Gender and Representation Bias in GPT-3 Generated Stories

Li Lucy
University of California, Berkeley
lucy3_li@berkeley.edu

David Bamman
University of California, Berkeley
dbamman@berkeley.edu

Abstract

Using topic modeling and lexicon-based word similarity, we find that stories generated by GPT-3 exhibit many known gender stereotypes. Generated stories depict different topics and descriptions depending on GPT-3's perceived gender of the character in a prompt, with feminine characters[1] more likely to be associated with family and appearance, and described as less powerful than masculine characters, even when associated with high power verbs in a prompt. Our study raises questions on how one can avoid unintended social biases when using large language models for storytelling.

1 Introduction

Advances in large language models have allowed new possibilities for their use in storytelling, such as machine-in-the-loop creative writing (Clark et al., 2018; Kreminski et al., 2020; Akoury et al., 2020) and narrative generation for games (Raley and Hua, 2020). However, fictional stories can reinforce real stereotypes, and artificially generated stories are no exception. Language models mimic patterns in their training data, parroting or even amplifying social biases (Bender et al., 2021).

An ongoing line of research examines the nature and effects of these biases in natural language generation (Sheng et al., 2020; Wallace et al., 2019; Shwartz et al., 2020). Language models generate different occupations and levels of respect for different genders, races, and sexual orientations (Sheng et al., 2019; Kirk et al., 2021). Abid et al. (2021) showed that GPT-3's association of Muslims and violence can be difficult to diminish, even when prompts include anti-stereotype content.

Our work focuses on representational harms in generated narratives, especially the reproduction of gender stereotypes found in film, television, and books. We use GPT-3, a large language model that has been released as a commercial product and thus has potential for wide use in narrative generation tasks (Brown et al., 2020; Brockman et al., 2020; Scott, 2020; Elkins and Chun, 2020; Branwen, 2020). Our experiments compare GPT-3's stories with literature as a form of domain control, using generated stories and book excerpts that begin with the same sentence.

Douloti understood some and didn't understand some. But he didn't care to understand. It was enough for him to know the facts of the situation and why his mother had left ...
Douloti understood some and didn't understand some. But more, she could tell that Nenn had sympathy for one who had given up life. Sister Nenn went on with her mending ...

Figure 1: GPT-3 can assign different gender pronouns to a character across different generations, as shown in this example using a prompt, in bold, pulled from Mahasweta Devi's *Imaginary Maps*.

We examine the topic distributions of books and GPT-3 stories, as well as the amount of attention given to characters' appearances, intellect, and power. We find that GPT-3's stories tend to include more masculine characters than feminine ones (mirroring a similar tendency in books), and identical prompts can lead to topics and descriptions that follow social stereotypes, depending on the prompt character's gender. Stereotype-related topics in prompts tend to persist further in a story if the character's gender aligns with the stereotype. Finally, using prompts containing different verbs, we are able to steer GPT-3 towards more intellectual, but not more powerful, characters. Code and materials to support this work can be found at https://github.com/lucy3/gpt3_gender.

2 Data

Our prompts are single sentences containing main characters sampled from 402 English contemporary fiction books, which includes texts from the

[1]We use "feminine character" to refer to characters with feminine pronouns, honorifics, or names, and ditto for "masculine character". See §3.1 for details.

Proceedings of the 3rd Workshop on Narrative Understanding, pages 48–55

Black Book Interactive Project, global Anglophone fiction, Pulitzer Prize winners, and bestsellers reported by *Publisher's Weekly* and the *New York Times*. We use BookNLP to find main characters and sentences containing them (Bamman et al., 2014). We define a main character as someone who is within their book's top 2% most frequent characters and mentioned at least 50 times. Every prompt is longer than 3 tokens, does not contain feminine or masculine pronouns, is from the main narrative and not dialogue, and contains only one single-token character name. This results in 2154 characters, with 10 randomly selected prompts each.

We use the GPT-3 API to obtain 5 text completions per prompt, with the `davinci` model, a temperature of 0.9, and a limit of 1800 tokens. A high temperature is often recommended to yield more "creative" responses (Alexeev, 2020; Branwen, 2020). We also pull excerpts that begin with each prompt from the original books, where each excerpt length is the average length of stories generated by that prompt. This human-authored text provides a control that contains the same main character names and initial content as GPT-3 data. The collection of generated stories contains over 161 million tokens, and the set of book excerpts contains over 32 million tokens.

