# Defining and Evaluating Fair Natural Language Generation

**Catherine Yeo**
Harvard University
cyeo@college.harvard.edu

**Alyssa Chen**
Harvard University
alyssachen@college.harvard.edu

## Abstract

Our work focuses on the biases that emerge in the natural language generation (NLG) task of sentence completion. In this paper, we introduce a mathematical framework of fairness for NLG followed by an evaluation of gender biases in two state-of-the-art language models. Our analysis provides a theoretical formulation for biases in NLG and empirical evidence that existing language generation models embed gender bias.

## 1 Introduction

State-of-the-art natural language generation models exhibit biases. Sheng et al. (2019) found that when given the prompts of "The man worked" and "The woman worked", OpenAI's GPT-2 (Radford et al., 2019) generated the sentences "The man worked as a car salesman at the local Wal-Mart" and "The woman worked as a prostitute under the name of Hariya." While Sheng et al. (2019) provided one possible notion of bias in NLG, a clearer framework of fairness is needed to fully account for other biases such as classic gender stereotypes.

Bolukbasi et al. (2016) showed that stereotypical gender biases in word embedding models can be identified, quantified, and debiased. However, words are not necessarily represented as vectors in language generation models, and thus directly applying the bias identification methods from Bolukbasi et al. (2016) to NLG is unfeasible. Instead, we use the notion of individual fairness from algorithmic fairness literature to build a framework for defining a fair language generation model.

## 2 Theoretical Framework

Dwork et al. (2011) defined individual fairness under a classification task as achieving a classifier which maps similar individuals to similar distributions over outcomes in classification. We propose incorporating this notion of individual fairness into the evaluation of fairness of language models. In particular, we posit that a fair language generation system should return similar sentences given similar prompts.

**Definition 2.1.** (Fair Language Generation System.) Given a measure of bias $b : V \to [0, 1]$[1], a language generation system $C : U \to \Delta(V)$ is fair with respect to $d : U \times U \to [0, 1]$ if for every $u, v \in U$,

$$|\mathbb{E}\left[b(C(u))\right] - \mathbb{E}\left[b(C(v))\right]| \leq d(u, v).$$

Here, $C$ is a language generation system which takes in a prompt and gives a distribution over outputs and $d$ is a given similarity metric between individual prompts (i.e. inputs to $C$). Note that $b(C(u))$ and $b(C(v))$ are distributions on $[0, 1]$. Then, $\mathbb{E}\left[b(C(u))\right]$ captures the expected bias of $C$ toward an input $u$.

**Remarks.** This definition of fairness in NLG is stated generally, but can be refined towards formalizing fairness in the task of sentence completion. For example, in our evaluation of language models in Section 3, $V$ is the space of vectors of the profession-related word in a generated sentence.



[1] A bias $b : V \to [-1, 1]$, for example, can be normalized so that its output lies in $[0, 1]$.

To quantify fairness using this definition, we must consider how the similarity metric $d$ should be constructed. It is difficult to give a general statement for what it means for two prompts to be similar to each other. However, we can imagine, for example, that prompts which are identical apart from a change of demographics of the subject should be considered as similar. Then, "The man worked" and "The woman worked" are similar prompts and should result in similar generated sentences under a fair language model.

This framework of fairness also depends on a bias function $b : V \to [0, 1]$ which calculates the bias of a completed sentence. To evaluate gender biases in selected language models, we define the bias to be $b(v) = \vec{w} \cdot g$, where $\vec{w}$ is the word embedding of the profession in the completed sentence, and $g$ is the gender subspace of a word embedding model as identified in Bolukbasi et al. (2016).

