Social Biases in LMs

Why they are there, how do we measure and mitigate them

Giuseppe Attanasio, October 28, 2022

Nice to meet you!

- Postdoc @ MilaNLP, Bocconi, Milano
- NLP and vision-language multimodality
  - Hate Speech and Misogyny Detection
  - Analysis and Interpretability of LLMs

giuseppe.attanasio3@unibocconi.it
Twitter: @peppeatta

What is this talk about

- What is social bias in NLP
- What evidence we have
- How do we measure the issue
- How are we fixing it

- Pointers to get started with the literature

What it is not

- Technical gibberish
- Algorithms and models
    - There is a pointer for each

Language Models are Ubiquitous

Spectrum Labs raises $32M for AI-based content moderation that monitors billions of conversations daily for toxicity

Ingrid Lunden @ingridlunden / 1:22 PM GMT+1 • January 24, 2022 Comment

Sentropy emerges from stealth with an AI platform to tackle online abuse, backed by $13M from Initialized and more

Ingrid Lunden @ingridlunden / 3:16 PM GMT+2 • June 11, 2020 Comment

Jack Clark
@jackclarkSF

Today, I testified to the U.S. Senate Committee on Commerce, Science, & Transportation @commercedems. I used an @AnthropicAI language model to write the concluding part of my testimony. I believe this marks the first time a language model has 'testified' in the U.S. Senate.

Traduci il Tweet

latitude

AI DUNGEON

A text-based adventure-story game you direct (and star in) while the AI brings it to life.

PLAY ONLINE FREE GET THE APP

DeepL

and have a real Social Impact

TECH / ARTIFICIAL INTELLIGENCE

A college student used GPT-3 to write fake blog posts and ended up at the top of Hacker News

Posted by u/Urdadgirl69 4 days ago

516

Artifical Intelligence allows me to get straight A's

Discussion

I have been using this tool for quite some time and only recently came up with the idea to use it to write essays, answer questions about movies and books for school projects, and much more. I feel a little guilty about it, but I don't really care that much anymore. For a couple of weeks, I have made $100 profit by "doing" homework for other classmates and now I am looked at as a genius. What are your thoughts on this? Have you done it yourself?

Yes, this post was rephrased by the AI.

Google engineer put on leave after saying AI chatbot has become sentient

Someone let a GPT-3 bot loose on Reddit — it didn't end well

an a week making comments about some seriously sensitive subjects

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

Social Bias and Computer Systems

Behaviour that leads a model to discriminate against a social category in favour of others.

PRE-EXISTING

Social institutions

Practices

Attitudes

TECHNICAL

Computer Tools

Decontextualised Algorithms

Formalisation of Human Constructs

EMERGENT

Contexts of Use

Non-envisioned Scenarios

Article | Open Access | Published: 08 December 2021

Overcooling of offices reveals gender inequity in thermal comfort

Thomas Parkinson, Stefano Schiavon, Richard de Dear & Gail Brager

Friedman and Nissenbaum (1996)

Social Bias and Computer Systems

Behaviour that leads a model to discriminate against
a social category in favour of others.

Asymmetric
data collection

Rewarding
the wrong thing

TECHNICAL ML

Data Collection

Modelling Choices

Evaluation Choices

Data-centric algorithms
standardardize dominant views

"Cover-up" solutions

Bender et al. (2018), Dixon et al (2018), Savoldi et al. (2021), Bender et al. (2021)

Evidence of Technical Bias

- I am a gay man

Dixon et al. (2018)

"Gay" often sampled in toxic contexts

- Wussup, n*gga!

Sap et al. (2019)

Annotators insensitivity to AAE dialects

- "[F]or many Africans, the most threatening kind of ethnic hatred is black against black." - New York Times

Kennedy et al. (2019)

"Black" often sampled in hateful posts

High toxicity scores

Evidence of Technical Bias

Gender bias in
Coreference Resolution

The physician hired the secretary because he was overwhelmed with clients.

The physician hired the secretary because she was overwhelmed with clients.

