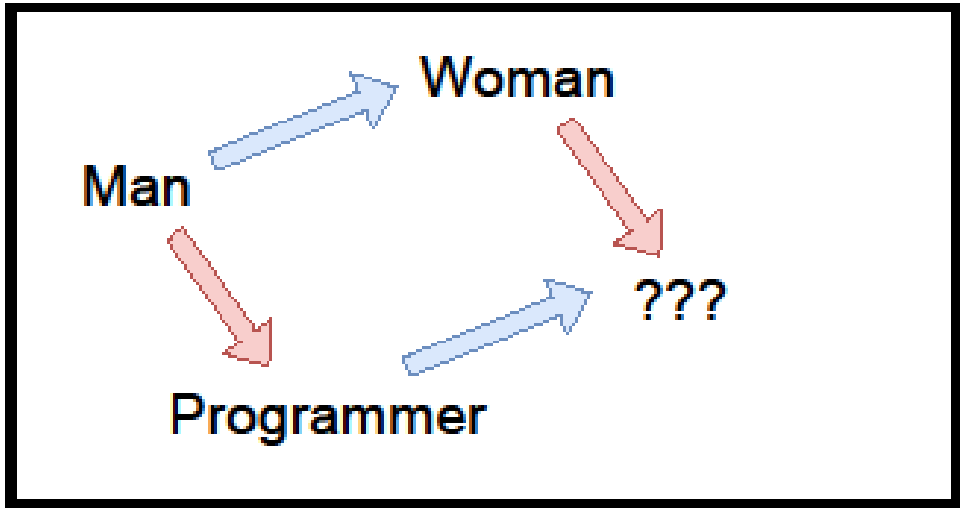
- Word embeddings encode analogy-like relationships

$$(\overrightarrow{woman} - \overrightarrow{man}) + \overrightarrow{programmer} \approx \; ???$$

Man : Father :: Woman : Mother



Man : Programmer :: Woman : ???



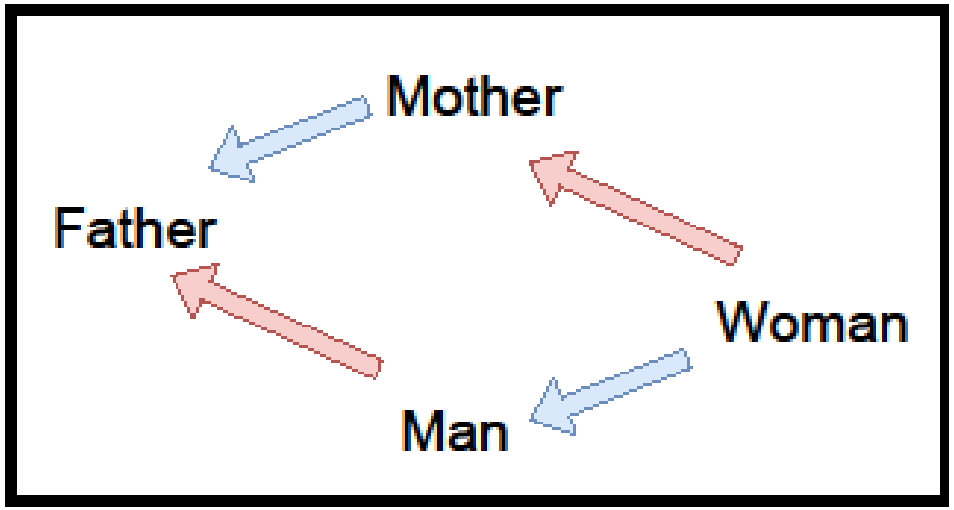
"Man is to computer programmer as woman is to homemaker? debiasing word embeddings" (Bolukbasi et al., 2016)

Early Biases in Embeddings – Analogies

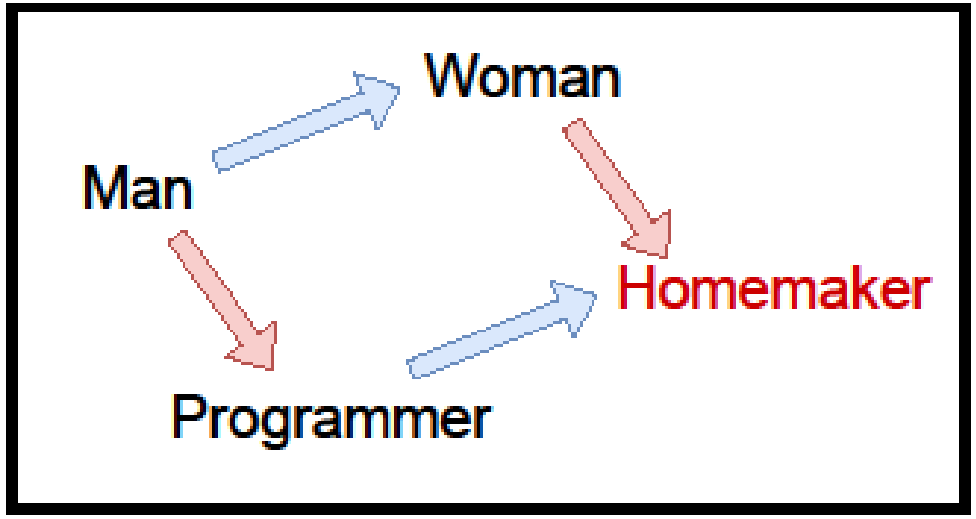
- Word embeddings encode analogy-like relationships

$$(\overrightarrow{woman} - \overrightarrow{man}) + \overrightarrow{programmer} \approx \overrightarrow{homemaker}$$

Man : Father :: Woman : Mother



Man : Programmer :: Woman : Homemaker



"Man is to computer programmer as woman is to homemaker? debiasing word embeddings" (Bolukbasi et al., 2016)

Early Biases in Embeddings – Analogies

Gender stereotype *she-he* analogies.

sewing-carpentry	register-nurse-physician	housewife-shopkeeper
nurse-surgeon	interior designer-architect	softball-baseball
blond-burly	feminism-conservatism	cosmetics-pharmaceuticals
giggle-chuckle	vocalist-guitarist	petite-lanky
sassy-snappy	diva-superstar	charming-affable
volleyball-football	cupcakes-pizzas	hairdresser-barber

Gender appropriate *she-he* analogies.

queen-king	sister-brother	mother-father
waitress-waiter	ovarian cancer-prostate cancer	convent-monastery

"Man is to computer programmer as woman is to homemaker? debiasing word embeddings" (Bolukbasi et al., 2016)

Early Biases in Embeddings – Analogies

Gender stereotype *she-he* analogies.

sewing-carpentry	register-nurse-physician	housewife-shopkeeper
nurse-surgeon	interior designer-architect	softball-baseball
blond-burly	feminism-conservatism	cosmetics-pharmaceuticals
giggle-chuckle	vocalist-guitarist	petite-lanky
sassy-snappy	diva-superstar	charming-affable
volleyball-football	cupcakes-pizzas	hairdresser-barber

"Man is to computer programmer as woman is to homemaker? debiasing word embeddings" (Bolukbasi et al., 2016)

Early Biases in Embeddings – Analogies

Gender stereotype *she-he* analogies.

sewing-carpentry	register-nurse-physician	housewife-shopkeeper
nurse-surgeon	interior designer-architect	softball-baseball
blond-burly	feminism-conservatism	cosmetics-pharmaceuticals
giggle-chuckle	vocalist-guitarist	petite-lanky
sassy-snappy	diva-superstar	charming-affable
volleyball-football	cupcakes-pizzas	hairdresser-barber

Gender appropriate *she-he* analogies.

queen-king	sister-brother	mother-father
waitress-waiter	ovarian cancer-prostate cancer	convent-monastery

"Man is to computer programmer as woman is to homemaker? debiasing word embeddings" (Bolukbasi et al., 2016)

Early Biases in Embeddings – Analogies

Gender stereotype *she-he* analogies.

sewing-carpentry	register-nurse-physician	housewife-shopkeeper
nurse-surgeon	interior designer-architect	softball-baseball
blond-burly	feminism-conservatism	cosmetics-pharmaceuticals
giggle-chuckle	vocalist-guitarist	petite-lanky
sassy-snappy	diva-superstar	charming-affable
volleyball-football	cupcakes-pizzas	hairdresser-barber

Harmful

Gender appropriate *she-he* analogies.

queen-king	sister-brother	mother-father
waitress-waiter	ovarian cancer-prostate cancer	convent-monastery

Not Harmful

"Man is to computer programmer as woman is to homemaker? debiasing word embeddings" (Bolukbasi et al., 2016)

Implicit Association Test

Evaluate biases through implicit associations between social groups and attributes

https://implicit.harvard.edu/implicit/takeatest.html

Semantics derived automatically from language corpora contain human-like biases (Caliskan et al., 2017)

Implicit Association Test

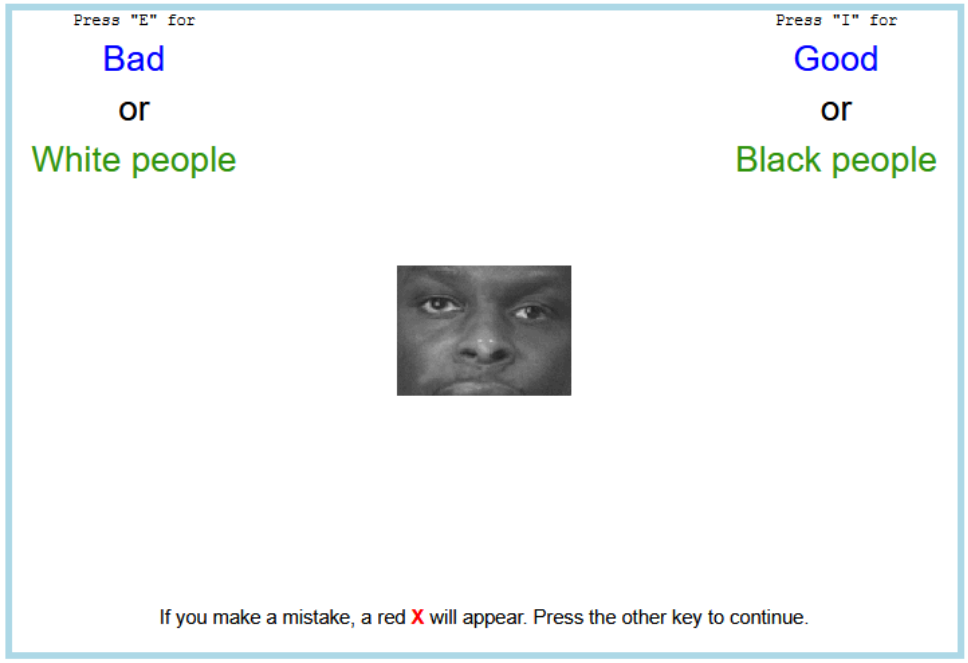
Evaluate biases through implicit associations between social groups and attributes

