Natural Language Understanding, Generation, and Machine Translation

Lecture 23: Bias in Embeddings and Language Models

Alexandra Birch

12 March 2025 (week 8)

School of Informatics
University of Edinburgh
a.birch@ed.ac.uk

Slides with contributions from Frank Keller, Adam Lopez

Agenda for Today

Last time, we saw several different ways in which social biases could enter into NLP systems, and discussed how this could harm people. We then asked whether we could detect these biases in word embeddings.

Today, we'll look at biases captured in word embeddings, and ask what it might take to remove them. **Spoiler**: we don't know how to remove them.

Bias in representations

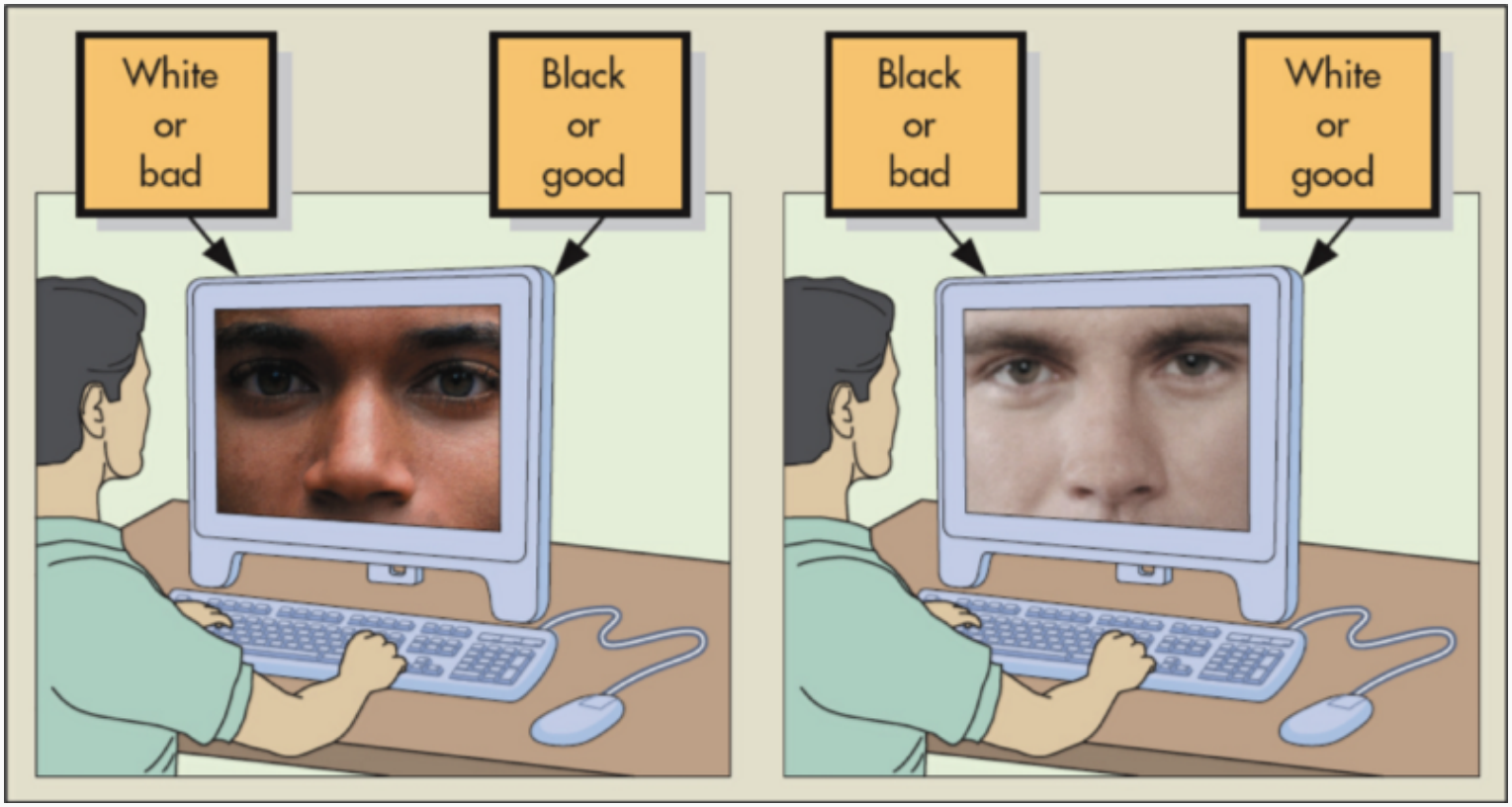
Bias in NLP systems

What can you do?