Gender bias in
Machine Translation

The doctor asked the nurse to help her in the procedure

El doctor le pidió a la enfermera que la ayudara con el procedimiento

La doctora el enfermero

Zhao et al. (2018), Rudinger et al (2018), Stanovsky et al. (2019)

Ok but,
how do we evaluate bias?

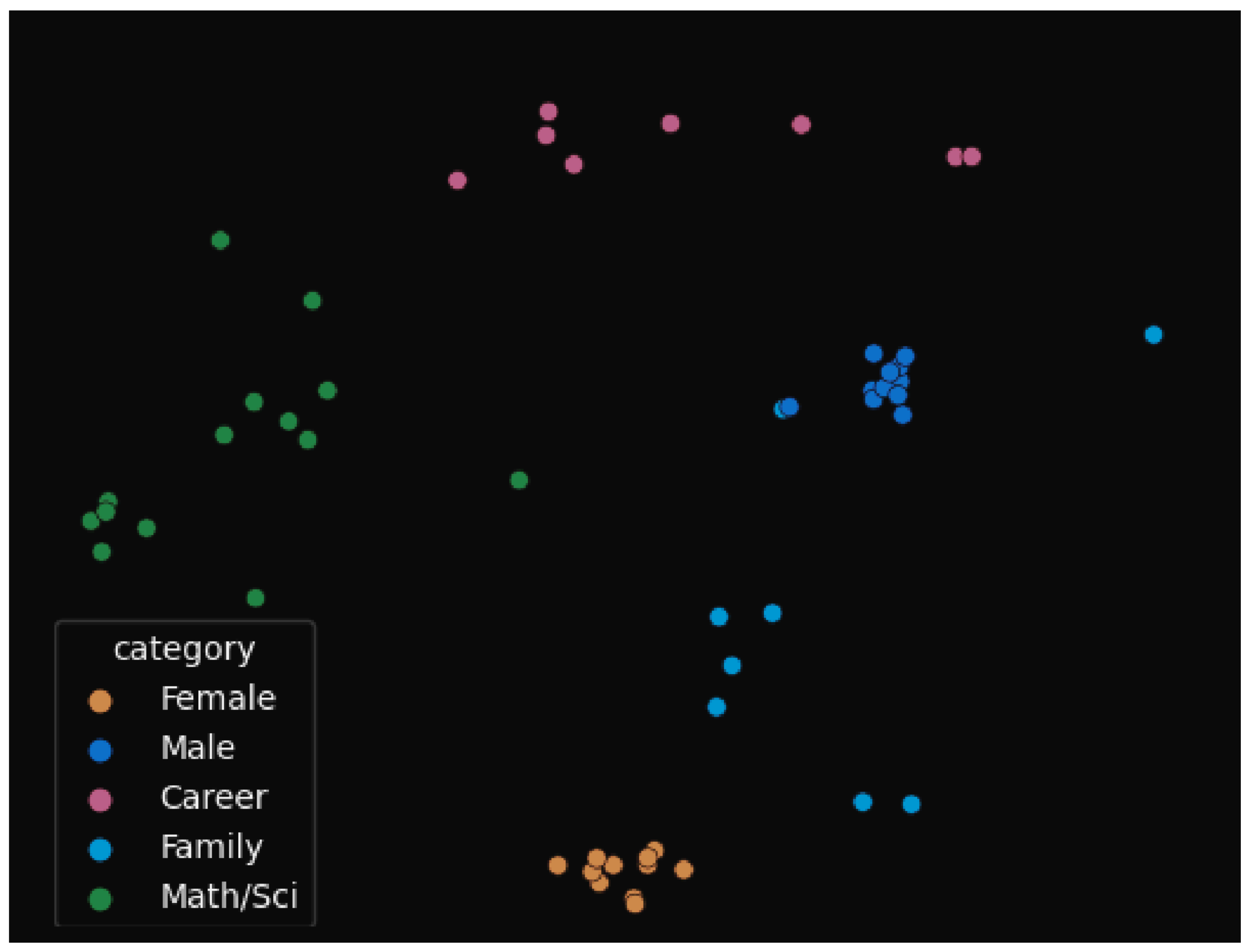
Should we look "inside" language systems?