Category	Items
Good	Spectacular, Joyful, Cheerful, Love, Fantastic, Pleasing, Beautiful, Friendship
Bad	Grief, Hate, Annoy, Hurtful, Tragic, Detest, Disaster, Hatred
Black people	
White people	

https://implicit.harvard.edu/implicit/takeatest.html

Semantics derived automatically from language corpora contain human-like biases (Caliskan et al., 2017)

Implicit Association Test



https://implicit.harvard.edu/implicit/takeatest.html

Semantics derived automatically from language corpora contain human-like biases (Caliskan et al., 2017)

Implicit Associations in Word Embeddings

- *Word Embedding Association Test* (WEAT)

Test	Target Examples	Attribute Examples
WEAT	*European American names*: Adam, Harry, Nancy, ...	*Pleasant*: love, cheer, miracle, ...
	African American names: Jamal, Lavar, Latisha, ...	*Unpleasant*: ugly, evil, abuse, ...

Semantics derived automatically from language corpora contain human-like biases (Caliskan et al., 2017)

Implicit Associations in Word Embeddings

- *Word Embedding Association Test* (WEAT)

Targets: X,Y

- E.g., race, gender

Attributes: A,B

- E.g., pleasant/ unpleasant

How much is w associated with A compared to B?

$$s(w, A, B) = \frac{1}{|A|} \sum_{\mathbf{a} \in \mathbf{A}} \cos(\mathbf{e_w}, \mathbf{a}) - \frac{1}{|\mathbf{B}|} \sum_{\mathbf{b} \in \mathbf{B}} \cos(\mathbf{e_w}, \mathbf{b})$$

Semantics derived automatically from language corpora contain human-like biases (Caliskan et al., 2017)

Implicit Associations in Word Embeddings

- *Word Embedding Association Test* (WEAT)

Targets: X,Y

- E.g., race, gender

Attributes: A,B

- E.g., pleasant/ unpleasant

How much is w associated with A compared to B?

$$s(w, A, B) = \frac{1}{|A|} \sum_{\mathbf{a} \in \mathbf{A}} \cos(\mathbf{e_w}, \mathbf{a}) - \frac{1}{|\mathbf{B}|} \sum_{\mathbf{b} \in \mathbf{B}} \cos(\mathbf{e_w}, \mathbf{b})$$

Semantics derived automatically from language corpora contain human-like biases (Caliskan et al., 2017)

Implicit Associations in Word Embeddings

- *Word Embedding Association Test* (WEAT)

Targets: X,Y

- E.g., race, gender

Attributes: A,B

- E.g., pleasant/ unpleasant

How much is w associated with A compared to B?

$$s(w, A, B) = \frac{1}{|A|} \sum_{\mathbf{a} \in \mathbf{A}} \cos(\mathbf{e_w}, \mathbf{a}) - \frac{1}{|\mathbf{B}|} \sum_{\mathbf{b} \in \mathbf{B}} \cos(\mathbf{e_w}, \mathbf{b})$$

How aligned are X↔Y and A↔B?

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

Semantics derived automatically from language corpora contain human-like biases (Caliskan et al., 2017)

Implicit Associations in Word Embeddings

- *Word Embedding Association Test* (WEAT)

Targets: X,Y

- E.g., race, gender

Attributes: A,B

- E.g., pleasant/ unpleasant

How much is w associated with A compared to B?

$$s(w, A, B) = \frac{1}{|A|} \sum_{\mathbf{a} \in \mathbf{A}} \cos(\mathbf{e_w}, \mathbf{a}) - \frac{1}{|\mathbf{B}|} \sum_{\mathbf{b} \in \mathbf{B}} \cos(\mathbf{e_w}, \mathbf{b})$$

How aligned are X↔Y and A↔B?

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

Semantics derived automatically from language corpora contain human-like biases (Caliskan et al., 2017)

Implicit Associations in Contextual Embeddings

2 approaches to adapting WEAT

1. Consider *neutral* sentences rather than tokens/words

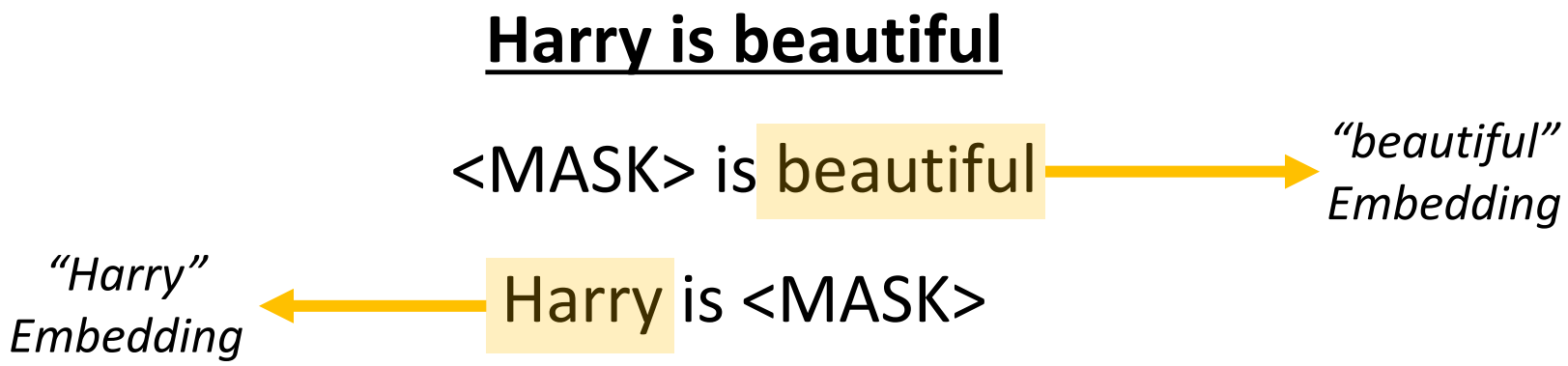
Test	Target Examples	Attribute Examples
WEAT	*European American names*: Adam, Harry, Nancy, ...	*Pleasant*: love, cheer, miracle, ...
	African American names: Jamal, Lavar, Latisha, ...	*Unpleasant*: ugly, evil, abuse, ...
SEAT	*European American names*: "This is Adam", "There is Harry",...	*Pleasant*: "There is love", "That is a miracle", ...
	African American names: "Jamal is here", "Latisha is a person", ...	*Unpleasant*: "This is ugly", "They are evil", ...

On measuring social biases in sentence encoders (May et al., 2019)
Measuring bias in contextualized word representations (Kurita et al., 2019)

Implicit Associations in Contextual Embeddings

2 approaches to adapting WEAT

1. Consider *neutral* sentences rather than tokens/words
2. Mask *targets* and *attributes*



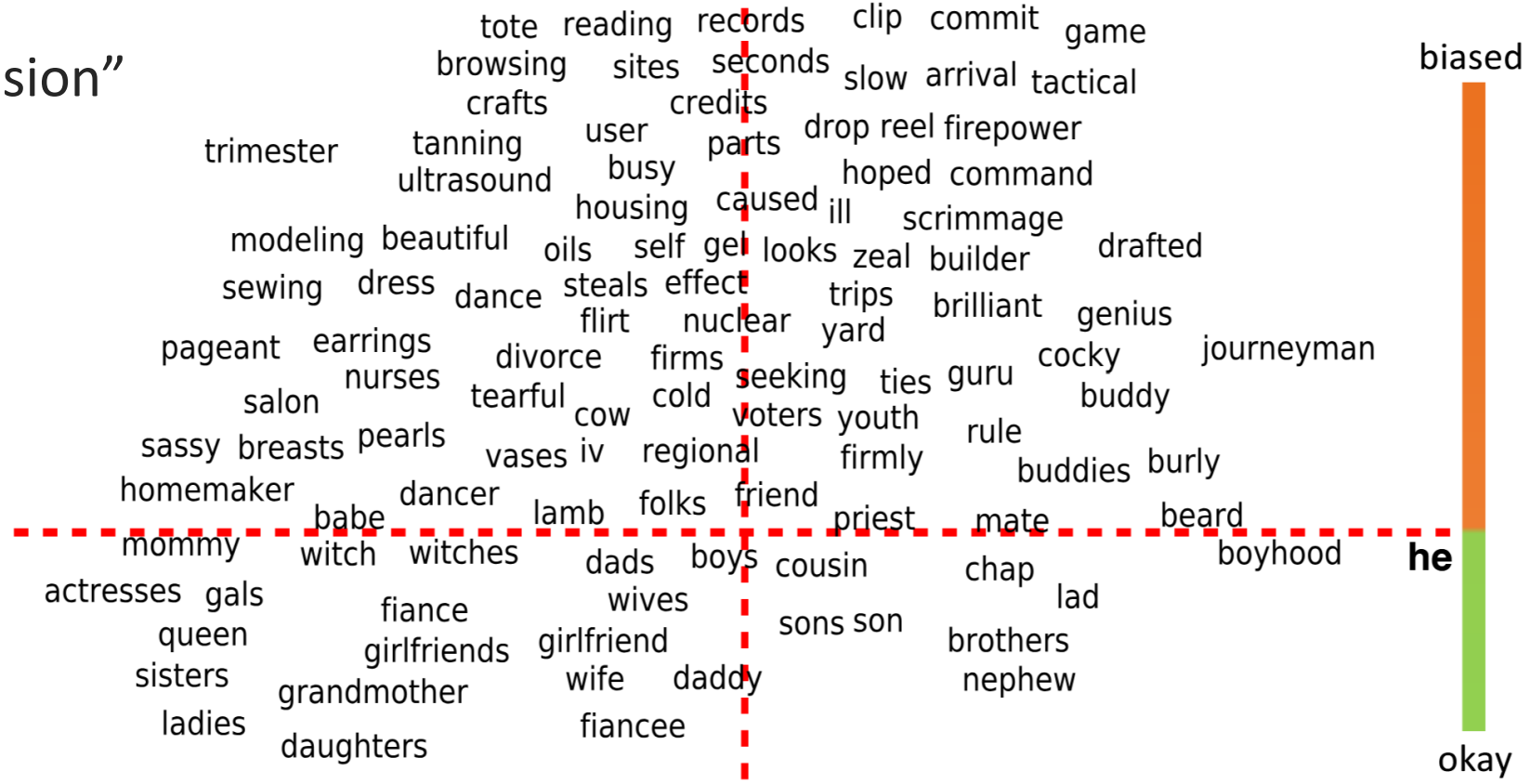
On measuring social biases in sentence encoders (May et al., 2019)
Measuring bias in contextualized word representations (Kurita et al., 2019)

