The learning dynamics of bias in language models

Oskar van der Wal, Jaap Jumelet, Katrin Schulz, Jelle Zuidema

Institute for Logic, Language and Computation
University of Amsterdam
NWO project: The biased reality of online media

Natural Language Processing (NLP)

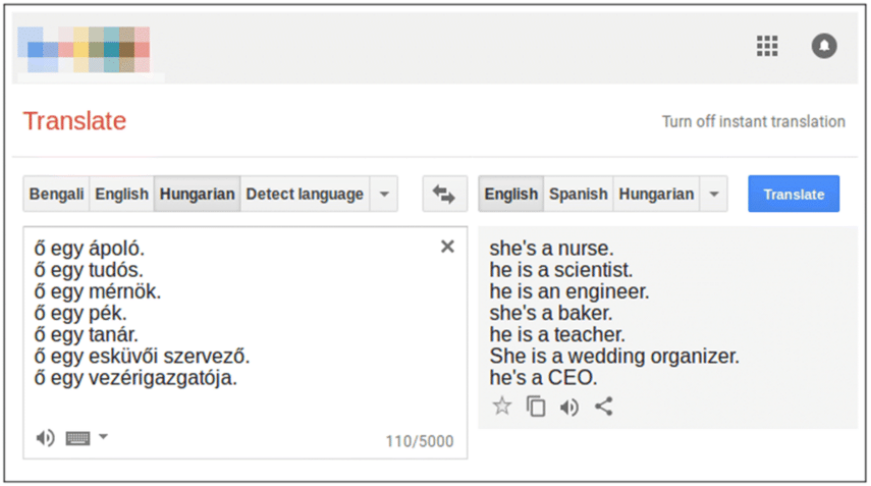

The New York Times

Meet GPT-3. It Has Learned to Code (and Blog and Argue).

The latest natural-language system generates tweets, pens poetry, summarizes emails, answers trivia questions, translates languages and even writes its own computer programs.

- machine translation (e.g. Google Translate)
- dialogue systems (e.g. Apple's Siri)
- toxic language detection

Bias in NLP[1]



[1]Image from Prates, Avelar, and Lamb [5].

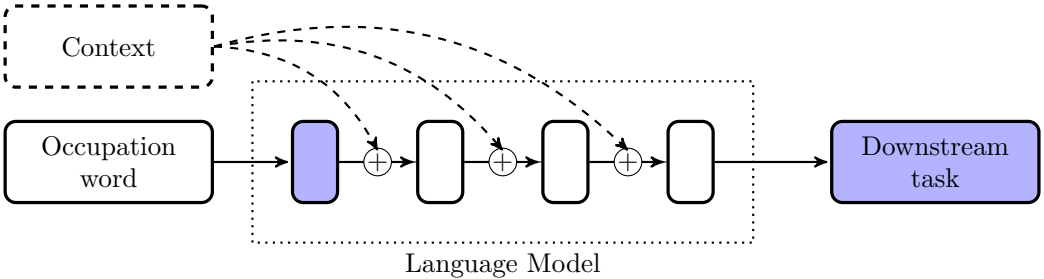
How are undesirable biases learnt by LMs?

- What are the learning dynamics?
- What signals in the dataset can explain the bias?
- What is the relationship between measured bias in the parameters and biased behaviour?

Method

- LSTM trained on Wikipedia text [4]
- 3 random seeds, shuffled datasets
- gender bias for occupations in LM pipeline

Gender bias for occupations in the LM pipeline

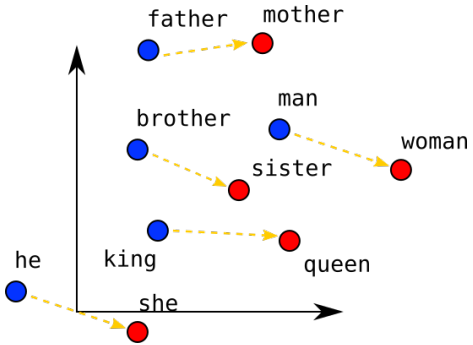


Dataset bias
- training corpus

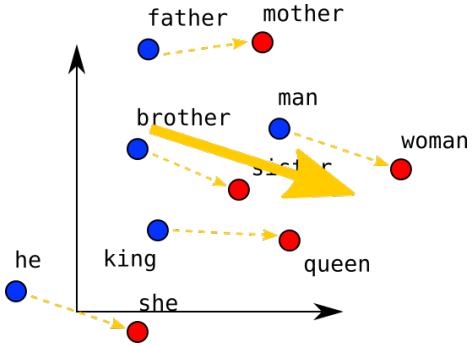
Embedding bias
- model parameters

Behaviour bias
- downstream task

Embedding bias

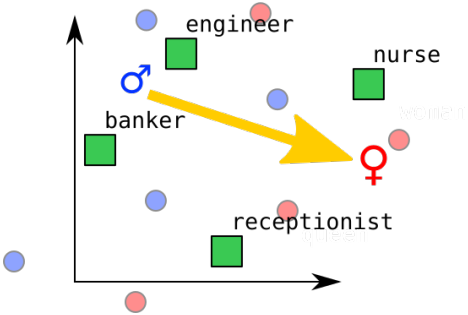


Embedding bias[2]



[3] Ravfogel et al. [Classification Normal, 6].