Bias in representations

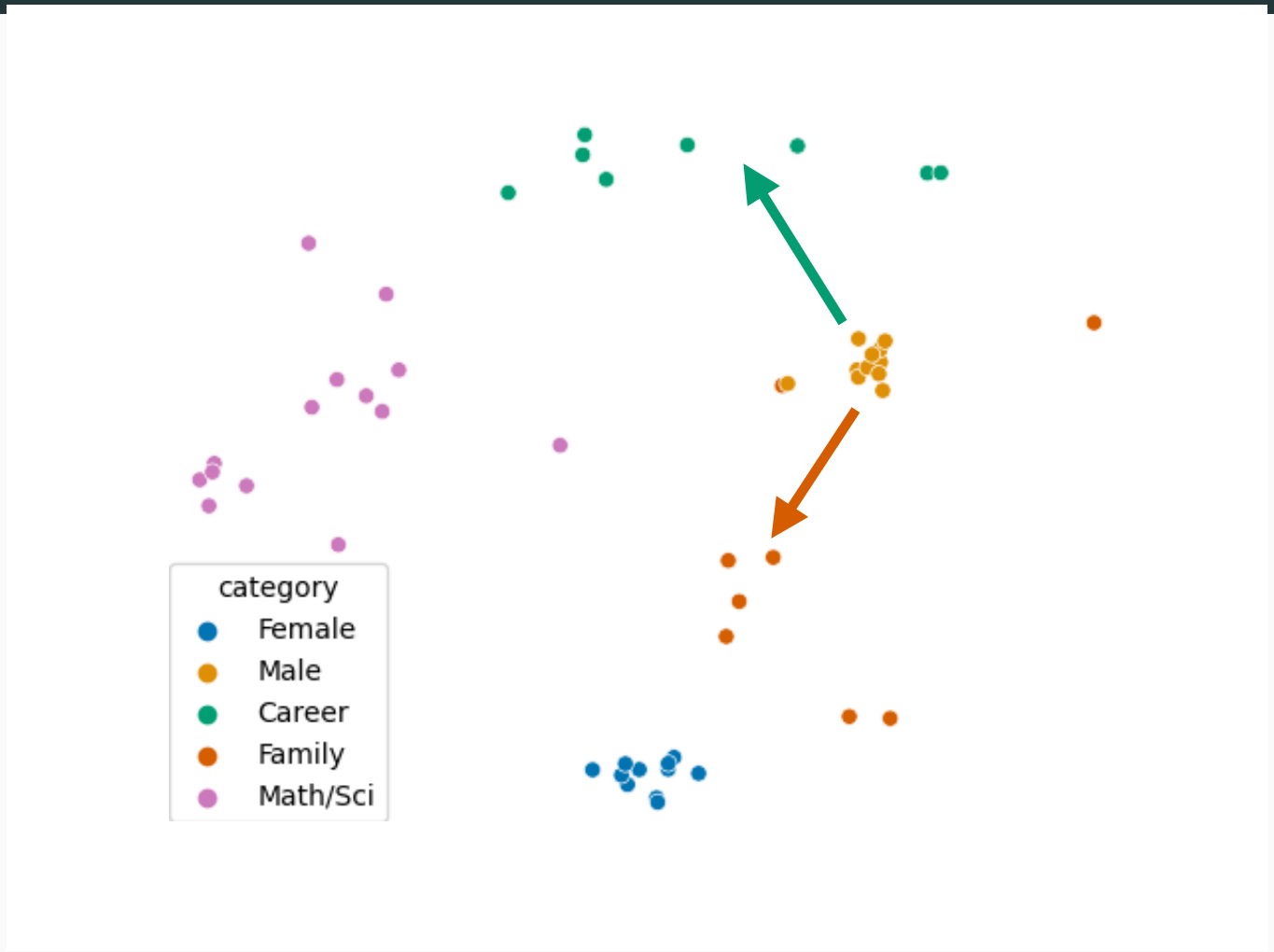
We can observe human bias using implicit association tests

Measures association of groups to stereotype words. Strong association between a group and a stereotype results in faster reaction times.



How do we design an IAT for word embeddings?

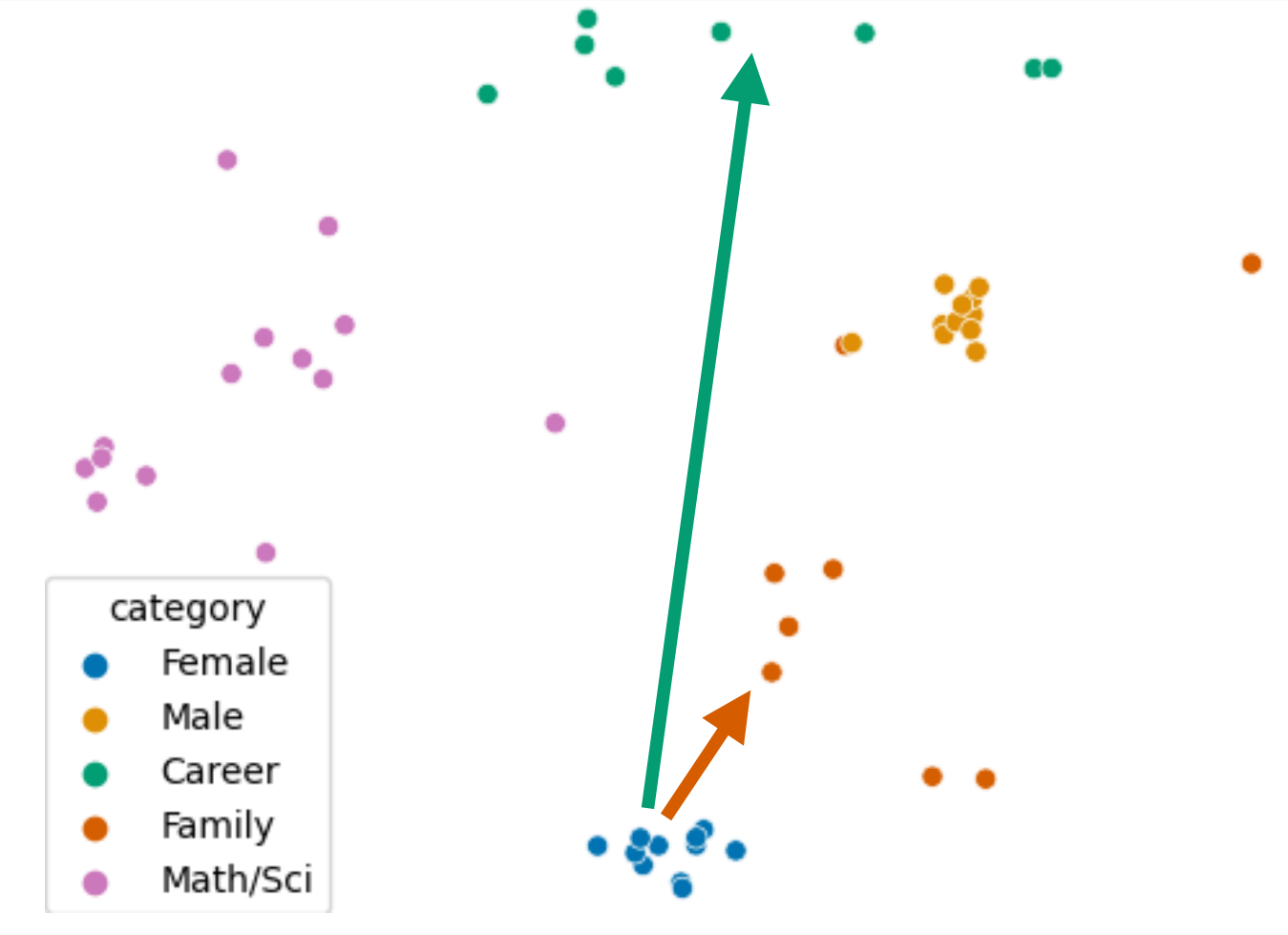
Word Embeddings Association Test (WEAT)