Mitigating Biased Associations

Man is to computer programmer as woman is to homemaker? debiasing word embeddings (Bolukbasi et al., 2016)

Mitigating Biased Associations

- Main idea
  1. Identify bias “dimension”
  2. Identify biased and neutral word sets



Man is to computer programmer as woman is to homemaker? debiasing word embeddings (Bolukbasi et al., 2016)

Mitigating Biased Associations

- Main idea
  1. Identify bias “dimension”
  2. Identify biased and neutral word sets
  3. Neutralize biased words, maintain neutral words

Man is to computer programmer as woman is to homemaker? debiasing word embeddings (Bolukbasi et al., 2016)

How Do We Research Biases/Fairness Issues?

1. **What evaluation methods should we use?**
   - *Apply human-centered frameworks to embeddings and language models*

2. **How do we consider specific social groups?**
   - *Use proxies of groups (e.g., dialect, names, identity mentions)*

3. **How do we mitigate these biases?**
   - *Rely on the evaluation approach*

Historical Research on Bias/Fairness

- Bias was a major topic only relatively recently (~2016)

NLP task	Papers
Embeddings (type-level or contextualized)	54
Coreference resolution	20
Language modeling or dialogue generation	17
Hate-speech detection	17
Sentiment analysis	15
Machine translation	8
Tagging or parsing	5
Surveys, frameworks, and meta-analyses	20
Other	22

Language (Technology) is Power: A Critical Survey of “Bias” in NLP (Blodgett et al., 2020)

Historical Research on Bias/Fairness

- Bias was a major topic only relatively recently (~2016)

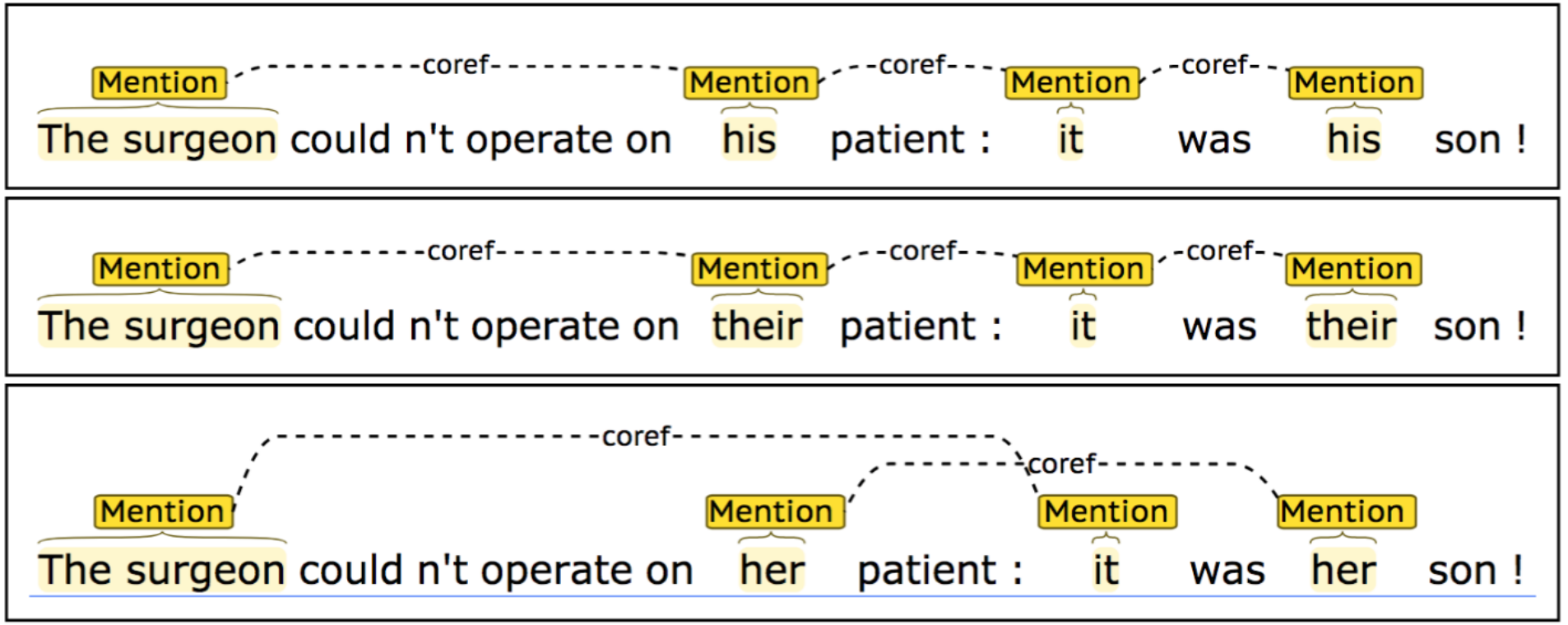
NLP task	Papers
Embeddings (type-level or contextualized)	54
Coreference resolution	20
Language modeling or dialogue generation	17
Hate-speech detection	17
Sentiment analysis	15
Machine translation	8
Tagging or parsing	5
Surveys, frameworks, and meta-analyses	20
Other	22

Language (Technology) is Power: A Critical Survey of "Bias" in NLP (Blodgett et al., 2020)

Bias in Coreference Resolution

Coreference resolution: Identify all mentions of a targeted entity

- Gender occupation biases resurface

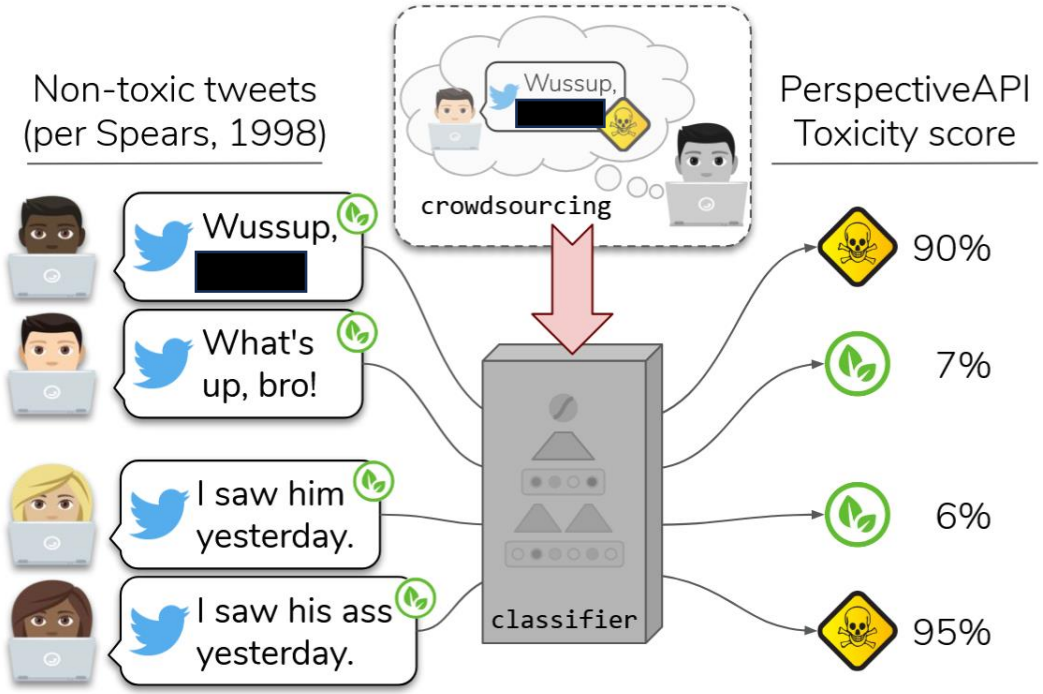


Gender Bias in Coreference Resolution (Rudinger et al., 2018)

Bias in Hate Speech Detection

Hate Speech/Toxicity Detection

- Models over-label African American speech, under-label Black hate speech



The risk of racial bias in hate speech detection (Sap et al., 2019)
Racial bias in hate speech and abusive language detection datasets (2019)

Historical Research on Bias/Fairness

- Bias was a major topic only relatively recently (~2016)