3 Text processing methods

We use BookNLP's tokenizer and dependency parser on our data (Underwood et al., 2018; Bamman et al., 2014), followed by coreference resolution on named entities using the model annotated and trained on literature by Bamman et al. (2020). Pronoun chains containing the same character name within the same story are combined.

3.1 Gender inference

Depending on the context, gender may refer to a person's self-determined identity, how they express their identity, how they are perceived, and others' social expectations of them (Cao and Daumé III, 2020; Ackerman, 2019). Gender inference raises many ethical considerations and carries a risk of harmful misgendering, so it is best to have individuals self-report their gender (Larson, 2017). However, fictional characters typically do not state their genders in machine-generated text, and GPT-3 may gender a character differently from the original book. Our study focuses on how GPT-3 may perceive a character's gender based on textual features. Thus, we infer conceptual gender, or gender used by a perceiver, which may differ from the gender experienced internally by an individual being perceived (Ackerman, 2019).

First, we use a character's pronouns (*he/him/his*, *she/her/hers*, *their/theirs*) as a rough heuristic for gender. For book character gender, we aggregate pronouns for characters across all excerpts, while for generated text, we assign gender on a per-story basis. Since coreference resolution can be noisy, we label a character as feminine if at least 75% of their pronouns are *she/her*, and a character as masculine if at least 75% of their pronouns are *he/his*. The use of pronouns as the primary gendering step labels the majority of main characters (Figure 2). This approach has several limitations. Gender and pronoun use can be fluid, but we do not determine which cases of mixed-gender pronouns are gender fluidity rather than coreference error. Coreference models are also susceptible to gender biases (Rudinger et al., 2018), and they are not inclusive of nonbinary genders and pronouns (Cao and Daumé III, 2020).

Out of 734,560 characters, 48.3% have no pronouns. For these characters, we perform a second step of estimating expected conceptual gender by name, first using a list of gendered honorifics if they appear.[2] Then, if a name has no pronouns or honorifics, we use U.S. birth names from 1990 to 2019 (Social Security Administration, 2020), labeling a name as a gender if at least 90% of birth names have that gender. This step also has limitations. The gender categories of names are not exact, and the association between a name and gender can change over time (Blevins and Mullen, 2015). Some cultures do not commonly gender names, and U.S. name lists do not always generalize to names from other countries. Still, humans and NLP models associate many names with gender and consequently, with gender stereotypes (Bjorkman, 2017; Caliskan et al., 2017; Nosek et al., 2002; Moss-Racusin et al., 2012). We assume that GPT-3 also draws on social connotations when generating and processing names. We hope that future work can further improve the respectful measurement of gender in fiction.

All book excerpts and generated stories are more likely to have masculine characters, and in ones with feminine main characters in the prompt, there is a slightly smaller gap between feminine and mas-

[2]The full of list of honorifics is in our Github repo.

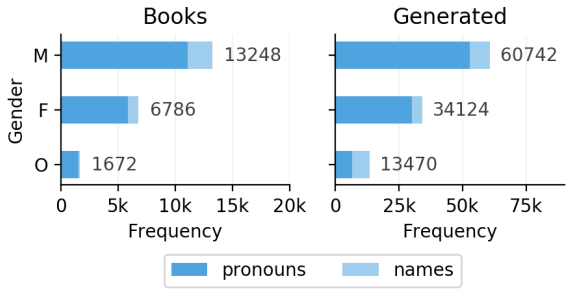

Figure 2: Frequency of masculine (*M*), feminine (*F*), and other (*O*) main prompt characters in our datasets. Bars are colored by gendering method.