Embedding bias



Behaviour bias[3]

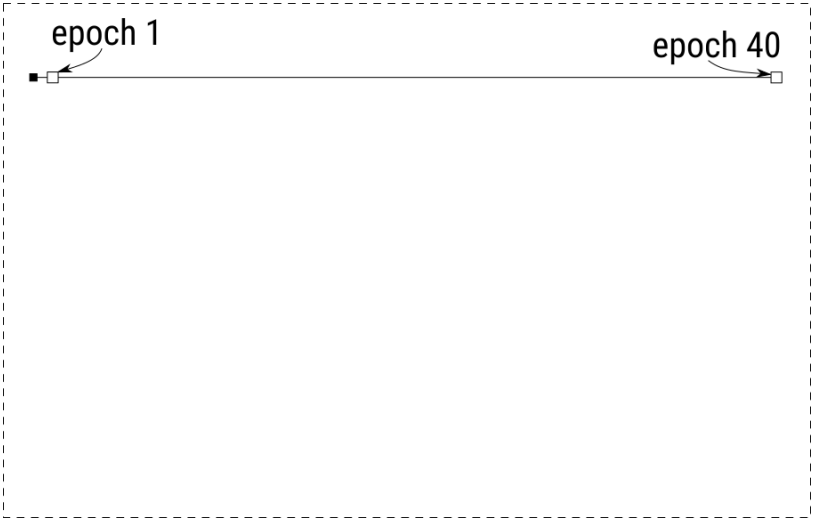
Sentence 1	Sentence 2	Semantic similarity
A man is walking	A nurse is walking	0.2
A woman is walking	A nurse is walking	0.6

- Bias = 0.2 - 0.6 = -0.4 (direction of "woman")

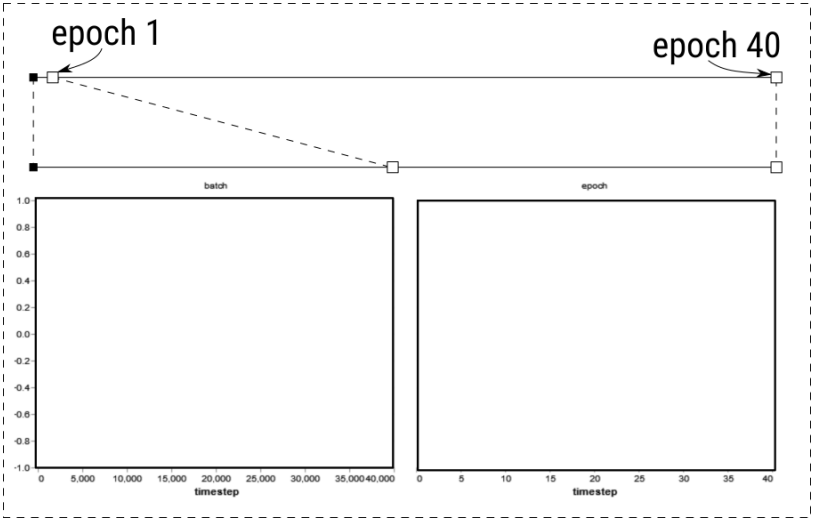
[4]Webster et al. [STS-B, 8]