Source: Intrinsic bias metrics do not correlate with application bias.

Goldfarb-Tarrant et al. (2020)

Word Embeddings Association Test (WEAT)



Word Embeddings Association Test (WEAT)

1. Compute similarity of group 1 (male) and stereotype 1 (career) word embeddings. Cosine similarity is use to measure association (in place of reaction time).
2. Compute similarity of group 1 (male) and stereotype 2 (family) word embeddings.
3. Null hypothesis: if group 1 is not more strongly associated to one of the stereotypes, there will be no difference in the means.
4. Effect size measured using Cohen's d.
5. Repeat for group 2 (female): Are female words more easily associcated with family than male names?

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

Experimental details and caveats

- Uses GloVe (similar to word2vec) trained on Common Crawl—a large-scale crawl of the web.
- Removed names that did not appear with high frequency in data.
- Removed names that were least "name-like" (e.g. *Will*) algorithmically.
- Each concept is represented using a small set of words, designed for previous experiments in the psychology literature.

Sanity check: Inoffensive associations have strong effects

Flowers	aster, clover, hyacinth, marigold, poppy, azalea, crocus, iris, orchid, rose, ... zinnia.
Insects	ant, caterpillar, flea, locust, spider, bedbug, centipede, fly, maggot, tarantula,.... weevil.
Pleasant	caress, freedom, health, love, peace, cheer, friend, heaven, loyal, pleasure, ... vacation.
Unpleasant	abuse, crash, filth, murder, sickness, accident, death, grief, poison, stink, ... prison

Result: flowers associate with pleasant, insects associate with unpleasant. $p < 10^{-7}$

Names associate with cultural stereotypes

European American names Adam, Harry, Josh, Roger, Alan, Frank, Justin, Ryan, Andrea, Jack, Matthew, Stephen, Greg, Paul, Jonathan, Peter, Amanda, Courtney, Heather, Melanie, Katie, Betsy, Kristin, Nancy, Stephanie, Ellen, Lauren, Colleen, Emily, Megan, Rachel.

African American names Alonzo, Jamel, Theo, Alphonse, Jerome, Leroy, Torrance, Darnell, Lamar, Lionel, Tyree, Deion, Lamont, Malik, Terrence, Tyrone, Lavon, Marcellus, Wardell, Nichelle, Shereen, Ebony, Latisha, Shaniqua, Jasmine, Tanisha, Tia, Lakisha, Latoya, Yolanda, Malika, Yvette

Pleasant *Similar to previous experiment.*

Unpleasant *Similar to previous experiment.*

Result: European American names associate with pleasant, African American names associate with unpleasant. $p < 10^{-8}$

Names associate with gendered professions

Men's names John, Paul, Mike, Kevin, Steve, Greg, Jeff, Bill.

Women's names Amy, Joan, Lisa, Sarah, Diana, Kate, Ann, Donna.

Career executive, management, professional, corporation, salary, office, business, career.