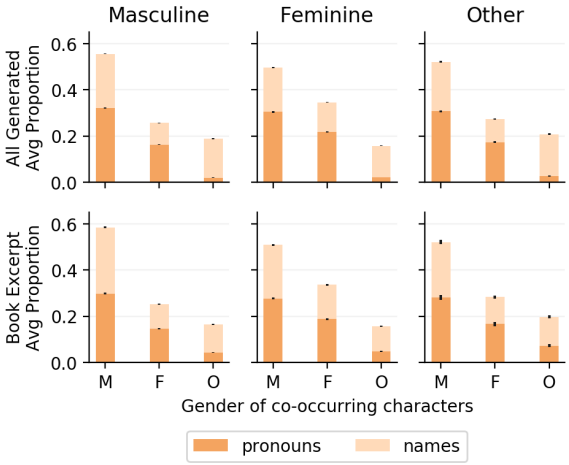

Figure 3: On average, there are more masculine characters in each GPT-3 story or book excerpt. Each column is the gender of the prompt character, and the bars are colored by gendering method. Error bars are 95% confidence intervals.

culine characters (Figure 3). This pattern persists even when only looking at pronoun-gendered characters, who are referred to multiple times and are likely to play larger roles. Our results echo previous work that show that English literature pays more attention to men in text (Underwood et al., 2018; Kraicer and Piper, 2018; Johns and Dye, 2019).

3.2 Matched stories

Prompts containing main characters of different genders may also contain different content, which can introduce confounding factors when isolating the effect of perceived gender on generated stories. We also run all our experiments on a subset of 7334 paired GPT-3 stories. Every prompt does not contain gendered pronouns and is used to generate multiple stories. GPT-3 may assign different gender pronouns to the main character in the same prompt across different stories (Table 1). We find cases where this occurs, randomly pairing stories with the same prompt, where one has the main character associated with feminine pronouns and another has them associated with masculine pronouns. In this setup, we exclude stories where the main character in the prompt is gendered by name.

4 Topic differences

Given this dataset of book excerpts and stories generated by GPT-3, we carry out several analyses to understand the representation of gender within them. We focus on overall content differences between stories containing prompt characters of different genders in this current section, and lexicon-based stereotypes in §5.

4.1 Method

Topic modeling is a common unsupervised method for uncovering coherent collections of words across narratives (Boyd-Graber et al., 2017; Goldstone and Underwood, 2014). We train latent Dirichlet allocation (LDA) on unigrams and bigrams from book excerpts and generated stories using MALLET, with 50 topics and default parameters. We remove character names from the text during training. For each topic t, we calculate $\Delta T(t) = P(t|F) - P(t|M)$, where $P(t|M)$ is the average probability of a topic occurring in stories with masculine main characters, and $P(t|F)$ is the analogous value for feminine main characters.

4.2 Results

Table 1 shows that generated stories place masculine and feminine characters in different topics, and in the subset of matched GPT-3 stories, these differences still persist (Pearson $r = 0.91$, $p < 0.001$). Feminine characters are more likely to be discussed in topics related to family, emotions, and body parts, while masculine ones are more aligned to politics, war, sports, and crime. The differences in generated stories follow those seen in books (Pearson $r = 0.84$, $p < 0.001$). Prompts with the same content can still lead to different narratives that are tied to character gender, suggesting that GPT-3 has internally linked stereotypical contexts to gender. In previous work, GPT-3's predecessor GPT-2 also places women in caregiving roles (Kirk et al., 2021), and character tropes for women emphasize maternalism and appearance (Gala et al., 2020).

We also use our trained LDA model to infer topic probabilities for each prompt, and examine prompts

topic	high probability words	all GPT-3	matched GPT-3
life	really, time, want, going, sure, lot, feel, little, life, things	0.018	0.010
family	baby, little, sister, child, girl, want, children, father, mom, mama	0.014	0.007
appearance	woman, girl, black, hair, white, women, looked, look, face, eyes	0.007	0.006
politics	people, country, government, president, war, american, world, chinese, political, united states	-0.008	-0.003
war	men, war, soldiers, soldier, general, enemy, camp, fight, battle, fighting	-0.008	-0.006
machines	plane, time, air, ship, machine, pilot, space, computer, screen, control	-0.008	-0.004

Table 1: Feminine and masculine main characters are associated with different topics, even in the matched prompt setup. These topics have the biggest ΔT in all GPT-3 stories, and these differences are statistically significant (t-test with Bonferroni correction, $p < 0.05$).