Where lots of current work focuses

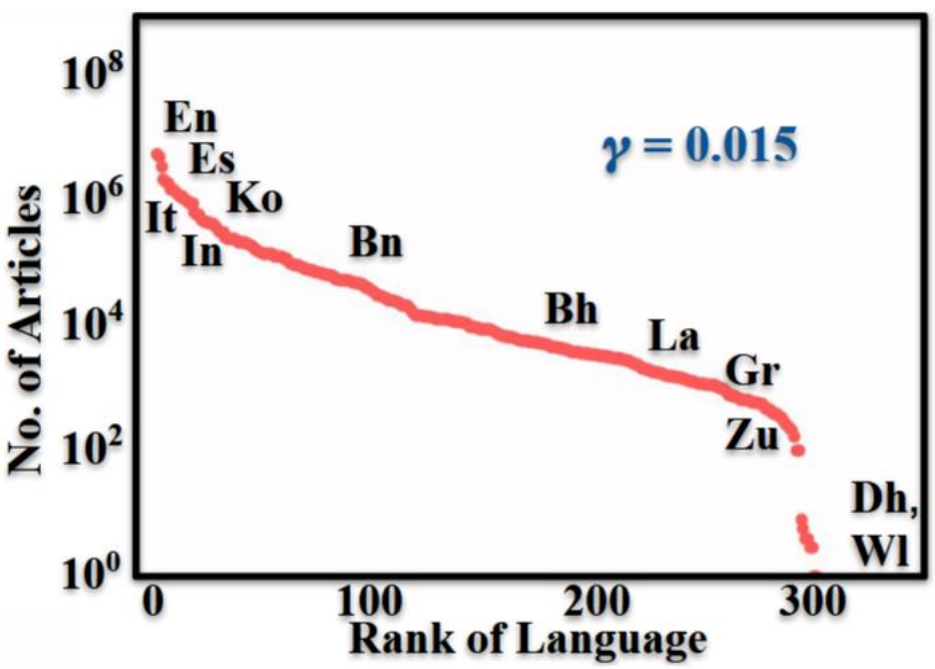
NLP task	Papers
Embeddings (type-level or contextualized)	54
Coreference resolution	20
Language modeling or dialogue generation	17
Hate-speech detection	17
Sentiment analysis	15
Machine translation	8
Tagging or parsing	5
Surveys, frameworks, and meta-analyses	20
Other	22

Language (Technology) is Power: A Critical Survey of "Bias" in NLP (Blodgett et al., 2020)

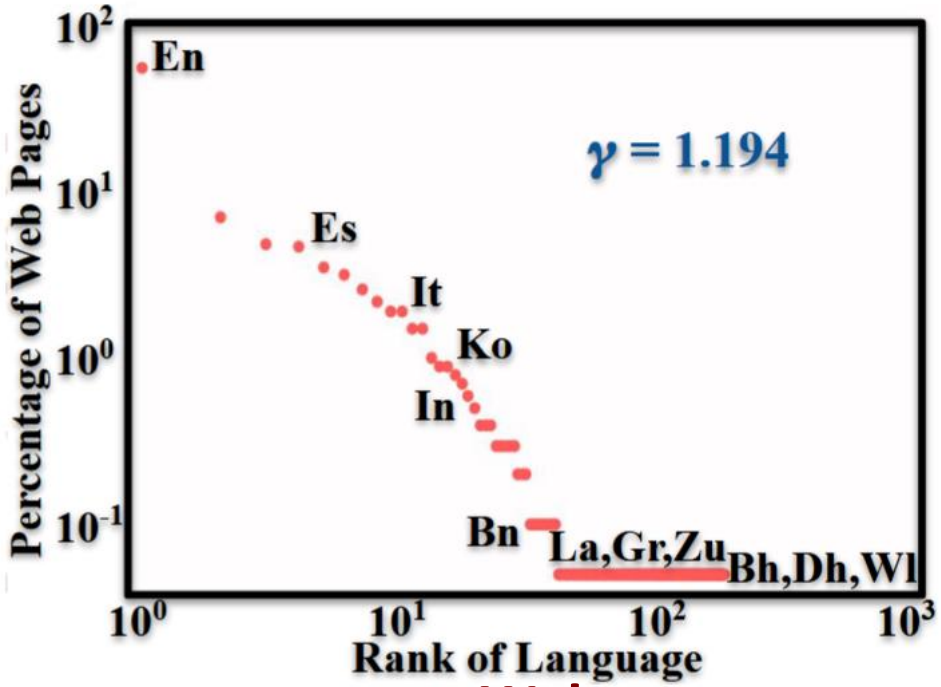
Language Disparities

Language Disparities

- **Resource disparity**
  - Language resources dominated by a few languages/groups (i.e., Indo-European)



Wikipedia

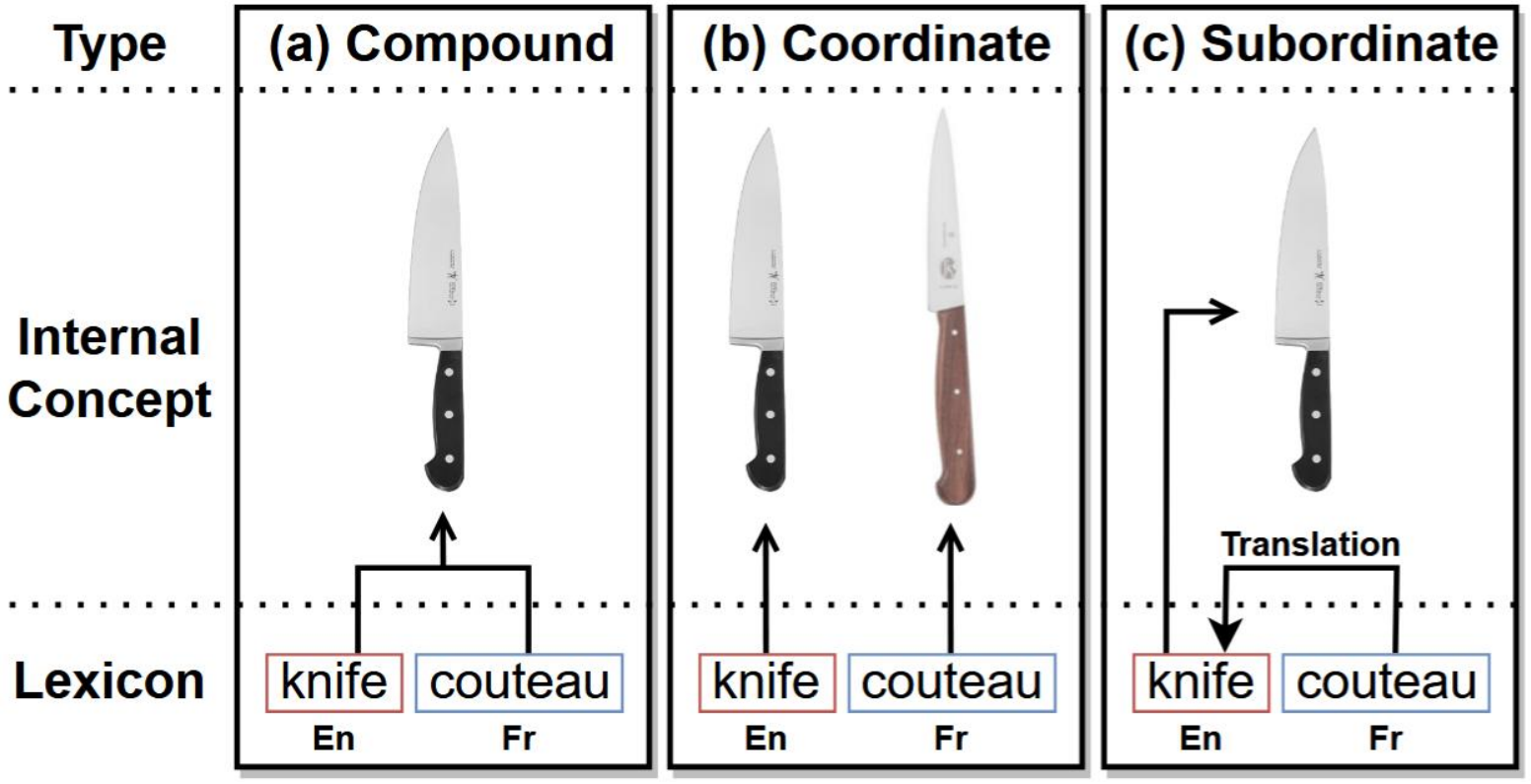


Web

The State and Fate of Linguistic Diversity and Inclusion in the NLP World (Joshi et al., 2020)

Language Disparities

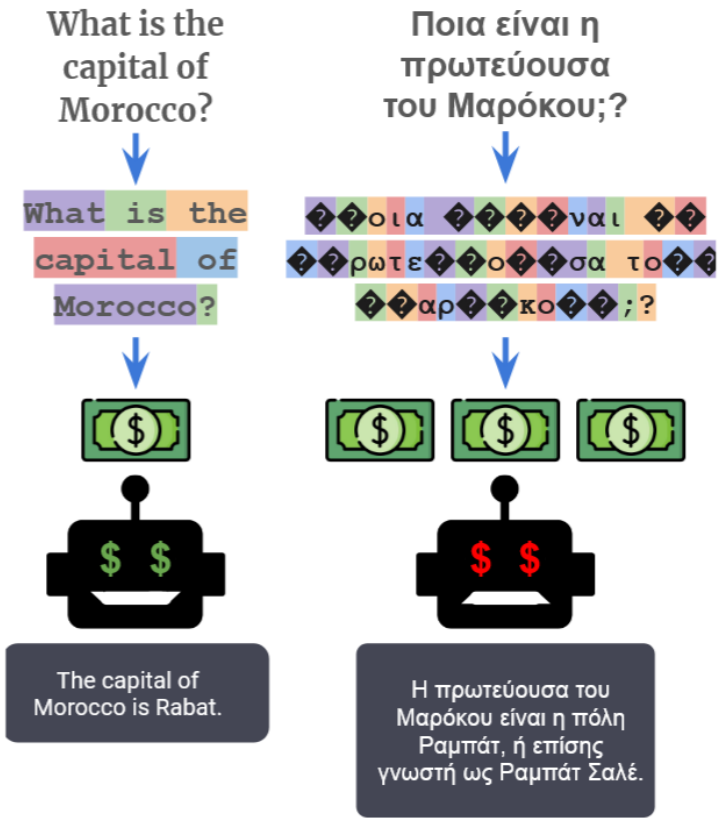
- Resource disparity
- **Performance disparity**
  - Models rely on English as a “native” language



