cases

doc_1		doc_2		decision	id
authors	• Oskar van der Wal • Jaap Jumelet • Katrin Schulz • Willem Zuidema	**authors**	• Oskar van der Wal • Jaap Jumelet • K. Schulz • Willem H. Zuidema	DUPLICATES	125
title	The Birth of Bias: A case study on the evolution of gender bias in an English language model	**title**	The Birth of Bias: A case study on the evolution of gender bias in an English language model		
publication_date	2022-07-21 00:59:04+00:00	**publication_date**	2022-07-21 00:00:00		
source	SupportedSources.ARXIV	**source**	SupportedSources.SEMANTIC_SCHOLAR		
journal	None	**journal**	ArXiv		
volume		**volume**	abs/2207.10245		
doi		**doi**	10.48550/arXiv.2207.10245		
urls	• http://arxiv.org/pdf/2207.10245v1 • http://arxiv.org/abs/2207.10245v1 • http://arxiv.org/pdf/2207.10245v1	**urls**	• https://www.semanticscholar.org/paper/d6ddc4f4c81c9565019be1983d37f9fbdf5bd057		
id	id-5511724643078565877	**id**	id7697536614282111548		
abstract	Detecting and mitigating harmful biases in modern language models are widely recognized as crucial, open problems. In this paper, we take a step back and investigate how language models come to be biased in the first place. We use a relatively small language model, using the LSTM architecture trained on an English Wikipedia corpus. With full access to the data and to the model parameters as they change during every step while training, we can map in detail how the representation of gender develops, what patterns in the dataset drive this, and how the model's internal state relates to the bias in a downstream task (semantic textual similarity). We find that the representation of gender is dynamic and identify different phases during training. Furthermore, we show that gender information is represented increasingly locally in the input embeddings of the model and that, as a consequence, debiasing these can be effective in reducing the downstream bias. Monitoring the training dynamics, allows us to detect an asymmetry in how the female and male gender are represented in the input embeddings. This is important, as it may cause naive mitigation strategies to introduce new undesirable biases. We discuss the relevance of the findings for mitigation strategies more generally and the prospects of generalizing our methods to larger language models, the Transformer architecture, other languages and other undesirable biases.	**abstract**	Detecting and mitigating harmful biases in modern language models are widely recognized as crucial, open problems. In this paper, we take a step back and investigate how language models come to be biased in the first place.We use a relatively small language model, using the LSTM architecture trained on an English Wikipedia corpus. With full access to the data and to the model parameters as they change during every step while training, we can map in detail how the representation of gender develops, what patterns in the dataset drive this, and how the modelâ€™s internal state relates to the bias in a downstream task (semantic textual similarity).We find that the representation of gender is dynamic and identify different phases during training.Furthermore, we show that gender information is represented increasingly locally in the input embeddings of the model and that, as a consequence, debiasing these can be effective in reducing the downstream bias.Monitoring the training dynamics, allows us to detect an asymmetry in how the female and male gender are represented in the input embeddings. This is important, as it may cause naive mitigation strategies to introduce new undesirable biases.We discuss the relevance of the findings for mitigation strategies more generally and the prospects of generalizing our methods to larger language models, the Transformer architecture, other languages and other undesirable biases.		
versions		**versions**			