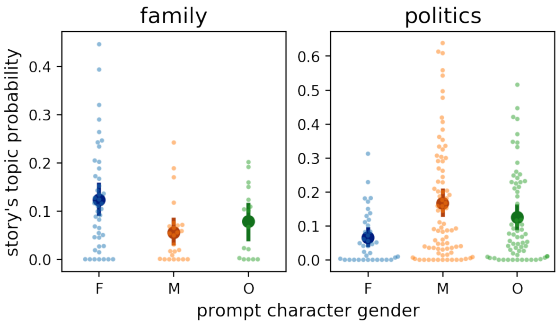

Figure 4: Prompt character gender is related the probability of a generated story continuing the *family* and *politics* topics. Each dot is a GPT-3 story, and the larger dots are means with 95% confidence intervals.

with a high (> 0.15) probability of a topic with gender bias, such as *politics* or *family*. We chose this threshold using manual inspection, and prompts that meet this threshold tended to have at least one topic-related word in them. When prompts contain the *family* topic, the resulting story tends to continue or amplify that topic more so if the main character is feminine (Figure 4). The reverse occurs when prompts have a high probability of *politics*: the resulting story is more likely to continue the topic if the main character is masculine. So, even if characters are in a prompt with anti-stereotypical content, it is still challenging to generate stories with topic probabilities at similar levels as a character with the stereotype-aligned gender.

5 Lexicon-based stereotypes

Now, we measure how much descriptions of characters correspond to a few established gender stereotypes. Men are often portrayed as strong, intelligent, and natural leaders (Smith et al., 2012; Sap et al., 2017; Fast et al., 2016b; Gala et al., 2020). Popular culture has increased its attention towards women in science, politics, academia, and law (Long et al., 2010; Inness, 2008; Flicker, 2003). Even so, depictions of women still foreground their physical appearances (Hoyle et al., 2019), and portray them as weak and less powerful (Fast et al., 2016b; Sap et al., 2017). Thus, our present study measures three dimensions of character descriptions: appearance, intellect, and power.

5.1 Method

Words linked to people via linguistic dependencies can be used to analyze descriptions of people in text (Fast et al., 2016b; Hoyle et al., 2019; Lucy et al., 2020; Bamman et al., 2013; Sap et al., 2017). These words can be aligned with lexicons curated by human annotators, such as Fast et al. (2016b)'s categories of adjectives and verbs, which were used to measure gender stereotypes in online fiction.

We train 100-dimensional word2vec embeddings (Mikolov et al., 2013) on lowercased, punctuationless generated stories and books, using default parameters in the `gensim` Python package. We extract adjectives and verbs using the dependency relations `nsubj` and `amod` attached to main character names and their pronouns in non-prompt text. For masculine and feminine characters, we only use their gender-conforming pronouns.

To gather words describing appearance, we combine Fast et al. (2016b)'s lexicons for *beautiful* and *sexual* (201 words). For words related to intellect, we use Fast et al. (2016a)'s Empath categories containing the word *intellectual* (98 words). For measuring power, we take Fast et al. (2016b)'s lexicons for *strong* and *dominant* (113 words), and contrast them with a union of their lexicons for *weak*, *dependent*, *submissive*, and *afraid* (141 words).

Counting lexicon word frequency can overemphasize popular words (e.g. *want*) and exclude related words. Therefore, we calculate semantic similarity instead. For appearance and intellect, we compute the average cosine similarity of a verb or adjective to every word in each lexicon. For power, we take a different approach, because antonyms tend be close in semantic space (Mrkšić et al., 2016). Previous work has used differences between antonyms to create semantic axes and compare words to these axes (Kozlowski et al., 2019; Turney and Littman, 2003; An et al., 2018). Let a

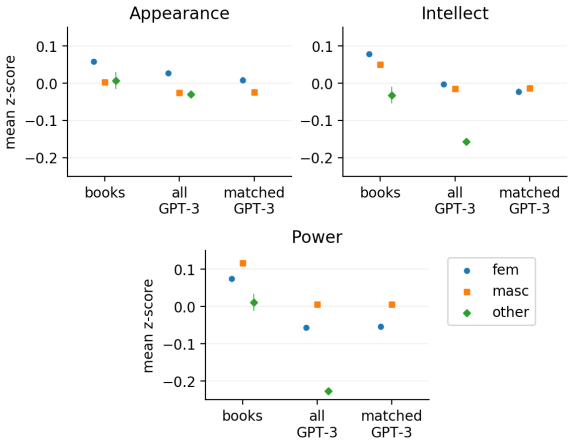

Figure 5: Appearance, intellect, and power scores across genders in books and GPT-3-generated stories. Error bars are 95% confidence intervals. All differences between feminine and masculine characters are significant (Welch's t-test, $p < 0.001$), except for intellect in matched GPT-3 stories.