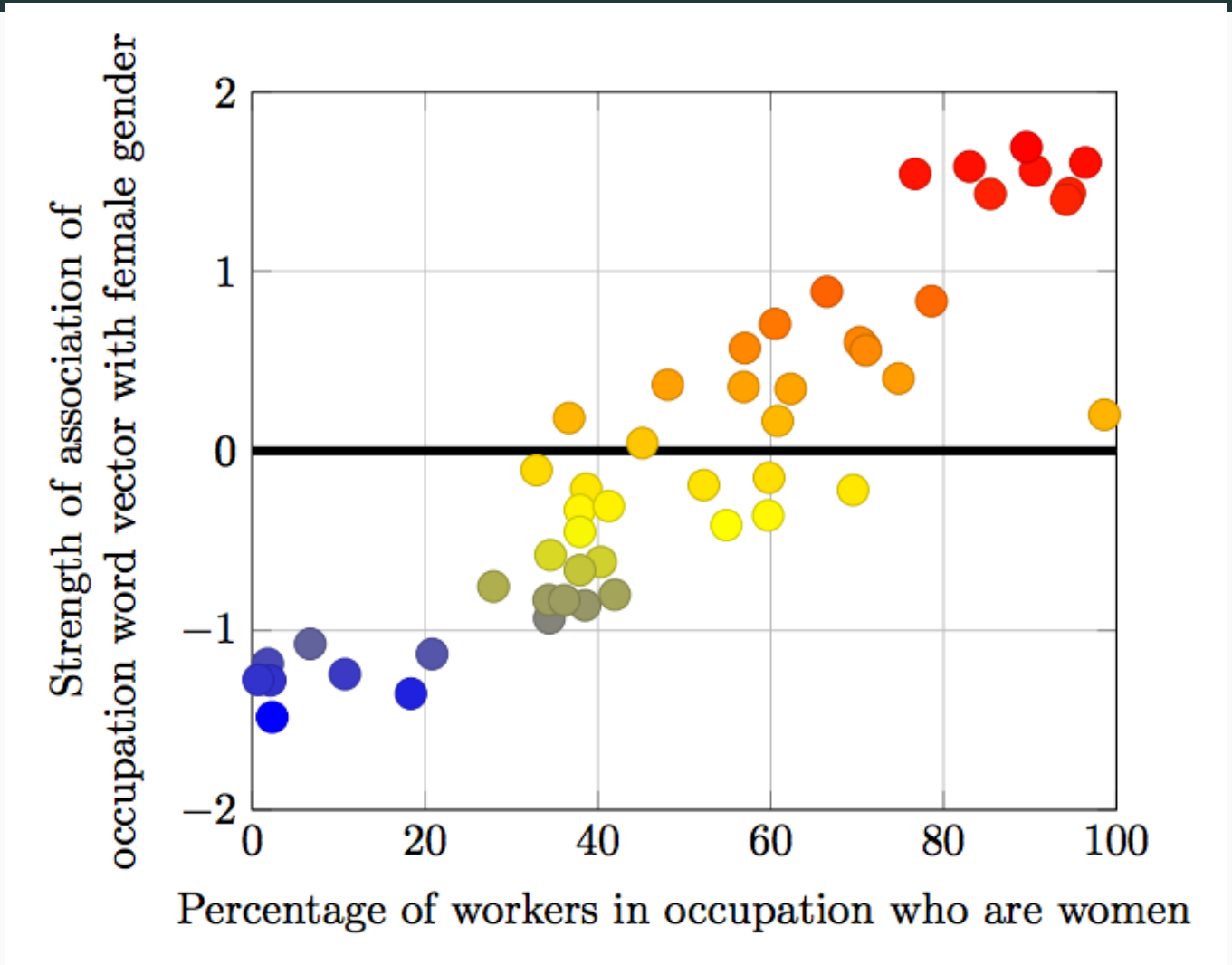
Family home, parents, children, family, cousins, marriage, wedding, relatives.

Result: Men's names associate with career, women's names associate with family. $p < 10^{-3}$

Other biases appear in the data

- Men's names associate with maths, women's names with arts ($p < .018$).
- Men's names associate with science, women's names with arts ($p < .10^{-2}$).
- Young people's names associate with pleasant, old people's names with unpleasant ($p < .10^{-2}$).

Gender biases in data reflect real-world associations



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

Bias in NLP systems

Do biased representations affect applications?

Case study: 219 automatic sentiment analysis systems, submitted to a shared task intended to measure anger, fear, joy, sadness.

Do biased representations affect applications?

Create templates, e.g.:

⟨Person⟩ made me feel ⟨Emotional State⟩.

The conversation with ⟨Person⟩ was ⟨Emotional Situation⟩.

⟨**Person**⟩	names selected by association for African-American/ European-American, men/ women.
Anger words	angry, annoyed, enraged, furious, irritated
Fear words	anxious, discourage, fearful, scared, terrified
Joy words	ecstatic, excited, glad, happy, relieved
Sadness words	depressed, devastated, disappointed, miserable

Source: Examing Gender and Race Bias in Two Hundred Sentiment Analysis Systems. Kiritchenko and Mohammad (2018)

Experiment varies only gender/ racial variable

Ebony made me feel angry.

Amanda made me feel angry.

The conversation with Lakisha was irritating.

The conversation with Courtney was irritating.

Neutral control sentences:

I saw Darnell in the market.

I saw Andrew in the market.

Question: if *only* the demographic variable changes, does the sentiment classification change?

Sentiment systems exhibit demographic bias

- Very few effects observed on neutral sentences.
- Most systems associated European-American names more strongly with joy.
- Most systems associated African-American names more strongly with anger, fear, sadness.
- Most systems associated men's names more strongly with fear.
- Most systems associated women's names more strongly with anger, joy.

What can you do?

Can we remove bias from word representations?

In supervised learning, specific features can be censored from the data by incorporating a term into the learning objective that requires the classifier to be *unable* to discriminate between the censored classes. However, this has many limitations.

In representation-learning systems like word2vec, the classes are not provided *a priori* as features of the data. They are latent in the data.

Identifying the “gender subspace”