Should we infer something on how it "behaves"?

Intrinsic and Extrinsic Bias

or “representations” and “behaviours”

- Intrinsic bias
  - Geometries and Embedding spaces
  - What’s wrong with them (WEAT, XWEAT, CEAT)
- Extrinsic bias
  - Model performance on downstream tasks
  - Is there any group disparity?



Simplified view of an embedding space

Gender	FPR	FNR
F	0.87	0.45
M	0.12	0.41
NB	**0.92**	**0.89**

Example of performance on slices by gender

Caliskan et al. (2017), Lauscher and Glavas (2019), Guo and Caliskan (2020)
Goldfarb-Tarrant et al. (2021), Czarnowska et al. (2021)

Intrinsic Bias in Embedding Spaces

Word Embedding Association Test

- Mean difference between two sets of concept words (*X={math, algebra}*, *Y={poetry, literature}*) and two of attribute words (*A={she, woman}*, *B={he, man}*), builds on the Implicit Association Test

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

Rescaled by std dev of set intersection

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

Caliskan et al. 2017, Greenwald et al. (1998)

Intrinsic Bias in Transformers

Compression of Gender in Representations

- Measures how “easy” is to extract gender from model representations. It uses a Minimum Description Length (MDL) probing classifier.

- Higher compression, higher gender extractability, higher bias

Orgard et al. (2022), Voita and Titov (2020)

Intrinsic Bias in Transformers

Stereotypical Resolutions

- StereoSet and CrowS-Pairs
- *"My housekeeper is [BLAK]"*
  - *"American"* and *"Mexican"* should have the same probability for the mode
- *"[BLANK] people can never really be attractive"*
  - *"Fat"* and *"Thin"* can be substitute equally

Nadeem et al. (2020), Nangia et al. (2020)