be a word in the lexicon related to strength and b be a word embedding from the lexicon related to weakness. We use An et al. (2018)'s SEMAXIS to calculate word x's score:

$$S(x) = \cos\left(x, \frac{1}{|A|}\sum_{a \in A} a - \frac{1}{|B|}\sum_{b \in B} b\right),$$

where a positive value means x is stronger, and a negative value means x is weaker. We z-score all three of our metrics, and average the scores for all words associated with characters of each gender.

5.2 Results

Book characters have higher power and intellect than generated characters, but relative gender differences are similar between the two datasets (Figure 5). As hypothesized, feminine characters are most likely to be described by their appearance, and masculine characters are most powerful. The gender differences between masculine and feminine characters for appearance and power persist in matched GPT-3 stories, suggesting that GPT-3 has internally linked gender to these attributes. The patterns for intellect show that feminine characters are usually highest, though the insignificant difference in matched GPT-3 stories ($p > 0.05$) suggests that this attribute may be more affected by other content than gender.

We also test the ability of prompts to steer GPT-3 towards stronger and more intellectual characters. We examine character descriptions in stories generated by prompts in which characters are the subject of high power verbs from Sap et al. (2017)'s connotation frame lexicon, which was created for the study of characters in film. We also examine GPT-3 stories with prompts where characters use cognitive verbs from Bloom's Taxonomy, which is used to measure student learning, such as *summarize*, *interpret*, or *critique* (Anderson et al., 2001). We match verbs based on their lemmatized forms.

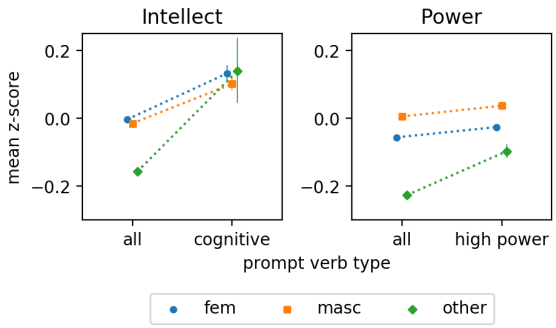

Figure 6: A comparison of stories generated by all prompts with stories generated by prompts where characters are linked to cognitive or high power verbs. Error bars are 95% confidence intervals.

We find that prompts containing cognitive verbs result in descriptions with higher intellect scores (Figure 6). Prompts containing high power verbs, however, do not lead to similar change, and non-masculine characters with high power verbs still have lower power on average than all masculine characters. Traditional power differentials in gender may be challenging to override and require more targeted prompts.

6 Conclusion

The use of GPT-3 for storytelling requires a balance between creativity and controllability to avoid unintended generations. We show that multiple gender stereotypes occur in generated narratives, and can emerge even when prompts do not contain explicit gender cues or stereotype-related content. Our study uses prompt design as a possible mechanism for mitigating bias, but we do not intend to shift the responsibility of preventing social harm from the creators of these systems to their users. Future studies can use causal inference and more carefully designed prompts to untangle the factors that influence GPT-3 and other text generation models' narrative outputs.

7 Acknowledgments

We thank Nicholas Tomlin, Julia Mendelsohn, and Emma Lurie for their helpful feedback on earlier versions of this paper. This work was supported by funding from the National Science Foundation (Graduate Research Fellowship DGE-1752814 and grant IIS-1942591).

References

Abubakar Abid, Maheen Farooqi, and James Zou. 2021. Persistent anti-Muslim bias in large language models. *arXiv preprint arXiv:2101.05783*.

Lauren Ackerman. 2019. Syntactic and cognitive issues in investigating gendered coreference. *Glossa: A Journal of General Linguistics*, 4(1).

Nader Akoury, Shufan Wang, Josh Whiting, Stephen Hood, Nanyun Peng, and Mohit Iyyer. 2020. STORIUM: A Dataset and Evaluation Platform for Machine-in-the-Loop Story Generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6470–6484, Online. Association for Computational Linguistics.