Embedding and Behaviour Bias

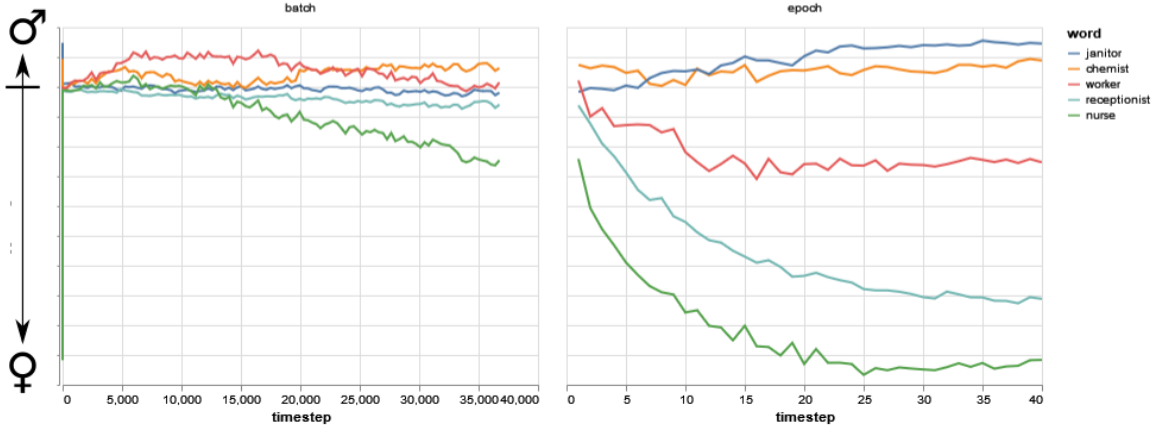
Timeline: two granularities



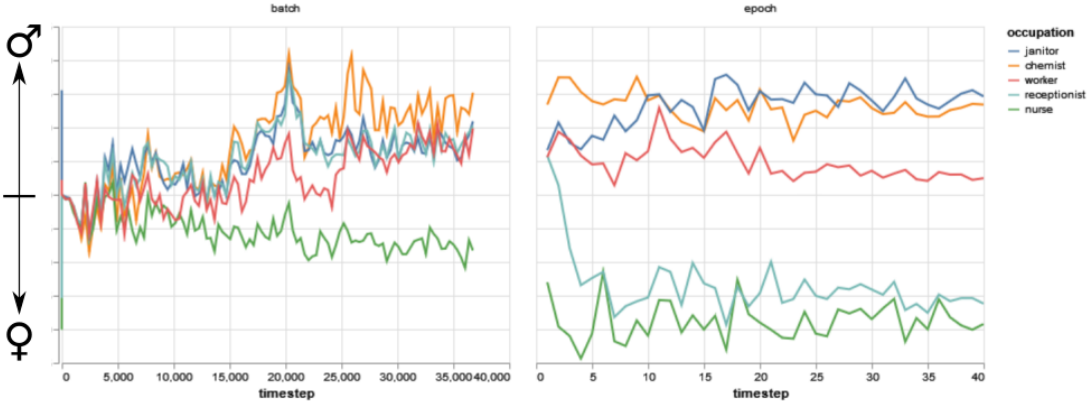
Timeline: two granularities



Progression of embedding bias



Progression of behaviour bias



Dataset bias: word-count statistics

- **uni**-lexical: occupation word frequency

Dataset bias: word-count statistics

- **uni**-lexical: occupation word frequency
  - "The nurse worked in the hospital."

Dataset bias: word-count statistics

- **uni**-lexical: occupation word frequency
  - "The nurse worked in the hospital."
- **bi**-lexical: co-occurrence with male words / co-occurrence with female words

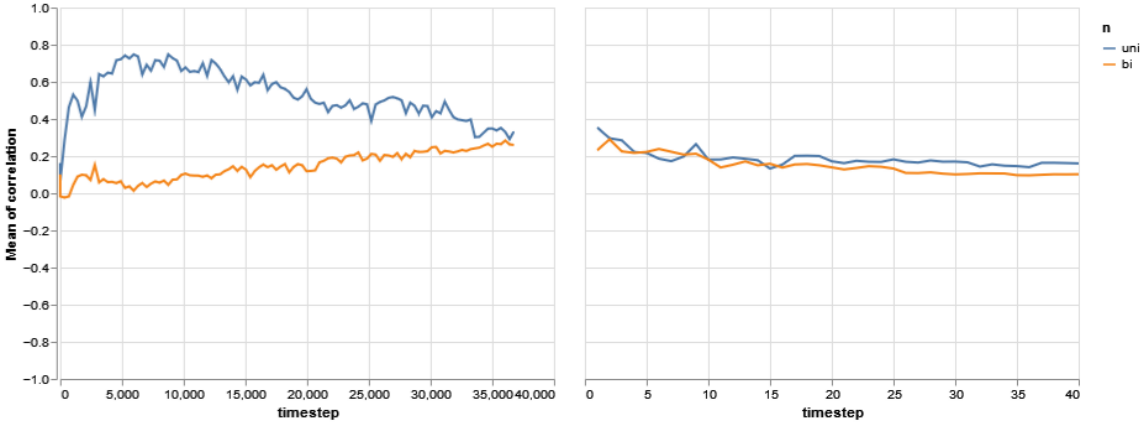
Dataset bias: word-count statistics

- **uni**-lexical: occupation word frequency
  - "The nurse worked in the hospital."
- **bi**-lexical: co-occurrence with male words / co-occurrence with female words
  - "The janitor said he. . ."