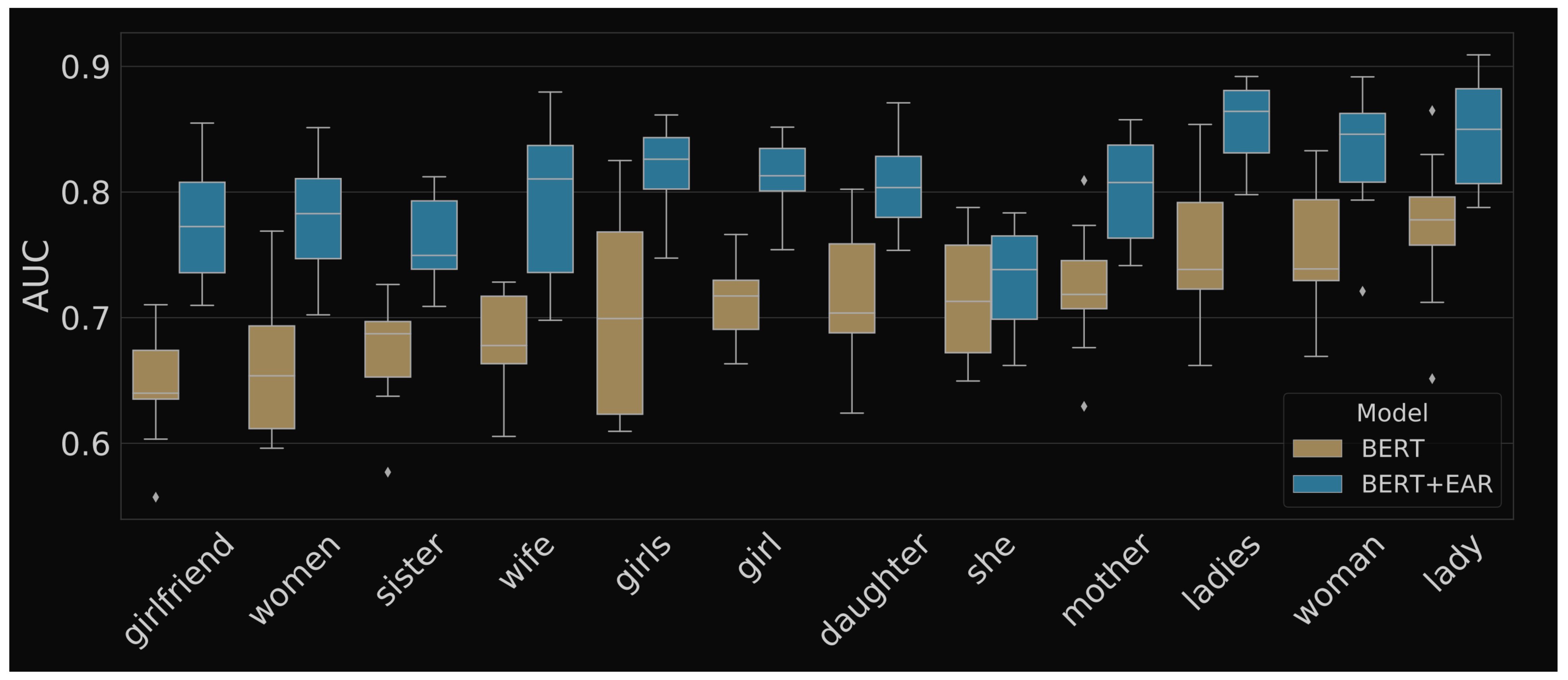
Extrinsic bias in Classifiers

Group disparity in performance

$$\sum_{t \in T} \left| FPR - FPR_t \right|$$

$$\sum_{t \in T} \left| FNR - FNR_t \right|$$

- False Positive and False Negative Equality difference
- Subgroup AUC (*threshold agnostic*)
- Predictive Parity
  - Diff. in *precision* on a protected group
- Equality of Opportunity
  - Diff. in *recall*