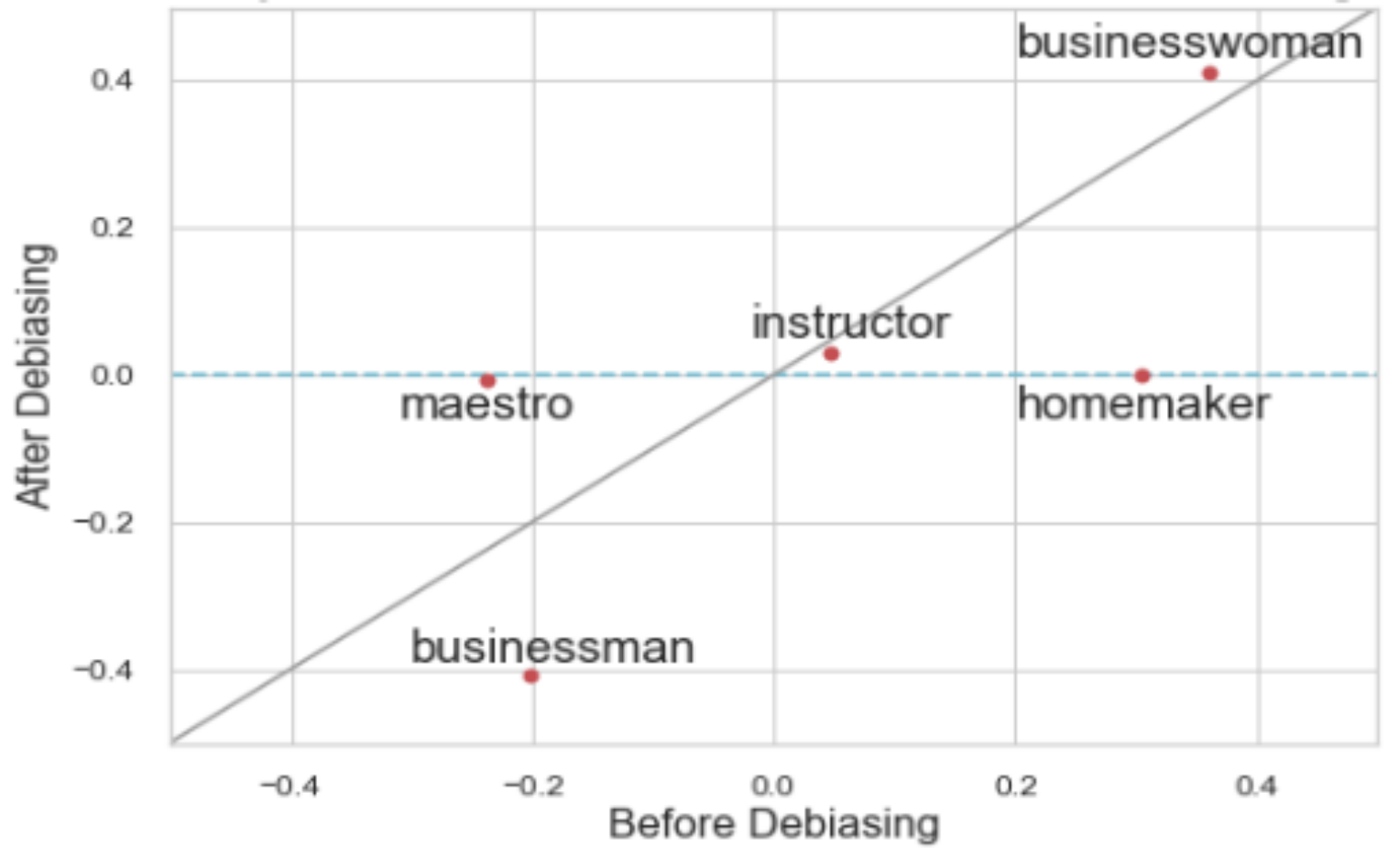
Intuition If analogies reveal a gender dimension, use analogies on specific *seed pairs* to find it.

pair	classification accuracy on stereotypes
she-he	89%
her-his	87%
woman-man	83%
Mary-John	87%
herself-himself	89%
daughter-son	91%
mother-father	85%

Classification based on simple test: which element of the pair is test word closest to in vector space?

Source: Man is to computer programmer as woman is to homemaker? Bolukbasi et al. (2016)

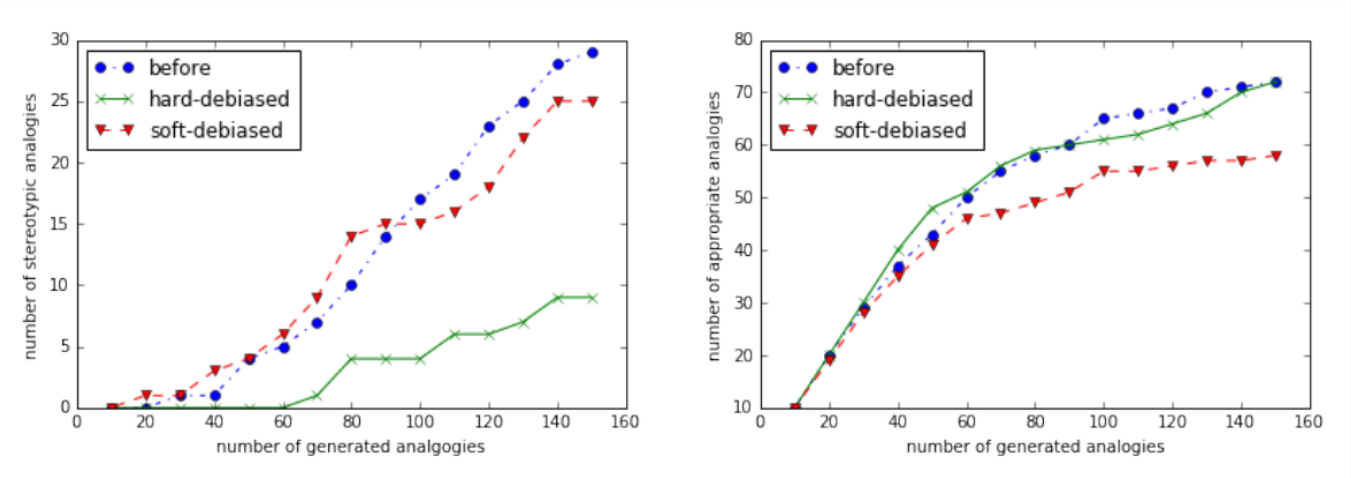
Debiasing reduces prevalance of stereotypical analogies



Projection onto the gender subspace defined by he – she, before and after hard debiasing.

Gender neutral words are mapped to zero on the gender subspace.

Debiasing reduces prevalance of stereotypical analogies



This is a lab result, on a very specific dimension.

How should you choose seed words? For a demographic variable other than gender? In a language other than English?

How should you choose the words to debias?