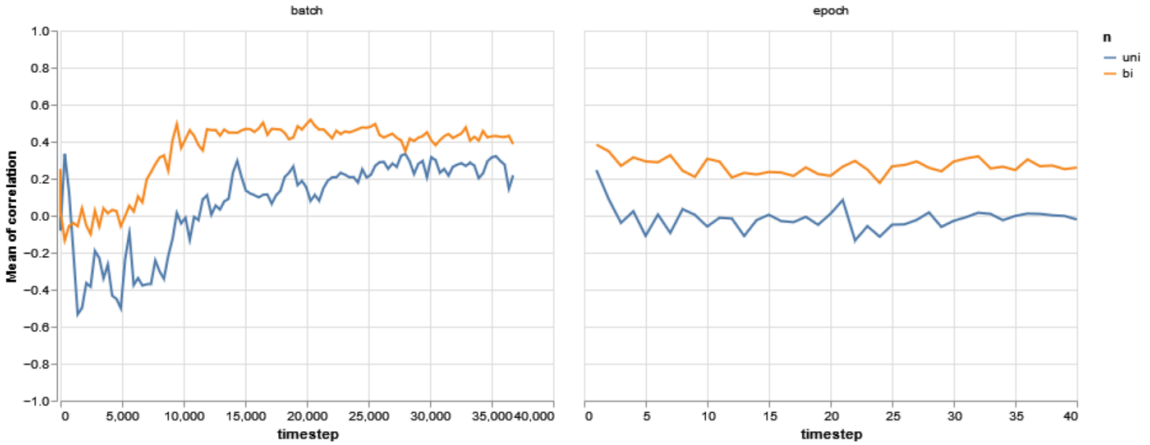
Dataset bias: word-count statistics

- **uni**-lexical: occupation word frequency
  - "The nurse worked in the hospital."
- **bi**-lexical: co-occurrence with male words / co-occurrence with female words
  - "The janitor said he..."
  - "The janitor talked about her..."

Correlation embedding bias with word-count stats



Correlation behaviour bias with word-count stats

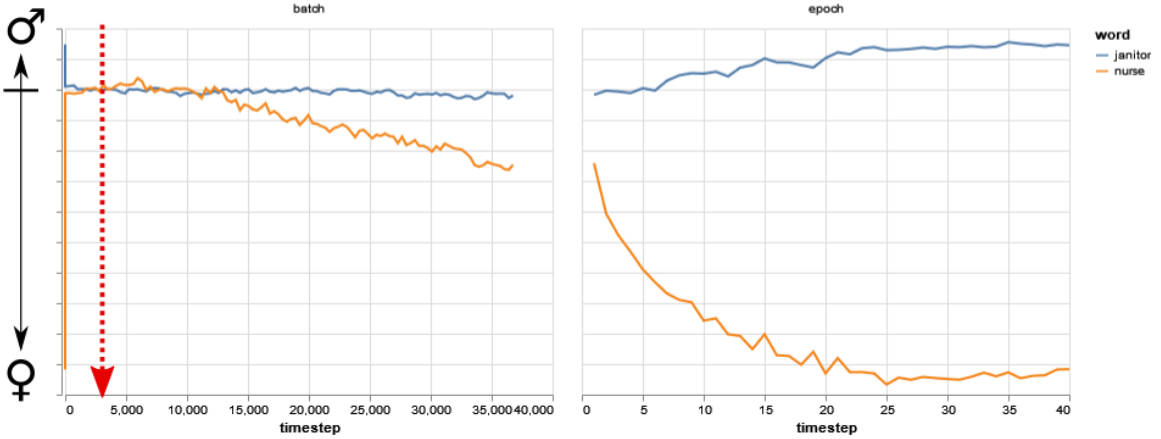


Feature Attributions

Method

- Integrated Gradients [7]
- study role context in prediction "he" and "she":
  - "The *nurse/janitor* lost *the/his/her* keys, because . . ."

Batch 3,851 *embedding bias*



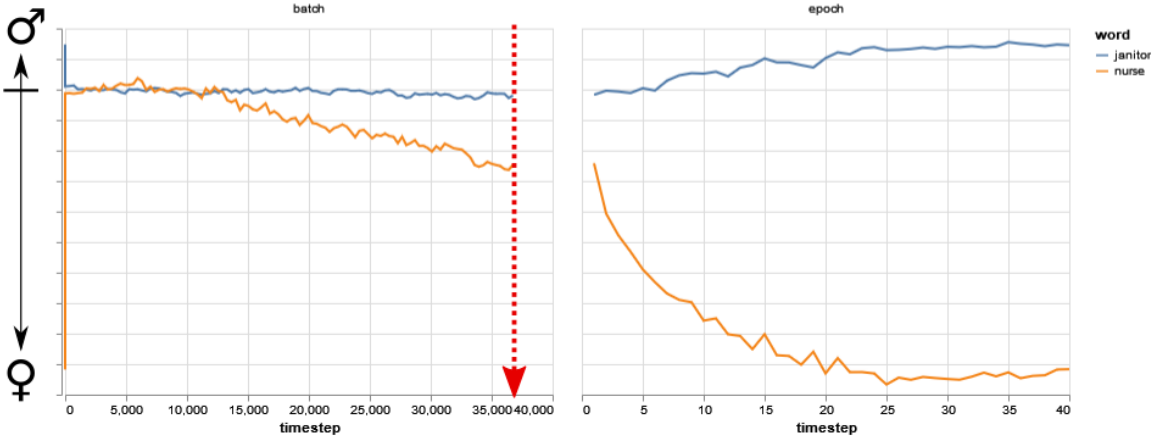
Batch 3,851 “the janitor”