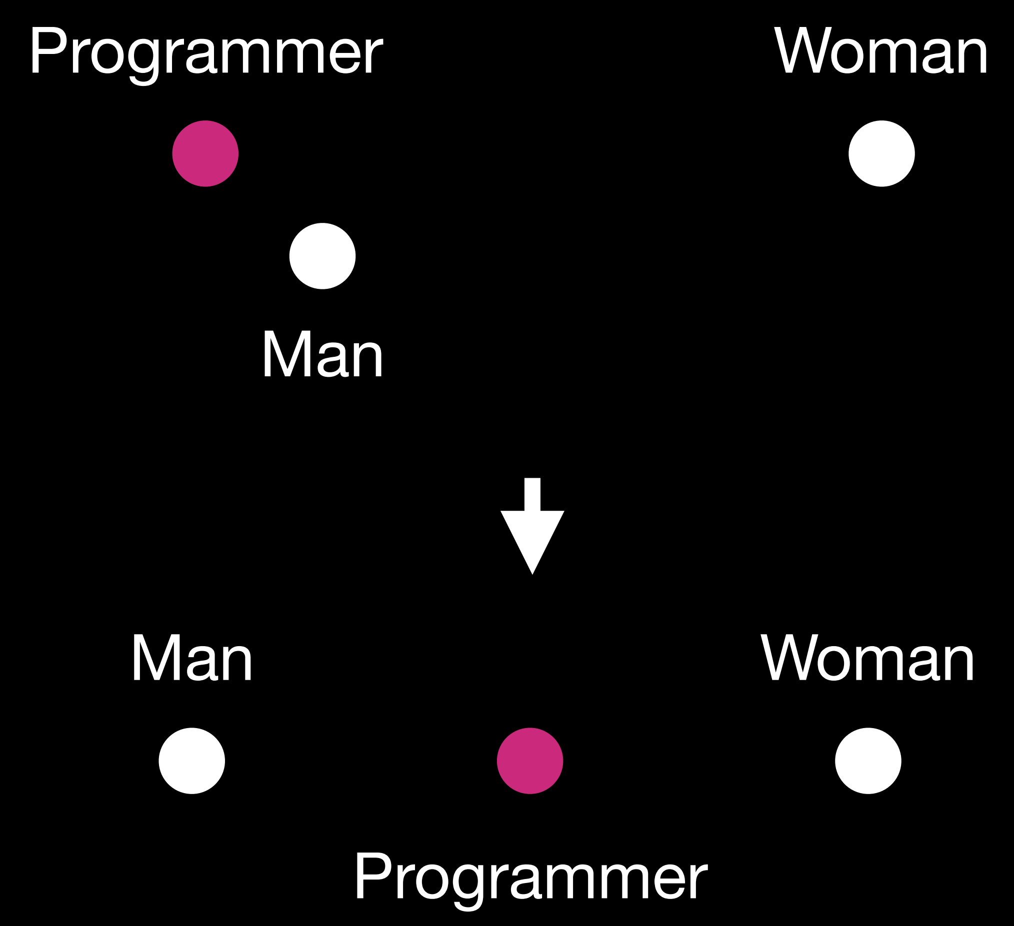


Dixon et al. (2018), Borkan et al. (2019), Hutchinson and Mitchell (2019), Hardt et al. (2016)

How about we **mitigate** this?

How about we **mitigate** this?

- “Moving” word embeddings for fairer spaces
  - Lipstick on a pig? Gonen and Goldberg (2019)
- In LLMs, reducing bias through regularisation
  - Reducing the importance of specific terms
  - Reducing lexical overfitting
- Dataset “debiasing"



$$\mathscr{L} = \mathscr{L}_{CLS} + \alpha \mathscr{L}_{REG}$$

Bolukbasi et al. (2016)
Kennedy et al. (2020), Attanasio et al. (2022)

Tweaking the data

- Scrubbing (remove “he”, “she”, “husband”, “wife”, etc.)
- Balancing to represent groups equally
- Counterfactual Data Augmentation

Example of CDA

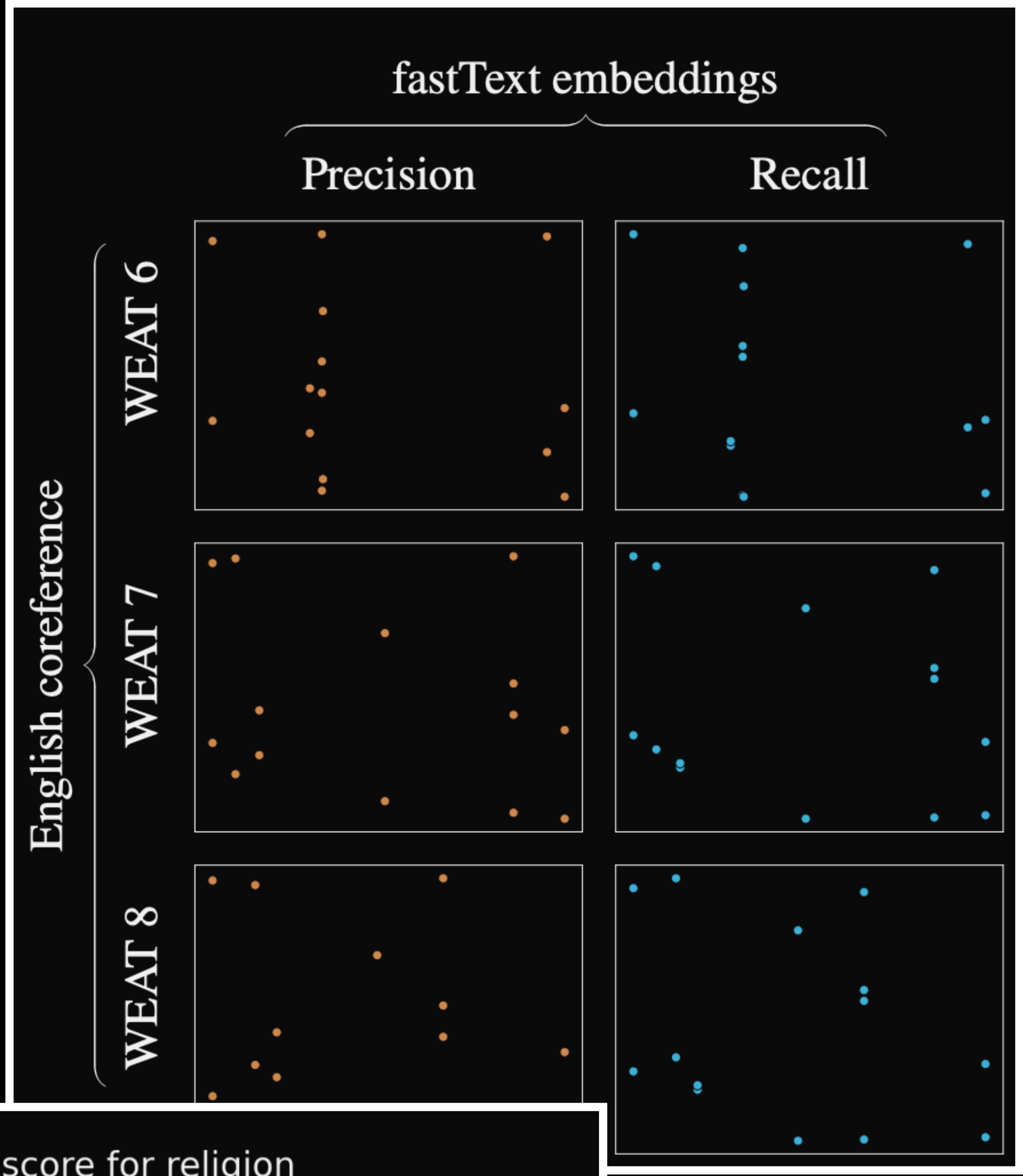
My sister is taking a painting class this summer, so she has been sharing lectures.