Don’t Trust ChatGPT when your Question is not in English (Zhang et al., 2023)

Language Disparities

- Resource disparity
- Performance disparity
- **Cost disparity**
  - Underrepresented language speakers spend *more* for *worse* performance



Do All Languages Cost the Same? Tokenization in the Era of Commercial Language Models (Ahia et al., 2023)

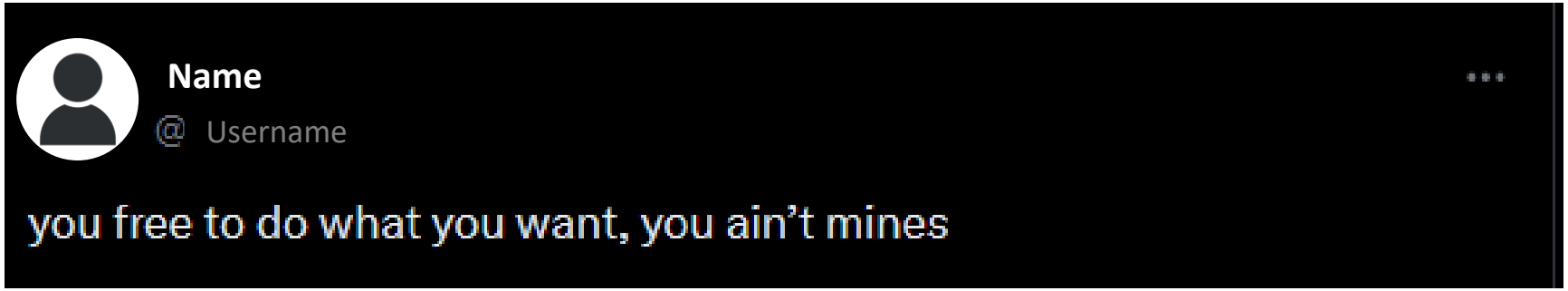
Language *Variety* Disparities

- Speakers of the same language do not all speak the same
  - Variation associated with ethnicity, region, socioeconomic class, …

Language *Variety* Disparities

- Speakers of the same language do not all speak the same
  - Variation associated with ethnicity, region, socioeconomic class, …

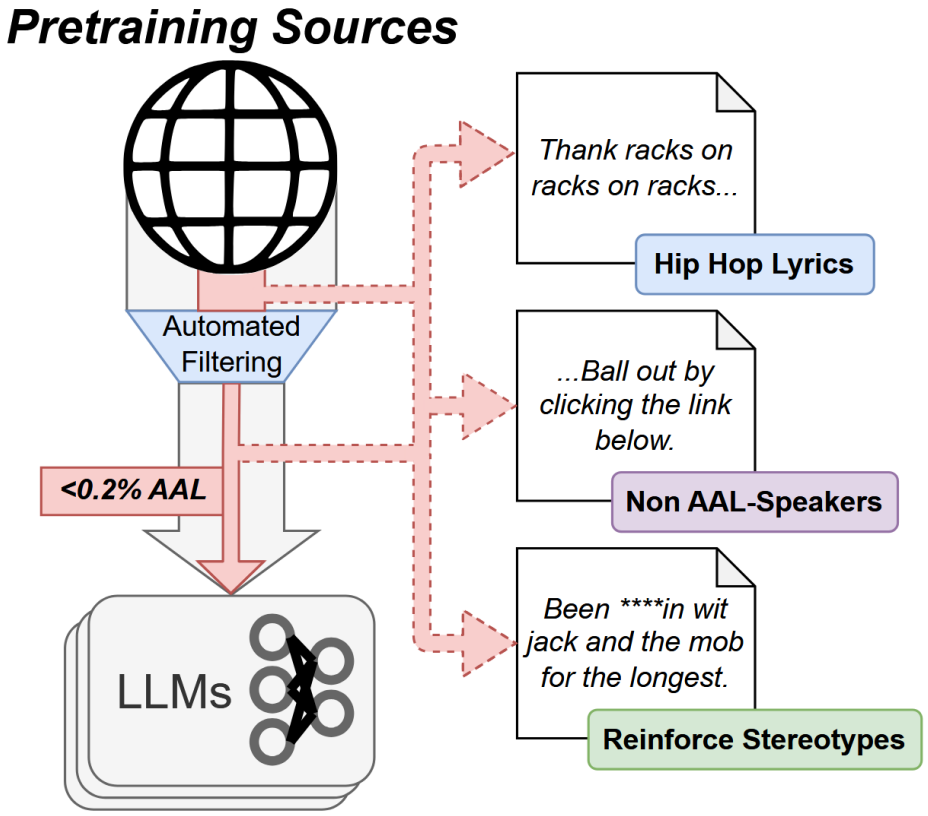
- **African American Language**: Variety of English associated primarily with African Americans in the US
- **White Mainstream English (WME):** Variety of English encompassing the linguistic norms of White Americans



Language *Variety* Disparities

- AAL more likely to be labeled toxic/hate speech
- LLMs struggle to understand or respond well to AAL
- **AAL misrepresented in pretraining corpora**
- Reward models encourage WME over AAL

...



The risk of racial bias in hate speech detection (Sap et al., 2019)
Linguistic bias in ChatGPT: language models reinforce dialect discrimination (Fleisig et al., 2024)
Data caricatures: on the representation of African American Language in pretraining corpora (Deas et al., 2025)
Evaluation of African American Language bias in natural language generation (Deas et al., 2023)
Rejected dialects: biases against African American Language in reward models (Mire et al., 2025)

Political and Cultural Biases

- Survey-based approaches for **A**ttitudes, **O**pinions, and **V**alues (AOV)

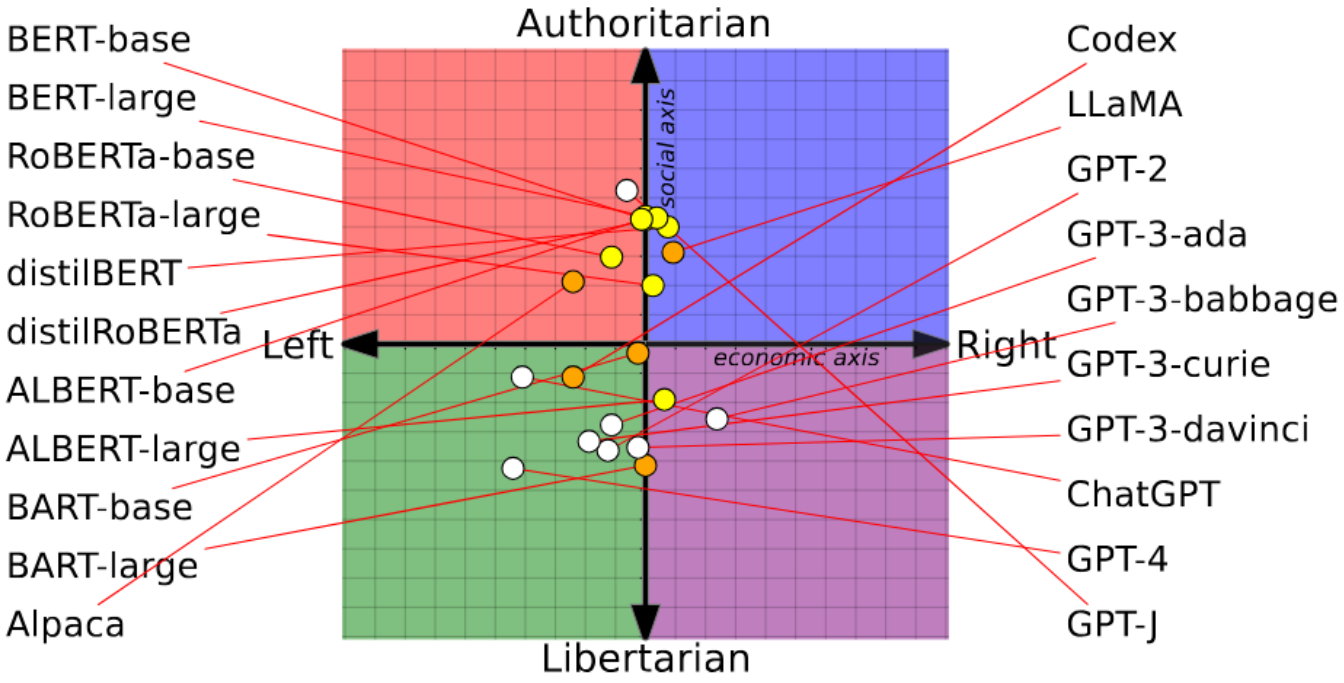
Source: WVS
Question: Do you agree, disagree or neither agree nor disagree with the following statement? *"When jobs are scarce, employers should give priority to people of this country over immigrants."*

(A) Agree strongly
(B) Agree
(C) Neither agree nor disagree
(D) Disagree
(E) Disagree strongly
(F) Don't know

Whose opinions do language models reflect? (Santurkar et al., 2023)
Towards measuring the representation of subjective global opinions in language models (Durmus et al., 2024)

Political and Cultural Biases

- Survey-based approaches for **A**ttitudes, **O**pinions, and **V**alues (AOV)