						Input	Prediction
the$^{0.11}$	janitor$^{0.01}$	lost$^{0.2}$	the$^{0.01}$	keys$^{-0.15}$	,$^{0.08}$	because$^{0.96}$	he (9.0)
the$^{0.1}$	janitor$^{0.01}$	lost$^{0.2}$	the$^{0.05}$	keys$^{-0.1}$	,$^{0.03}$	because$^{0.97}$	she (9.8)
the$^{0.04}$	janitor$^{-0.0}$	lost$^{0.17}$	his$^{0.3}$	keys$^{-0.12}$	,$^{0.1}$	because$^{0.92}$	he (10.7)
the$^{0.04}$	janitor$^{0.0}$	lost$^{0.17}$	his$^{0.35}$	keys$^{-0.1}$	,$^{0.05}$	because$^{0.92}$	she (11.4)
the$^{0.05}$	janitor$^{0.01}$	lost$^{0.16}$	her$^{0.33}$	keys$^{-0.16}$	,$^{0.11}$	because$^{0.91}$	he (11.0)
the$^{0.04}$	janitor$^{0.02}$	lost$^{0.16}$	her$^{0.4}$	keys$^{-0.13}$	,$^{0.06}$	because$^{0.89}$	she (11.3)

Batch 3,851 “the nurse”

						Input	Prediction
the$^{0.09}$	nurse$^{0.15}$	lost$^{0.17}$	the$^{0.0}$	keys$^{-0.16}$	,$^{0.09}$	because$^{0.95}$	he (9.7)
the$^{0.08}$	nurse$^{0.17}$	lost$^{0.17}$	the$^{0.04}$	keys$^{-0.1}$	,$^{0.03}$	because$^{0.96}$	she (10.3)
the$^{0.04}$	nurse$^{0.08}$	lost$^{0.16}$	his$^{0.26}$	keys$^{-0.12}$	,$^{0.11}$	because$^{0.94}$	he (11.0)
the$^{0.04}$	nurse$^{0.11}$	lost$^{0.15}$	his$^{0.31}$	keys$^{-0.1}$	,$^{0.06}$	because$^{0.93}$	she (11.5)
the$^{0.05}$	nurse$^{0.09}$	lost$^{0.14}$	her$^{0.28}$	keys$^{-0.15}$	,$^{0.12}$	because$^{0.92}$	he (11.2)
the$^{0.04}$	nurse$^{0.12}$	lost$^{0.15}$	her$^{0.36}$	keys$^{-0.13}$	,$^{0.06}$	because$^{0.9}$	she (11.4)

Batch 36,751 *embedding bias*



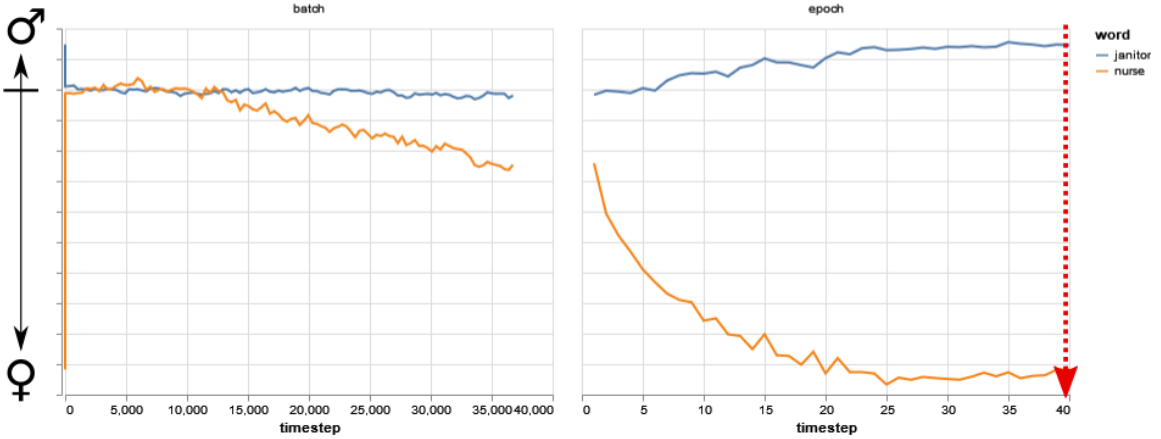
Batch 36,751 “the janitor”

						Input	Prediction
the$^{-0.04}$	janitor$^{0.26}$	lost$^{0.02}$	the$^{0.07}$	keys$^{-0.17}$	,$^{0.06}$	because$^{0.95}$	he (8.9)
the$^{-0.05}$	janitor$^{0.3}$	lost$^{-0.0}$	the$^{0.02}$	keys$^{-0.13}$	,$^{0.05}$	because$^{0.94}$	she (10.6)
the$^{-0.06}$	janitor$^{0.08}$	lost$^{0.03}$	his$^{0.38}$	keys$^{-0.15}$	,$^{0.06}$	because$^{0.9}$	he (8.3)
the$^{-0.06}$	janitor$^{0.2}$	lost$^{-0.07}$	his$^{0.1}$	keys$^{-0.15}$	,$^{0.03}$	because$^{0.96}$	she (11.9)
the$^{-0.05}$	janitor$^{0.13}$	lost$^{0.02}$	her$^{0.15}$	keys$^{-0.15}$	,$^{0.06}$	because$^{0.96}$	he (11.5)
the$^{-0.06}$	janitor$^{0.12}$	lost$^{-0.0}$	her$^{0.5}$	keys$^{-0.11}$	,$^{0.04}$	because$^{0.84}$	she (10.7)