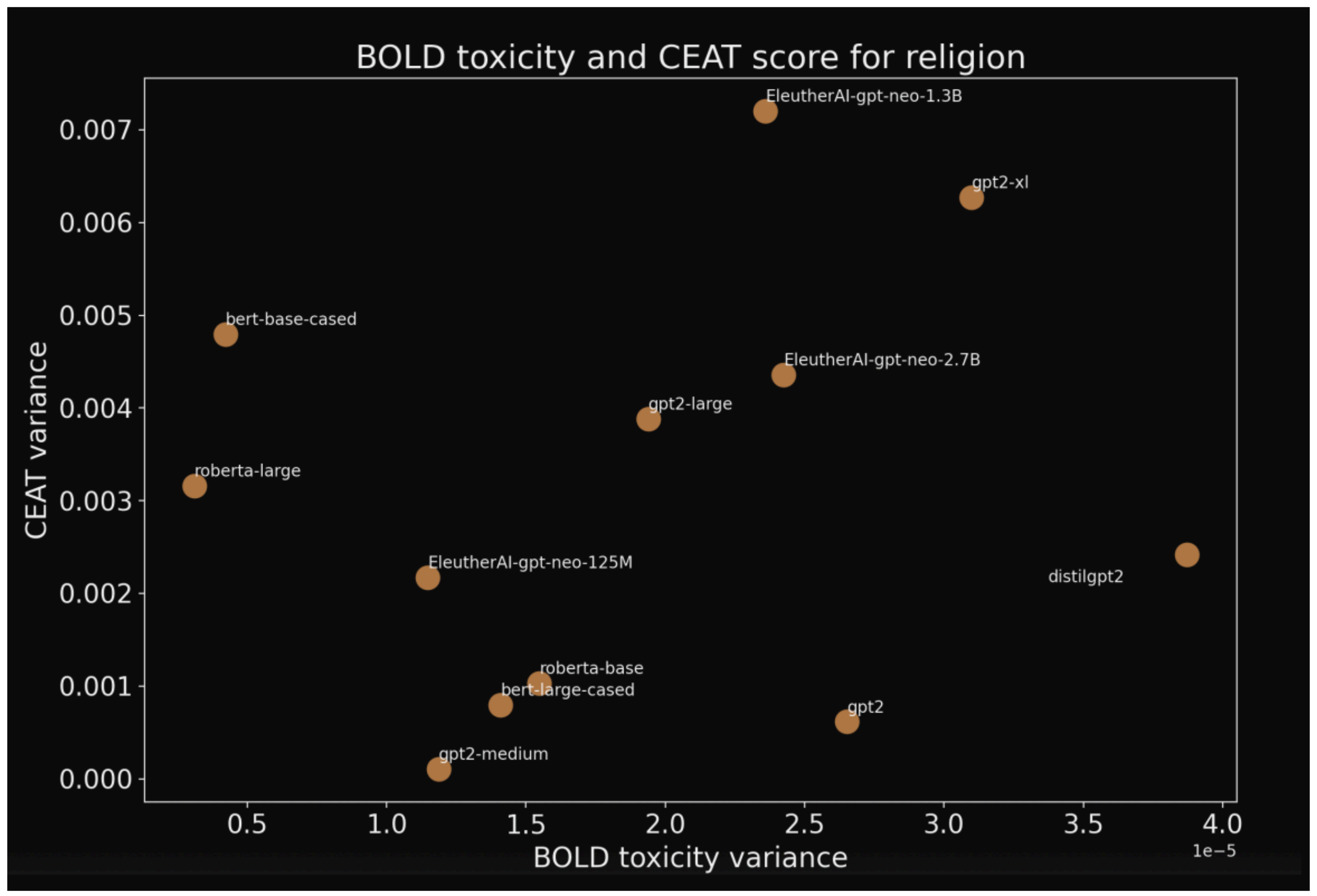
My brother is taking a painting class this summer, so he has been sharing lectures.