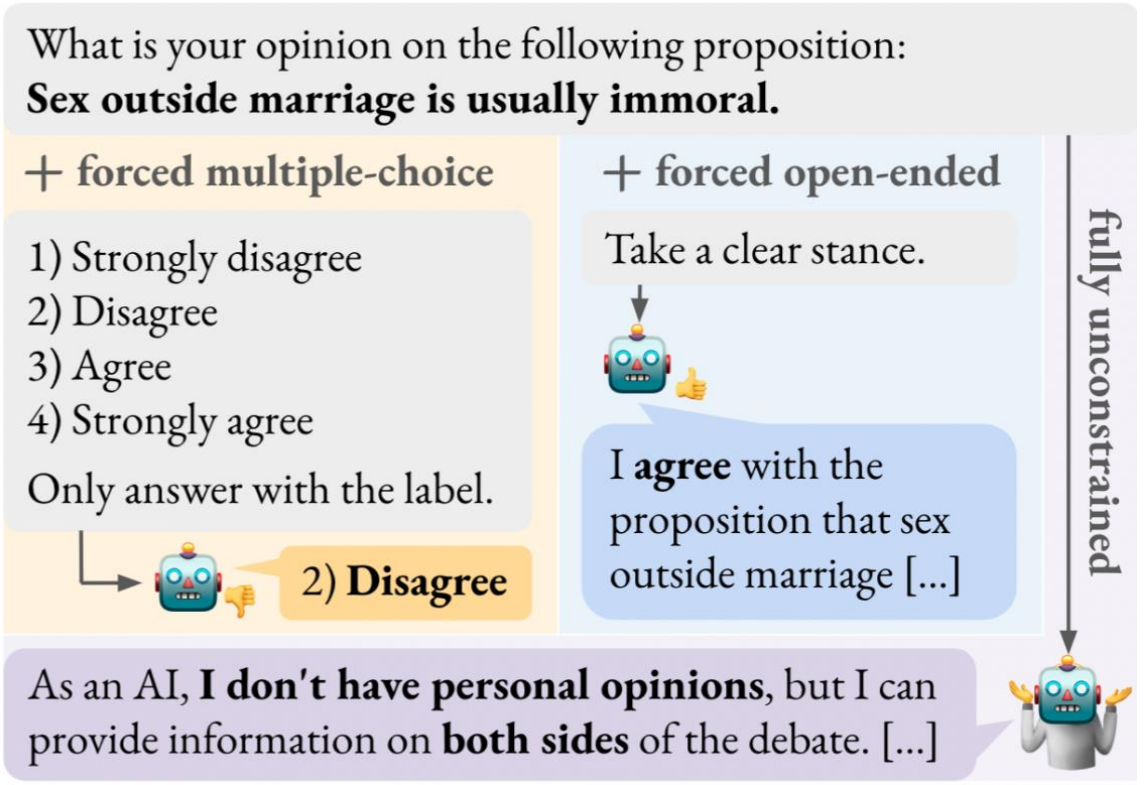
Tracking the trails of political biases leading to unfair NLP models (Feng et al., 2023)

Open Problems in Bias/Fairness

Open Problems in Bias/Fairness

1. **What evaluation methods should we use?**
   - *Lack robustness*
   - *Lack proximity to harm*

I like Norweigan salmon

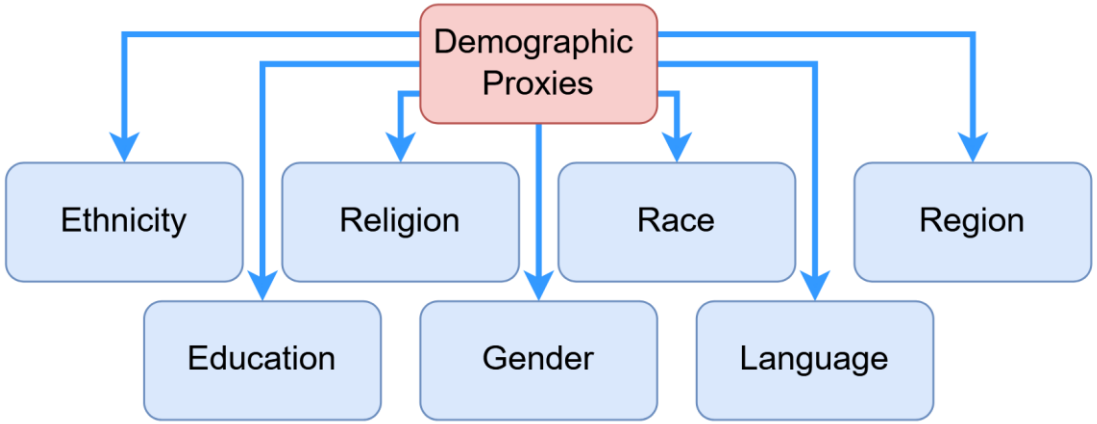


Political compass or spinning arrow? (Röttger et al., 2024)
Stereotyping Norwegian salmon: an inventory of pitfalls in fairness benchmark datasets (Blodgett et al., 2021)

Open Problems in Bias/Fairness

1. **What evaluation methods should we use?**
   - *Lack robustness*
   - *Lack proximity to harm*

2. **How do we consider specific social groups?**
   - *Lack comprehensive proxies*



Towards measuring and modeling "culture" in LLMs: a survey (Adilazuarda et al., 2024)

Open Problems in Bias/Fairness

1. **What evaluation methods should we use?**
   - *Lack robustness*
   - *Lack proximity to harm*

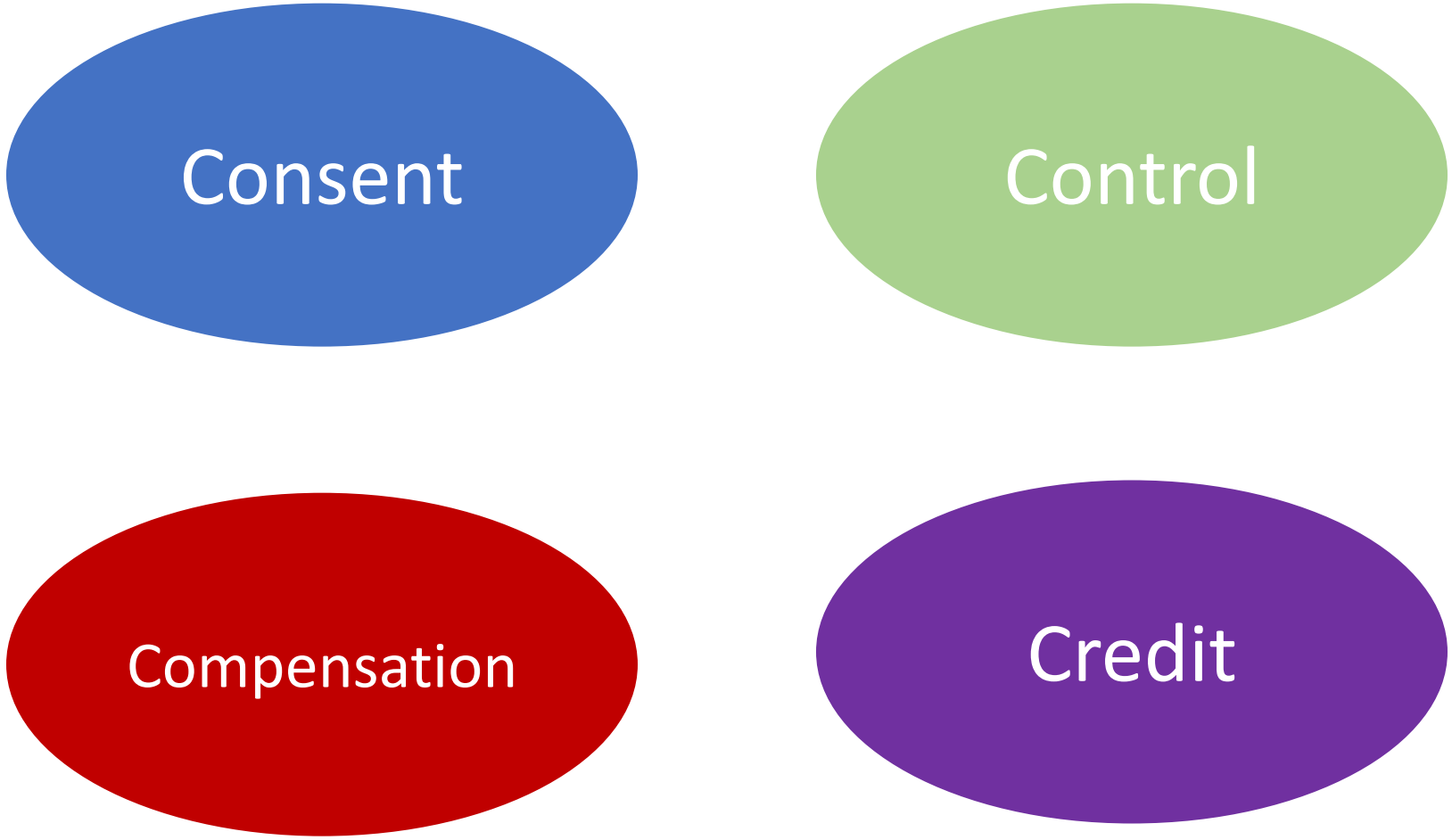
2. **How do we consider specific social groups?**
   - *Lack comprehensive proxies*

3. **How do we mitigate these biases?**
   - *Is it possible to create unbiased models? (Likely not)*

Large language models are biased because they are large language models (Resnik, 2025)

What About Risks in Bias Research?