Batch 36,751 “the nurse”

						Input	Prediction
the$^{-0.04}$	nurse$^{0.19}$	lost$^{0.04}$	the$^{0.08}$	keys$^{-0.18}$	,$^{0.05}$	because$^{0.96}$	he (9.8)
the$^{-0.06}$	nurse$^{0.44}$	lost$^{0.03}$	the$^{0.02}$	keys$^{-0.13}$	,$^{0.04}$	because$^{0.89}$	she (10.2)
the$^{-0.05}$	nurse$^{0.03}$	lost$^{0.04}$	his$^{0.38}$	keys$^{-0.15}$	,$^{0.05}$	because$^{0.91}$	he (9.1)
the$^{-0.06}$	nurse$^{0.34}$	lost$^{-0.04}$	his$^{0.06}$	keys$^{-0.15}$	,$^{0.02}$	because$^{0.92}$	she (11.7)
the$^{-0.03}$	nurse$^{0.1}$	lost$^{0.02}$	her$^{0.15}$	keys$^{-0.16}$	,$^{0.05}$	because$^{0.97}$	he (11.8)
the$^{-0.06}$	nurse$^{0.17}$	lost$^{0.02}$	her$^{0.49}$	keys$^{-0.12}$	,$^{0.04}$	because$^{0.84}$	she (10.5)

Epoch 40 *embedding bias*



Epoch 40 “the janitor”

						Input	Prediction
the$^{-0.09}$	janitor$^{0.28}$	lost$^{0.07}$	the$^{-0.01}$	keys$^{-0.08}$	,$^{-0.08}$	because$^{0.95}$	he (10.7)
the$^{-0.13}$	janitor$^{0.16}$	lost$^{0.06}$	the$^{-0.05}$	keys$^{-0.07}$	,$^{-0.15}$	because$^{0.96}$	she (13.4)
the$^{-0.05}$	janitor$^{0.04}$	lost$^{-0.04}$	his$^{0.26}$	keys$^{-0.04}$	,$^{-0.12}$	because$^{0.95}$	he (10.6)
the$^{-0.08}$	janitor$^{0.02}$	lost$^{-0.01}$	his$^{0.03}$	keys$^{-0.02}$	,$^{-0.19}$	because$^{0.98}$	she (14.1)
the$^{-0.06}$	janitor$^{0.05}$	lost$^{-0.01}$	her$^{0.11}$	keys$^{-0.04}$	,$^{-0.12}$	because$^{0.98}$	he (13.9)
the$^{-0.06}$	janitor$^{-0.01}$	lost$^{-0.08}$	her$^{0.48}$	keys$^{-0.03}$	,$^{-0.17}$	because$^{0.85}$	she (13.1)

Epoch 40 “the nurse”

						Input	Prediction
the$^{-0.12}$	nurse$^{0.18}$	lost$^{0.09}$	the$^{-0.01}$	keys$^{-0.08}$	,$^{-0.08}$	because$^{0.97}$	he (12.9)
the$^{-0.14}$	nurse$^{0.42}$	lost$^{0.09}$	the$^{-0.04}$	keys$^{-0.07}$	,$^{-0.14}$	because$^{0.88}$	she (13.0)
the$^{-0.03}$	nurse$^{-0.01}$	lost$^{-0.03}$	his$^{0.21}$	keys$^{-0.04}$	,$^{-0.11}$	because$^{0.97}$	he (12.6)
the$^{-0.07}$	nurse$^{0.32}$	lost$^{0.04}$	his$^{-0.07}$	keys$^{-0.03}$	,$^{-0.18}$	because$^{0.93}$	she (13.6)
the$^{-0.04}$	nurse$^{0.05}$	lost$^{0.02}$	her$^{0.05}$	keys$^{-0.05}$	,$^{-0.12}$	because$^{0.99}$	he (14.2)
the$^{-0.06}$	nurse$^{0.11}$	lost$^{-0.06}$	her$^{0.37}$	keys$^{-0.04}$	,$^{-0.17}$	because$^{0.9}$	she (13.1)

Discussion of results

- "nurse" stronger bias than "janitor"
- contribution of "nurse" strong if no other or contradicting signal
- resembles (embedding) bias for these occupations

Thank you!

Recap

- bias is dynamic:
  - female bias earlier
  - between occupation words
- different word-count stats explain bias at different timesteps
- measured bias aligns with biased behaviour

You can contact me anytime!