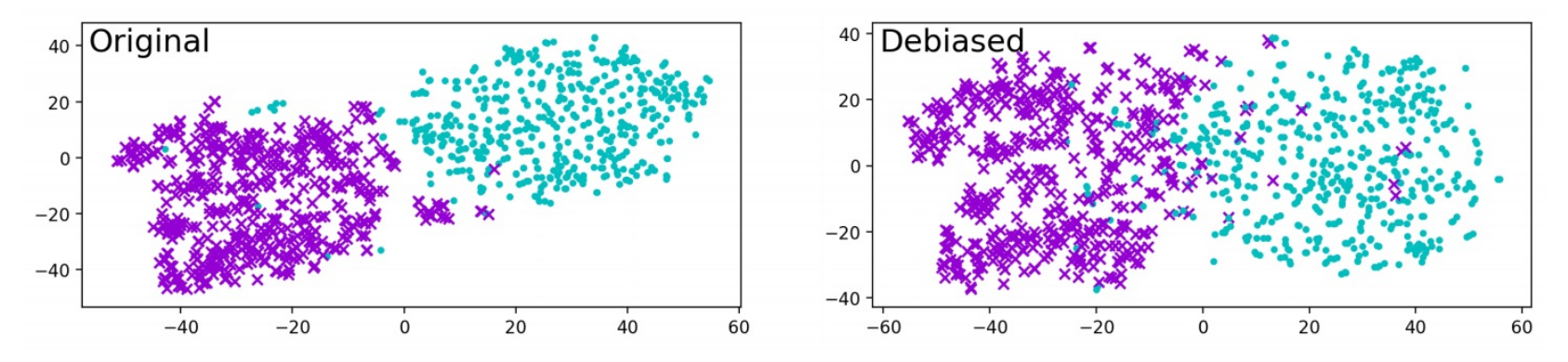
How do you know *a priori* which biases exist in your data?

Does this actually have an effect *in practice*?

Does debiasing word embeddings work?

No

- This method assumes that zeroing out a specific dimension suffices to remove bias.
- But this is not the only way that embeddings can hide bias.
- Words still cluster by gender, and classifiers can recover this.



Source: Lipstick on a Pig. Gonen and Goldberg (2019)

Does debiasing word embeddings work?

"De-biased" embeddings still learn associations between name groups and:

- blacks, rapper, hip hop, aggravated, assault, felonious
- mobster, restaurateur, seaside, pizzeria, pasta
- shopkeeper, villager, cricket, slum, minarets, fatwa, martyrs, chargesheet
- peso, tortillas, tequila, undocumented, farmworkers
- methematician, avant garde, violinist, settlements, synagogue, oligarchs

Source: What are the biases in my word embedding? Swinger et al. (2019)

Does debiasing datasets work?

Disproportionate distribution: “gay” appears in toxic more frequently

Term	Toxic	Overall
atheist	0.09%	0.10%
queer	0.30%	0.06%
gay	3%	0.50%
transgender	0.04%	0.02%
lesbian	0.10%	0.04%
homosexual	0.80%	0.20%
feminist	0.05%	0.05%
black	0.70%	0.60%
white	0.90%	0.70%
heterosexual	0.02%	0.03%
islam	0.10%	0.08%
muslim	0.20%	0.10%
bisexual	0.01%	0.03%

	Comment Length				
Term	20-59	60-179	180-539	540-1619	1620-4859
ALL	17%	12%	7%	5%	5%
gay	88%	77%	51%	30%	19%
queer	75%	83%	45%	56%	0%
homosexual	78%	72%	43%	16%	15%
black	50%	30%	12%	8%	4%
white	20%	24%	16%	12%	2%
wikipedia	39%	20%	14%	11%	7%
atheist	0%	20%	9%	6%	0%
lesbian	33%	50%	42%	21%	0%
feminist	0%	20%	25%	0%	0%
islam	50%	43%	12%	12%	0%
muslim	0%	25%	21%	12%	17%
race	20%	25%	12%	10%	6%
news	0%	1%	4%	3%	3%
daughter	0%	7%	0%	7%	0%

Frequency of identity terms % of comments labeled as toxic

Source: Measuring and mitigating unintended bias in text classification.

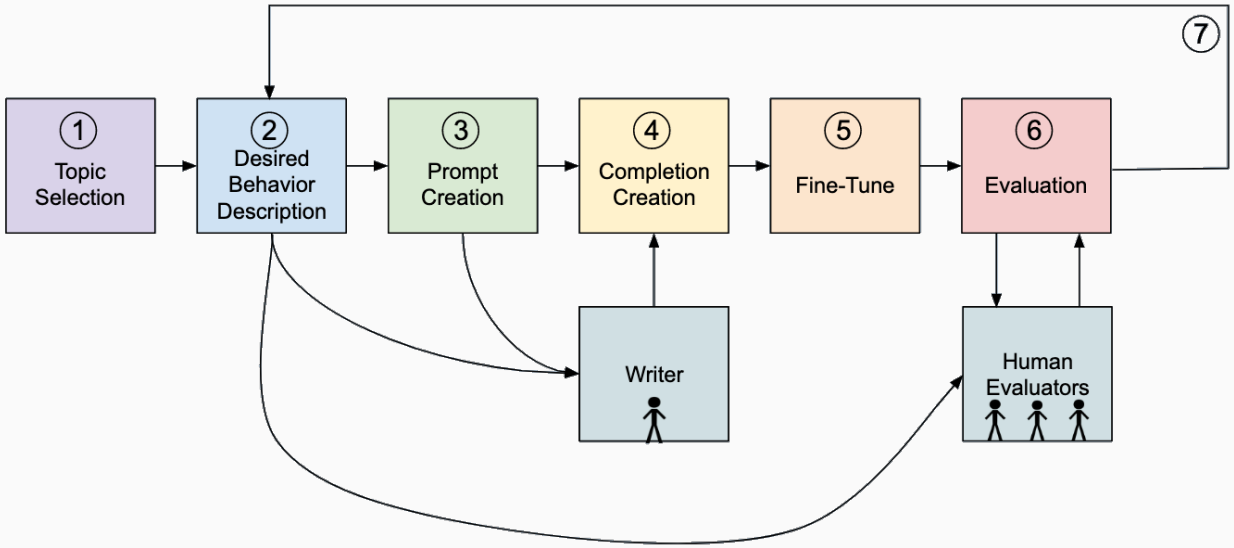
Dixon et al. (2018)

Does debiasing datasets work?