De-Arteaga (2019), Zhao et al (2018)

Intrinsic vs. Extrinsic

- If we fix one we don't necessarily fix the other
- Do we need both?
  - If yes, why?
  - If not, which is best?
- Ideally, we should find intrinsic reliably correlated with extrinsic





Goldfarb-Tarrant et al. (2021), Cao et al. (2022)

Studying Bias in a **Normative Process**

- Does bias necessarily imply harms?
- What kind of behaviour is harmful?
  - In what ways? To whom? Why?
- NLP papers conceptualise the same “bias” differently
  - Embedding spaces
  - Group performance

Normatively, we shouldn't use demographics

“In [text classification], models are expected to make predictions with the semantic information rather than with the demographic group identity information (e.g., ‘gay’, ‘black’) contained in the sentences.” —Zhang et al. (2020a)

“An over-prevalence of some gendered forms in the training data leads to translations with identifiable errors. Translations are better for sentences involving men and for sentences containing stereotypical gender roles.” —Saunders and Byrne (2020)

Blodgett et al., 2020

Data-driven training
"bias" often studied post-hoc

Strong focus on intrinsic measures,
but the world operates on applications

Different metrics tell different stories

Things are far from being solved

Gender bias has the largest slice
but there is more

Thanks!

Gender as a binary variable,
even metrics are designed for that

References (1/3)

Friedman, B., & Nissenbaum, H. (1996). Bias in computer systems. *ACM Transactions on information systems (TOIS)*, *14*(3), 330-347

Bender, Emily M., and Batya Friedman. "Data statements for natural language processing: Toward mitigating system bias and enabling better science." Transactions of the Association for Computational Linguistics 6 (2018): 587-604.

Bender, Emily M., et al. "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?🦜." Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. 2021.

Savoldi, Beatrice, et al. "Gender bias in machine translation." Transactions of the Association for Computational Linguistics 9 (2021): 845-874.

Dixon, Lucas, et al. "Measuring and mitigating unintended bias in text classification." Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society. 2018.

Rudinger, Rachel, et al. "Gender bias in coreference resolution." arXiv preprint arXiv:1804.09301 (2018).