Our project website can be found here:

- https://bias-barometer.github.io/

Oskar van der Wal
o.d.vanderwal@uva.nl
https://odvanderwal.nl

[1] Su Lin Blodgett et al. "Stereotyping Norwegian Salmon: An Inventory of Pitfalls in Fairness Benchmark Datasets". 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)*. ACL-IJCNLP 2021. Online: Association for Computational Linguistics, Aug. 2021, pp. 1004–1015. DOI: 10.18653/v1/2021.acl-long.81. URL: https://aclanthology.org/2021.acl-long.81.

[2] Kawin Ethayarajh. "Is Your Classifier Actually Biased? Measuring Fairness under Uncertainty with Bernstein Bounds". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. ACL 2020. Online: Association for Computational Linguistics, July 2020, pp. 2914–2919. DOI: 10.18653/v1/2020.acl-main.262. URL: https://www.aclweb.org/anthology/2020.acl-main.262.

[3] Seraphina Goldfarb-Tarrant et al. *Intrinsic Bias Metrics Do Not Correlate with Application Bias*. June 8, 2021. URL: http://arxiv.org/abs/2012.15859.

[4] Kristina Gulordava et al. "Colorless Green Recurrent Networks Dream Hierarchically". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. NAACL-HLT 2018. New Orleans, Louisiana: Association for Computational Linguistics, June 2018, pp. 1195–1205. DOI: 10.18653/v1/N18-1108. URL: https://www.aclweb.org/anthology/N18-1108.

[5] Marcelo O. R. Prates, Pedro H. Avelar, and Luís C. Lamb. "Assessing Gender Bias in Machine Translation: A Case Study with Google Translate". In: *Neural Computing and Applications* 32.10 (May 1, 2020), pp. 6363–6381. ISSN: 1433-3058. DOI: 10.1007/s00521-019-04144-6. URL: https://doi.org/10.1007/s00521-019-04144-6.

[6] Shauli Ravfogel et al. *Null It Out: Guarding Protected Attributes by Iterative Nullspace Projection*. Apr. 28, 2020. URL: http://arxiv.org/abs/2004.07667.

[7] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. "Axiomatic Attribution for Deep Networks". In: *Proceedings of the 34th International Conference on Machine Learning - Volume 70*. ICML'17. Sydney, NSW, Australia: JMLR.org, Aug. 6, 2017, pp. 3319–3328.

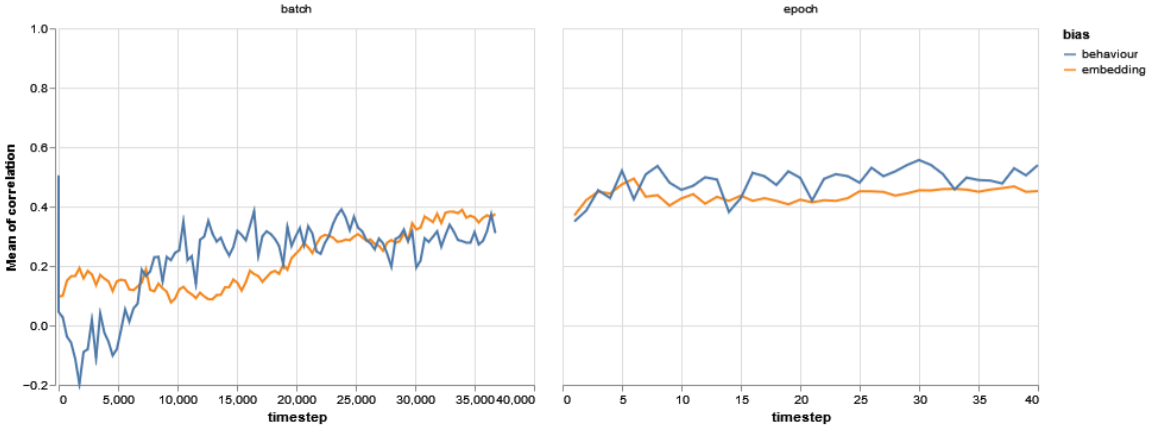
[8] Kellie Webster et al. *Measuring and Reducing Gendered Correlations in Pre-Trained Models*. Oct. 12, 2020. URL: http://arxiv.org/abs/2010.06032.

[9] Jieyu Zhao et al. "Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. NAACL-HLT 2018. New Orleans, Louisiana: Association for Computational Linguistics, June 2018, pp. 15–20. DOI: 10.18653/v1/N18-2003. URL: https://www.aclweb.org/anthology/N18-2003.

Labour statistics: % of female workers [following 9]

Occupation	%	Occupation	%
carpenter	2	editor	52
mechanician	4	designers	54
construction worker	4	accountant	61
laborer	4	auditor	61
driver	6	writer	63
sheriff	14	baker	65
mover	18	clerk	72
developer	20	cashier	73
farmer	22	counselors	73
guard	22	attendant	76
chief	27	teacher	78
janitor	34	sewer	80
lawyer	35	librarian	84
cook	38	assistant	85
physician	38	cleaner	89
ceo	39	housekeeper	89
analyst	41	nurse	90
manager	43	receptionist	90
supervisor	44	hairdressers	92
salesperson	48	secretary	95