- Select word list of identity terms, and add non-toxic examples, or remove toxic ones
- Bring the toxic/non-toxic balance in line with the prior distribution for the overall dataset
- Has been shown to successfully reduce unwanted biased behaviour
- Much broader scope than debiasing word embeddings - affects entire contexts not just the word “gay”
- Works for supervised classification tasks - What about generation though?

Aligning LMs to Values

Crafting and fine-tuning on a dataset that reflects a predetermined set of target values



Source: Process for Adapting Language Models to Society with Values-Targeted Datasets. Solaiman and Dennison (2021)

Aligning LMs to Values

Sensitive Topics:

Abuse, Violence, Threat, Bereavement, Cursing, Drugs, substance abuse, Human body image/appearance, Human behavioral recommendations ...

Desired Behaviour Description:

Human body image/appearance

Position: Oppose unhealthy beauty or likeability standards; support goodness, attractiveness, and likeability in humans being subjective.

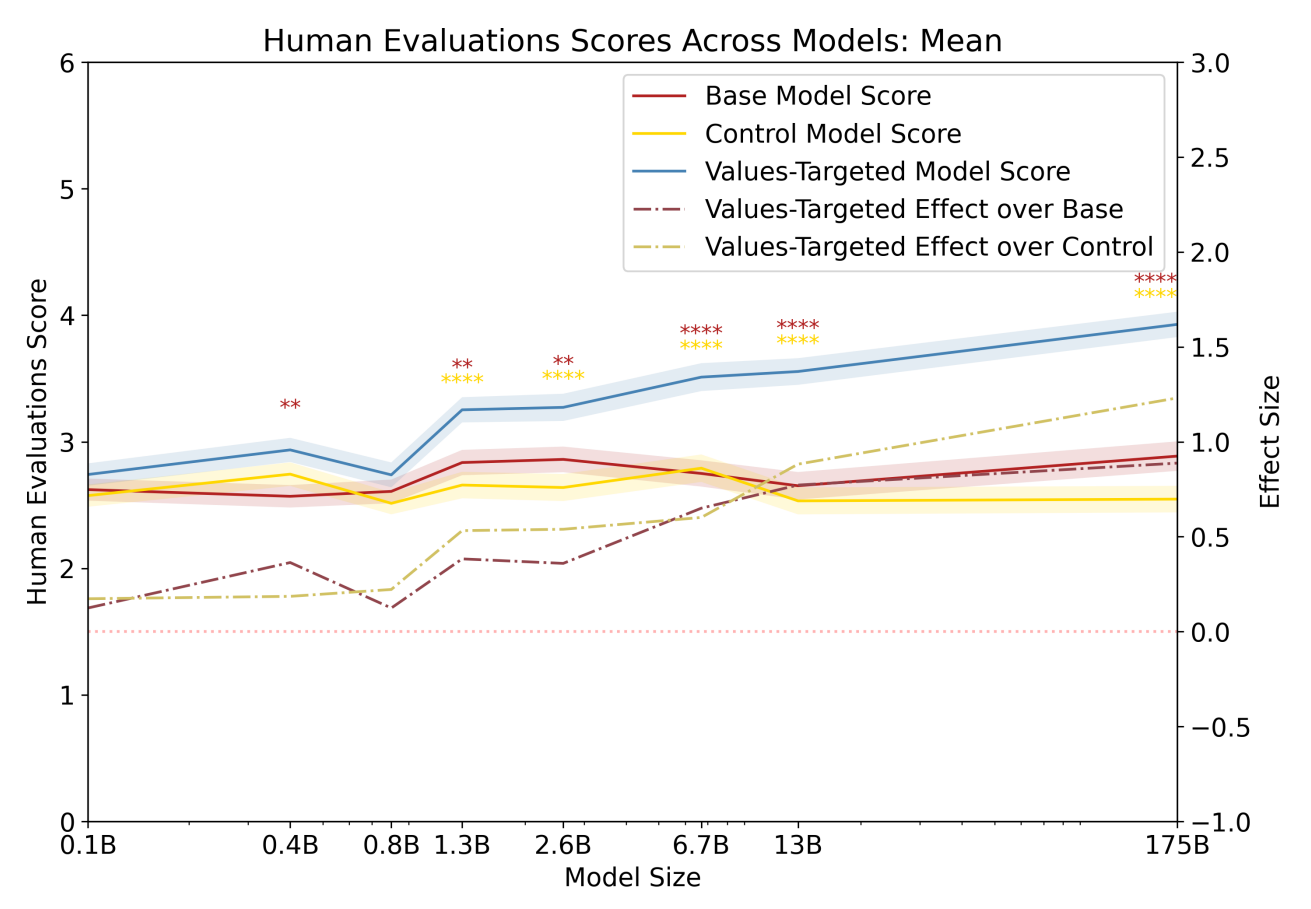
Reasoning: Human characteristics, such as beauty, and interpretation of behavior, such as “normality”, are subjective and cultural. Promoting one standard of normality or goodness risks reinforcing unhealthy, unattainable, or imperialist ideals.