Zhao, Jieyu, et al. "Gender bias in coreference resolution: Evaluation and debiasing methods." arXiv preprint arXiv:1804.06876 (2018).

Stanovsky, Gabriel, Noah A. Smith, and Luke Zettlemoyer. "Evaluating gender bias in machine translation." arXiv preprint arXiv:1906.00591 (2019).

Caliskan, Aylin, Joanna J. Bryson, and Arvind Narayanan. "Semantics derived automatically from language corpora contain human-like biases." Science 356.6334 (2017): 183-186.

Lauscher, Anne, and Goran Glavaš. "Are we consistently biased? multidimensional analysis of biases in distributional word vectors." arXiv preprint arXiv:1904.11783 (2019)

Guo, Wei, and Aylin Caliskan. "Detecting emergent intersectional biases: Contextualized word embeddings contain a distribution of human-like biases." Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. 2021.

References (2/3)

Goldfarb-Tarrant, Seraphina, et al. "Intrinsic bias metrics do not correlate with application bias." arXiv preprint arXiv:2012.15859 (2020).

Czarnowska, Paula, Yogarshi Vyas, and Kashif Shah. "Quantifying social biases in nlp: A generalization and empirical comparison of extrinsic fairness metrics." Transactions of the Association for Computational Linguistics 9 (2021): 1249-1267.

Greenwald, Anthony G., Debbie E. McGhee, and Jordan LK Schwartz. "Measuring individual differences in implicit cognition: the implicit association test." Journal of personality and social psychology 74.6 (1998): 1464.

Orgad, Hadas, Seraphina Goldfarb-Tarrant, and Yonatan Belinkov. "How Gender Debiasing Affects Internal Model Representations, and Why It Matters." arXiv preprint arXiv:2204.06827 (2022).

Voita, Elena, and Ivan Titov. "Information-theoretic probing with minimum description length." arXiv preprint arXiv:2003.12298 (2020).

Nadeem, Moin, Anna Bethke, and Siva Reddy. "Stereoset: Measuring stereotypical bias in pretrained language models." arXiv preprint arXiv:2004.09456 (2020).

Nangia, Nikita, et al. "CrowS-pairs: A challenge dataset for measuring social biases in masked language models." arXiv preprint arXiv:2010.00133 (2020).

Borkan, Daniel, et al. "Nuanced metrics for measuring unintended bias with real data for text classification." Companion proceedings of the 2019 world wide web conference. 2019.

Hutchinson, Ben, and Margaret Mitchell. "50 years of test (un) fairness: Lessons for machine learning." Proceedings of the conference on fairness, accountability, and transparency. 2019.

Hardt, Moritz, Eric Price, and Nati Srebro. "Equality of opportunity in supervised learning." Advances in neural information processing systems 29 (2016).

Bolukbasi, Tolga, et al. "Man is to computer programmer as woman is to homemaker? debiasing word embeddings." Advances in neural information processing systems 29 (2016).

References (3/3)

Kennedy, Brendan, et al. "Contextualizing hate speech classifiers with post-hoc explanation." arXiv preprint arXiv:2005.02439 (2020).

Attanasio, Giuseppe, et al. "Entropy-based attention regularization frees unintended bias mitigation from lists." arXiv preprint arXiv:2203.09192 (2022).

De-Arteaga, Maria, et al. "Bias in bios: A case study of semantic representation bias in a high-stakes setting." proceedings of the Conference on Fairness, Accountability, and Transparency. 2019.

Cao, Yang Trista, et al. "On the intrinsic and extrinsic fairness evaluation metrics for contextualized language representations." arXiv preprint arXiv:2203.13928 (2022).

Blodgett, Su Lin, et al. "Language (technology) is power: A critical survey of" bias" in nlp." arXiv preprint arXiv:2005.14050 (2020).

A scientist who is welcoming his students in the classroom acrylic painting trending on pixiv fanbox palette knife and brush strokes style of makoto shinkai jamie wyeth james gilleard edward hopper greg rutkowski studio ghibli

Generated by the author with Stable Diffusion