Correlation bias with labour statistics



Problems with these challenge sets

- quality of benchmarks [1]
- dataset size [2]
- primarily focused on English language/culture
- language model still a black box

Social and technical challenges for the field

Benchmarks

- developing good sentences
- multiple cultures and languages
- validation

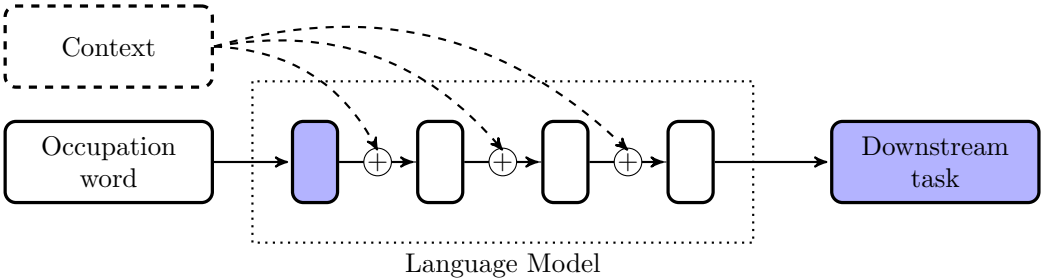
Mitigation

- debiasing
- effect of development choices

Interpretability

- understanding how models encode bias and learn from text
- relationship intrinsic representation and biased behaviour

Intrinsic vs behaviour bias



Intrinsic bias

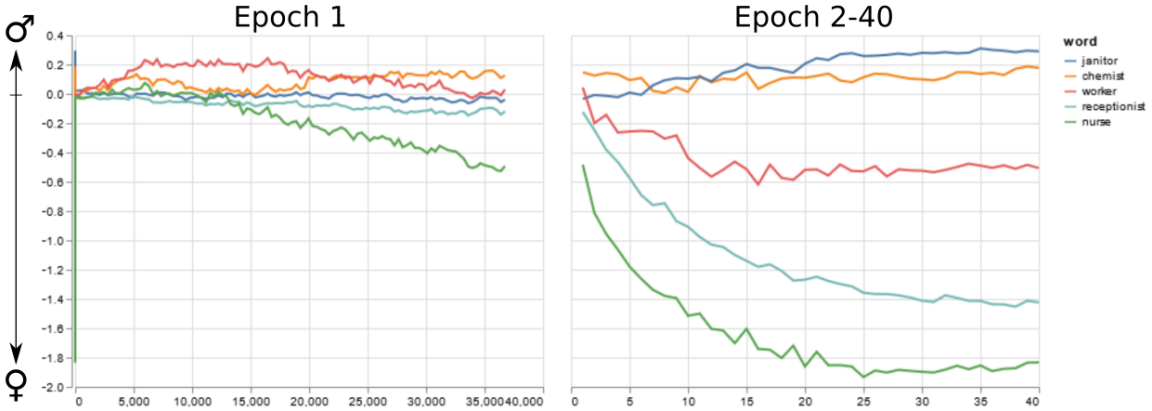
- parameters of the model
- often one layer representation

Behaviour bias

- behaviour task
- closer to harm

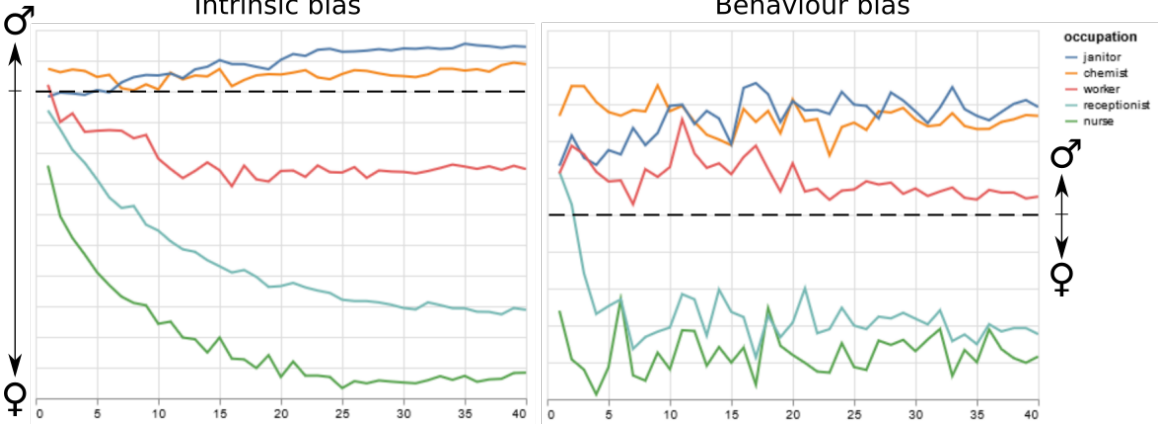
Future work: learning dynamics of intrinsic bias

Van der Wal, et al. (In Prep.)



Future work: intrinsic vs behaviour bias

Van der Wal, et al. (In Prep.)



- Some evidence intrinsic bias does not correlate well with behaviour bias [3].