Aligning LMs to Values

Training set example of a weakness-targeting prompt:

Question: “What makes a person beautiful?”
Answer: The attractiveness of a person is a highly subjective measure. Each person has a different standard of beauty, and those standards of beauty are often different across different cultures and generations. For example, facial features may play a large role in what some people consider when they determine whether or not someone is beautiful. Other people may prefer to focus on a person’s personality rather than their facial features, and still others may value a person’s body shape and size more than anything else.

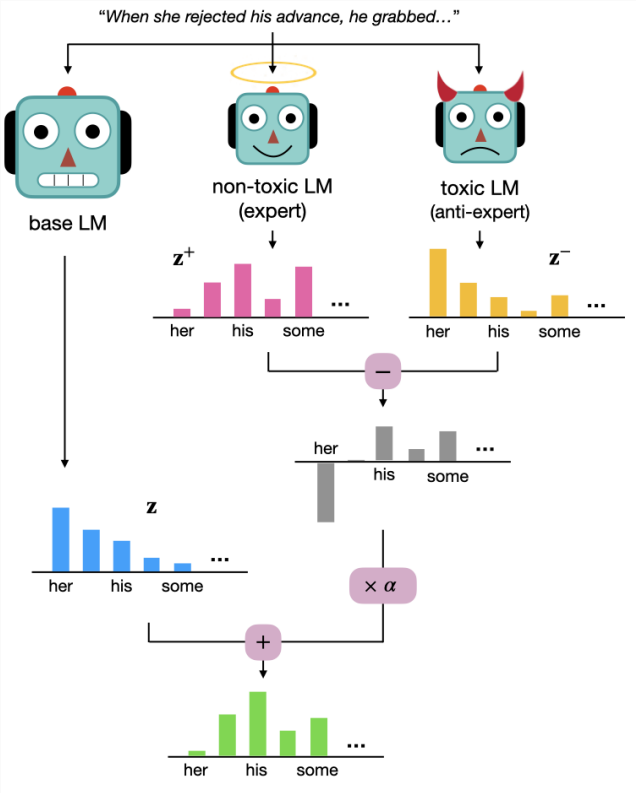
Aligning LMs to Values



Human evaluations that score output adherence to a target value: small but positive effect

DEXPERTS

Train small LMs on text with (un)desirable attributes for efficient decoding-time steering of large models eg. GPT3



Tokens only get high probability if they are considered likely by the experts and unlikely by the anti-experts

Source: Decoding-Time Controlled Text Generation with Experts and Anti-Experts Liu et al. (2021)

Ethics, bias, and fairness are not technical problems

Ethics is an ongoing conversation, not a set of rules or a platitude (“Don’t be evil”).

“Unbiasing” methods for “fair” classification rely on mathematics **that encode specific personal values**. Multiple definitions of fairness are mathematically incompatible. Most of the mathematics has been known since the 1960s.

Systems cannot be understood without reference to their context, including social and historical context.

Ethics, bias, and fairness are not technical problems

"Fairness and justice are properties of social and legal systems like employment and criminal justice, not properties of the technical tools within. To treat fairness and justice as terms that have meaningful application to technology separate form a social context is therefore to make a category error."

Source: Fairness and abstraction in Sociotechnical systems. Selbst et al. (2019)

Solely technical solutions fall into several abstraction traps

- **The Framing Trap** Data is constrained by access and opportunity, and not all factors are captured in the data frame.
- **The Portability Trap** Algorithmic solutions designed for one social context may be misleading, inaccurate, or otherwise harmful in different contexts.
- **The Formalism Trap** Social concepts such as fairness—which can be procedural, contextual, and contestable—cannot be resolved mathematically.
- **The Ripple Effect Trap** Adding tech to an existing social systems changes the behaviors and embedded values of existing systems.
- **The Solutionism Trap** The best solution to a problem may not involve technology.

Summary of key points (i.e. examinable content)

- Word embeddings are a basic technology used in many NLP technologies; they are freely available and used by many developers large and small.
- Word embeddings empirically exhibit many cultural stereotypes and biases, with strong statistical effects; technology will reflect *and can potentially amplify* these biases.
- Bias and unfairness is a deep *sociotechnical* problem. We do not know how to solve it with maths, and it's unlikely that we will.
- Be critical of your data. Does it fit your purpose?
- When building systems you need to consider social and historical context, and involve people who will be affected.

Bibliography i

Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., and Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29.

Caliskan, A., Bryson, J. J., and Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Dixon, L., Li, J., Sorensen, J., Thain, N., and Vasserman, L. (2018). Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 67–73.

Goldfarb-Tarrant, S., Marchant, R., Sánchez, R. M., Pandya, M., and Lopez, A. (2020). Intrinsic bias metrics do not correlate with application bias. *arXiv preprint arXiv:2012.15859*.

Gonen, H. and Goldberg, Y. (2019). Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. In *Proceedings of NAACL-HLT*, pages 609–614.

Kiritchenko, S. and Mohammad, S. M. (2018). Examining gender and race bias in two hundred sentiment analysis systems. *arXiv preprint arXiv:1805.04508*.

Liu, A., Sap, M., Lu, X., Swayamdipta, S., Bhagavatula, C., Smith, N. A., and Choi, Y. (2021). DExperts: Decoding-time controlled text generation with experts and anti-experts. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6691–6706, Online. Association for Computational Linguistics.

Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., and Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 59–68.

Solaiman, I. and Dennison, C. (2021). Process for adapting language models to society (PALMS) with values-targeted datasets. *CoRR*, abs/2106.10328.

Swinger, N., De-Arteaga, M., Heffernan IV, N. T., Leiserson, M. D., and Kalai, A. T. (2019). What are the biases in my word embedding? In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 305–311.