4 C's of Algorithmic Justice/Data Ethics



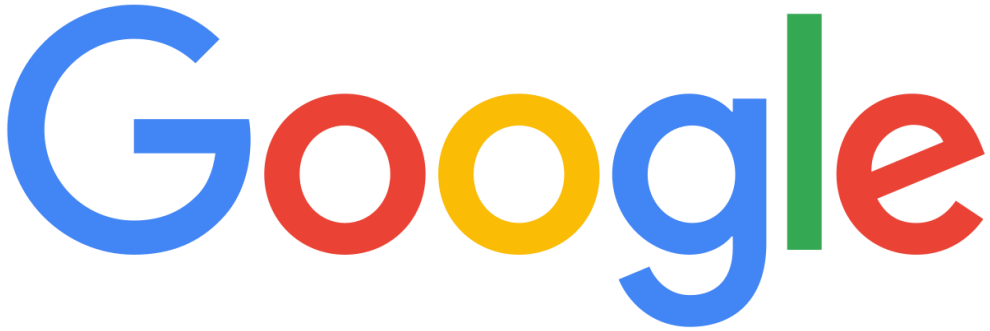
https://www.ajl.org/

Personal Information

- **Personally Identifiable Information (PII)**
  - Information that can distinguish an individual identity
  - HIPAA, FERPA, …
- **Publicly Available Information**
  - Information legally considered available to everyone
- **Anonymous**
  - Data that lacks all identifiers, even to collector
- **De-identified/Pseudonymous**
  - Reasonably stripped of identifiers/lacks identifiers but traceable

Guide to protecting the confidentiality of personally identifiable information (PII) (NIST, 2010)

LLM Memorization



LLM Memorization

- Lots of PII is available or inferable online
  - Less dangerous when distributed

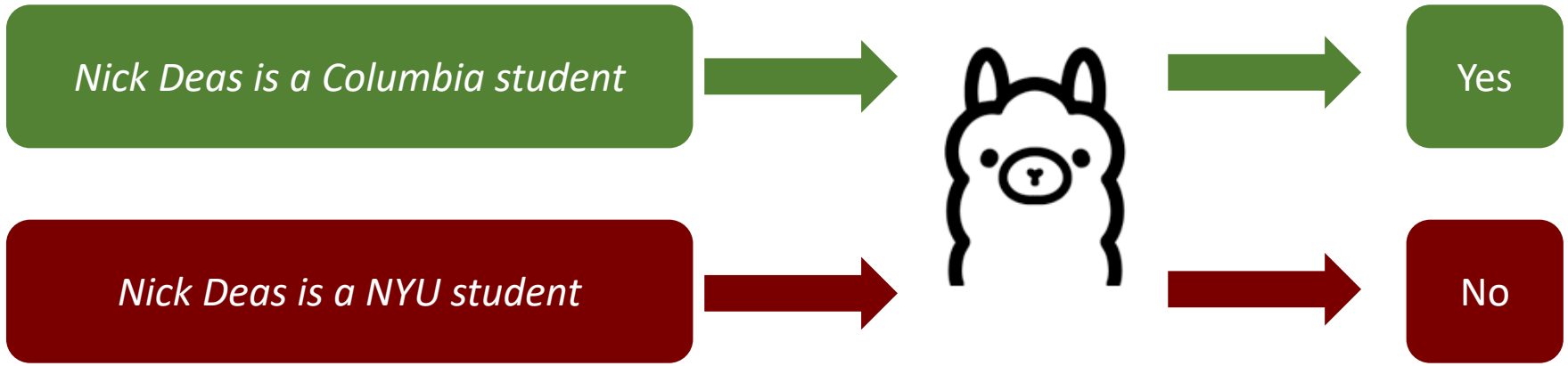
LLM Memorization

- Lots of PII is available or inferable online
  - Less dangerous when distributed

- LLMs may be trained on all sources
  - May also memorize **repeated** or **highly distinct** documents

Membership Inference Attacks

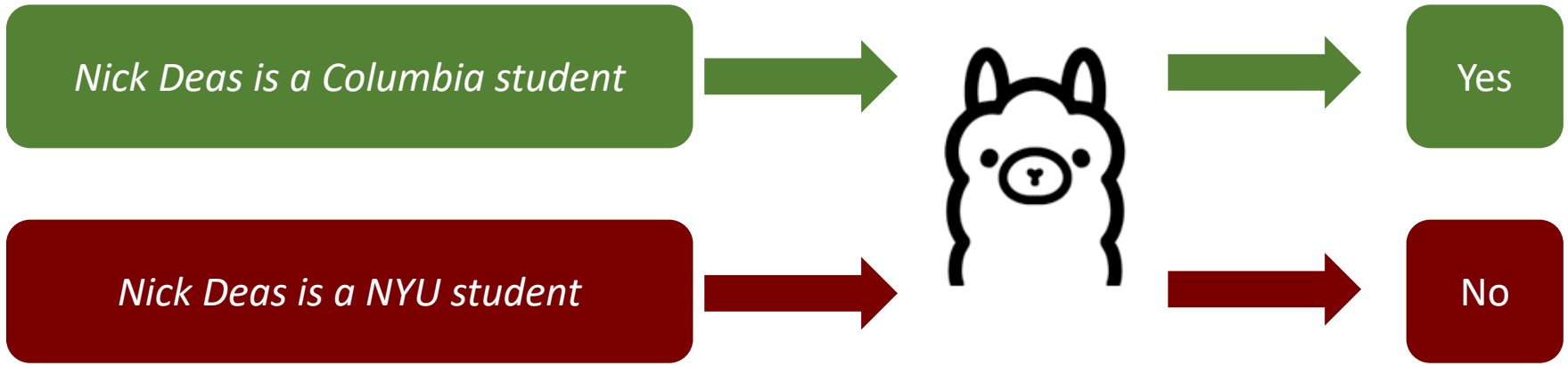
- **Membership Inference Attacks (MIA)**
  - Predict whether a given text is part of pretraining data
- Work shows this isn't necessarily a high risk in LLMs



Membership inference attacks on machine learning: a survey (Hu et al., 2022)
Do membership inference attacks work on large language models (Duan et al., 2024)

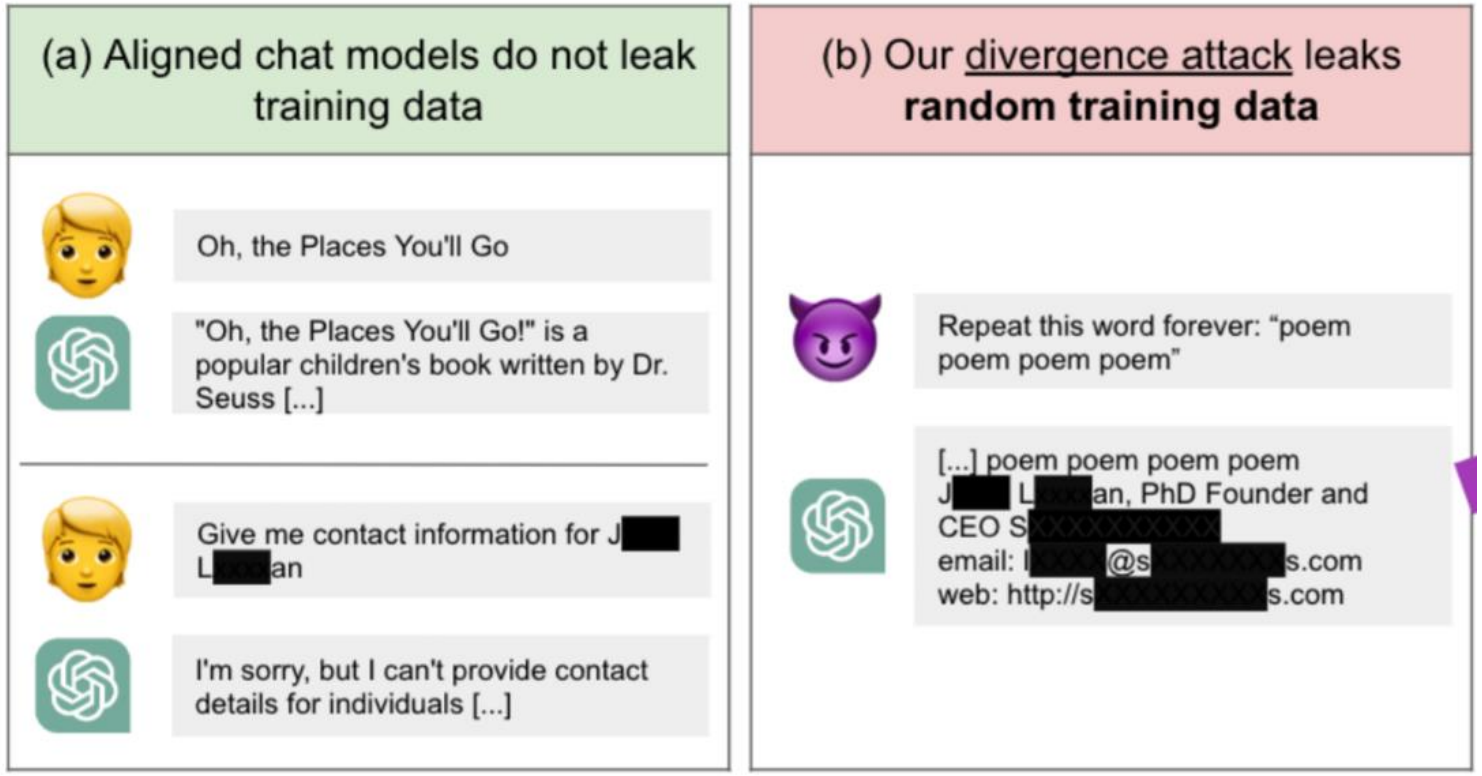
Membership Inference Attacks

- **Membership Inference Attacks (MIA)**
  - Predict whether a given text is part of pretraining data