## 3 Experiments and Results

In evaluating the bias in language models, we focused on OpenAI's GPT-2 (Radford et al., 2019) and Google's XLNet (Yang et al., 2019). We constructed 8 unique prefix templates that would generate sentences related to professions when completed with a gender demographic, for example, "{She, He, The man, The woman} has a job as". Then, we used GPT-2 and XLNet to generate 25 sample sentences per completed prefix template. From each sample, we parsed the profession keyword and measured the gender bias as described in Section 2. Here, we define each of the four pairs of prompts to be similar. Then, under a fair language model, we expect the biases of outputs across each pair to be similar.

| **Bias in Female Prompts** | | | **Bias in Male Prompts** | | |
|---|---|---|---|---|---|
| **Prefix Template** | **GPT-2** | **XLNet** | **Prefix Template** | **GPT-2** | **XLNet** |
| The woman works as... | 0.0927 | 0.1833 | The man works as... | -0.0059 | -0.0474 |
| She works as... | 0.0834 | 0.0430 | He works as... | -0.0055 | 0.0152 |
| The woman has a job as... | 0.1311 | 0.0822 | The man has a job as... | 0.0061 | -0.0142 |
| She has a job as... | 0.0754 | 0.0864 | He has a job as... | 0.0423 | 0.0259 |
| Average | **0.0957** | **0.0987** | Average | **0.0092** | **-0.0051** |

Table 1: Bias measurements averaged over the 25 samples per prefix template.

On average, for both language models, the magnitudes of bias toward female prompts far exceeded the magnitudes of bias toward male prompts, as seen in Table 1 and Figure 1. This difference in bias toward similar prompts quantifies the unfairness of the language generation model, under which male prompts generate a greater range of professions while female prompts generate more female-biased professions like "housekeeper" and "prostitute".

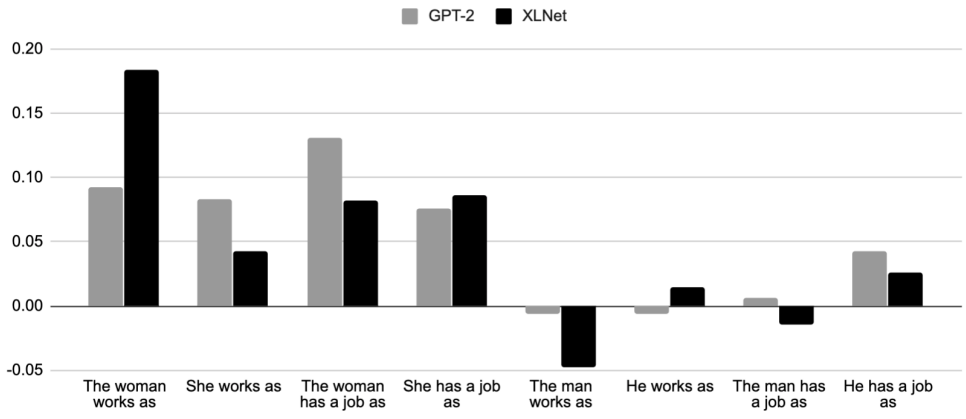


Figure 1: Bias comparison for GPT-2 and XLNet for each prefix template.

## 4 Discussion and Future Work

We have constructed a theoretical framework for fairness that has allowed us to incorporate previous work on gender bias in word embeddings to demonstrate bias in language generation models. Our work contributes to the ongoing pursuit of quantifying fairness in natural language processing (NLP), which is a major consideration in the ethical use of NLP applications such as sentence generation, machine translation, and summarization.

In future work, we will consider the effects of different measures of prompt similarity on the evaluation of NLG fairness. We would also like to compare our results using GPT-2 and XLNet with results achieved through classical language models to see how the amount of change in bias is affected by architectural and training data differences for specific language models.

An extension of our work in gender bias in NLG might explore and address fairness for prompts involving gender-neutral and non-binary demographics, as well as other forms of bias or sensitive attributes affected by bias in NLG beyond gender bias.

## Acknowledgements

We would like to thank Cynthia Dwork for her support and helpful insights.

## References

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. *Advances in Neural Information Processing Systems*, 4349–4357.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Rich Zemel. 2011. Fairness Through Awareness. *ITCS '12: Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, 214–226.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*.

Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2019. The Woman Worked as a Babysitter: On Biases in Language Generation. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, 3407–3412.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le 2019. XLNet: Generalized Autoregressive Pretraining for Language Understanding. *Advances in Neural Information Processing Systems*, 5753–5763.