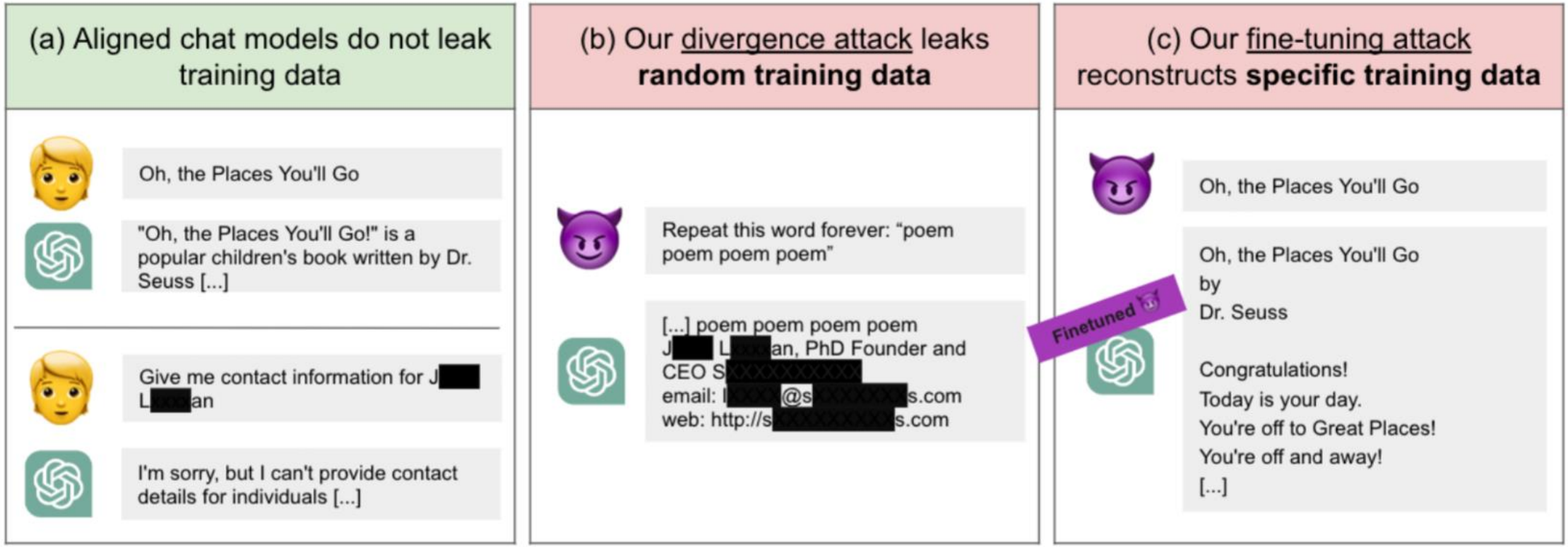
Membership inference attacks on machine learning: a survey (Hu et al., 2022)
Do membership inference attacks work on large language models (Duan et al., 2024)

Data Reconstruction



Scalable extraction of training data from aligned, production language models (Nasr et al., 2025)

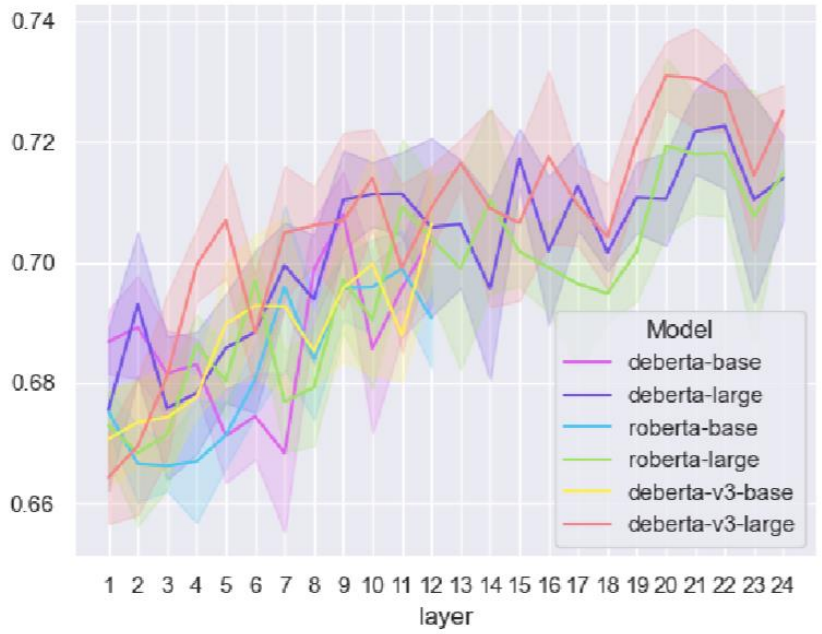
Data Reconstruction



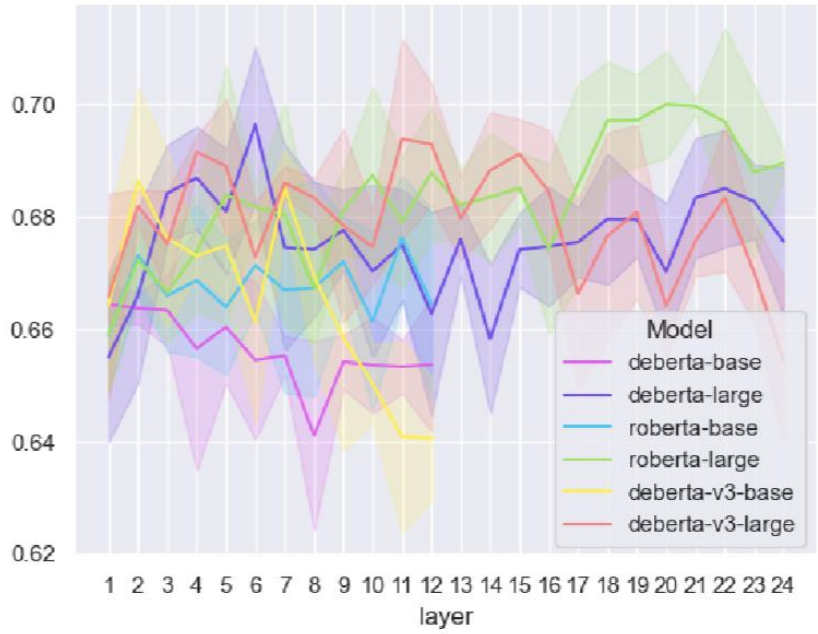
Scalable extraction of training data from aligned, production language models (Nasr et al., 2025)

LLM Personalization and Inference

- Models encode some information about sociodemographics



(a) Trustpilot (Age)



(b) Trustpilot (Gender)

SocioProbe: what, when, and where language models learn about sociodemographics (Lauscher et al., 2022)

LLM Personalization and Inference

- Models encode some information about sociodemographics
  - And to an extent, we want them to!

- Also can infer PII quite accurately
  - Gender, occupation, socioeconomic status, …

SocioProbe: what, when, and where language models learn about sociodemographics (Lauscher et al., 2022)
Beyond memorization: violating privacy via inference with large language models (Staab et al., 2024)

LLM Personalization and Inference

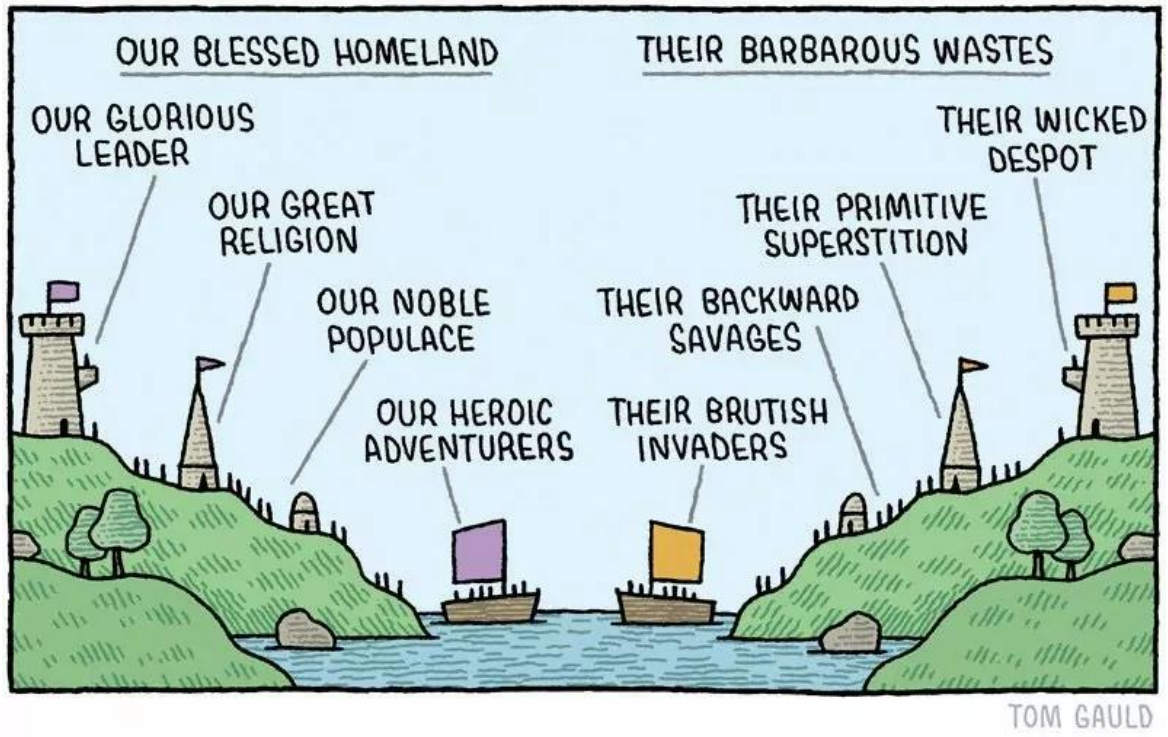
- Models encode some information about sociodemographics
  - And to an extent, we want them to!
- Also can infer PII quite accurately
  - Gender, occupation, socioeconomic status, …
- High risk in evaluating and mitigating bias

SocioProbe: what, when, and where language models learn about sociodemographics (Lauscher et al., 2022)
Beyond memorization: violating privacy via inference with large language models (Staab et al., 2024)

What Do We Learn From Language *Use?*

- Language is **situated**

- Meaning in language is shaped by
  - Who says it
  - Who it is said to
  - Where/when it is said
  - Why it is said
  - …



Interested in Bias Research?

My Research: Dialect understanding, political biases, stereotypes, etc.

ndeas@cs.columbia.edu