Vladimir Alexeev. 2020. GPT-3: Creative potential of NLP. Towards Data Science.

Jisun An, Haewoon Kwak, and Yong-Yeol Ahn. 2018. SemAxis: A lightweight framework to characterize domain-specific word semantics beyond sentiment. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2450–2461, Melbourne, Australia. Association for Computational Linguistics.

L.W. Anderson, B.S. Bloom, D.R. Krathwohl, P. Airasian, K. Cruikshank, R. Mayer, P. Pintrich, J. Raths, and M. Wittrock. 2001. *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives*. Longman.

David Bamman, Olivia Lewke, and Anya Mansoor. 2020. An annotated dataset of coreference in English literature. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 44–54, Marseille, France. European Language Resources Association.

David Bamman, Brendan O'Connor, and Noah A. Smith. 2013. Learning latent personas of film characters. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 352–361, Sofia, Bulgaria. Association for Computational Linguistics.

David Bamman, Ted Underwood, and Noah A. Smith. 2014. A Bayesian mixed effects model of literary character. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 370–379, Baltimore, Maryland. Association for Computational Linguistics.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? 🦜. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

Bronwyn M Bjorkman. 2017. Singular *they* and the syntactic representation of gender in English. *Glossa: A Journal of General Linguistics*, 2(1).

Cameron Blevins and Lincoln Mullen. 2015. Jane, john... leslie? a historical method for algorithmic gender prediction. *DHQ: Digital Humanities Quarterly*, 9(3).

Jordan Boyd-Graber, Yuening Hu, and David Mimno. 2017. Applications of topic models. *Foundations and Trends® in Information Retrieval*, 11(2-3):143–296.

Gwern Branwen. 2020. GPT-3 creative fiction.

Greg Brockman, Mira Murati, Peter Welinder, and OpenAI. 2020. OpenAI API. OpenAI Blog.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

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

Yang Trista Cao and Hal Daumé III. 2020. Toward gender-inclusive coreference resolution. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4568–4595, Online. Association for Computational Linguistics.

Elizabeth Clark, Anne Spencer Ross, Chenhao Tan, Yangfeng Ji, and Noah A. Smith. 2018. Creative writing with a machine in the loop: Case studies on slogans and stories. In *23rd International Conference on Intelligent User Interfaces*, IUI '18, page 329–340, New York, NY, USA. Association for Computing Machinery.

Katherine Elkins and Jon Chun. 2020. Can GPT-3 pass a writer's Turing Test? *Journal of Cultural Analytics*.

Ethan Fast, Binbin Chen, and Michael S Bernstein. 2016a. Empath: Understanding topic signals in large-scale text. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 4647–4657.

Ethan Fast, Tina Vachovsky, and Michael Bernstein. 2016b. Shirtless and dangerous: Quantifying linguistic signals of gender bias in an online fiction writing community. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 10.

Eva Flicker. 2003. Between brains and breasts—women scientists in fiction film: On the marginalization and sexualization of scientific competence. *Public Understanding of Science*, 12(3):307–318.

Dhruvil Gala, Mohammad Omar Khursheed, Hannah Lerner, Brendan O'Connor, and Mohit Iyyer. 2020. Analyzing gender bias within narrative tropes. In *Proceedings of the Fourth Workshop on Natural Language Processing and Computational Social Science*, pages 212–217, Online. Association for Computational Linguistics.

Andrew Goldstone and Ted Underwood. 2014. The quiet transformations of literary studies: What thirteen thousand scholars could tell us. *New Literary History*, 45(3):359–384.

Alexander Miserlis Hoyle, Lawrence Wolf-Sonkin, Hanna Wallach, Isabelle Augenstein, and Ryan Cotterell. 2019. Unsupervised discovery of gendered language through latent-variable modeling. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1706–1716, Florence, Italy. Association for Computational Linguistics.

Sherrie A. Inness. 2008. *Geek Chic: Smart Women in Popular Culture*. Palgrave Macmillan.

Brendan T. Johns and Melody Dye. 2019. Gender bias at scale: Evidence from the usage of personal names. *Behavior Research Methods*, 51(4).

Hannah Kirk, Yennie Jun, Haider Iqbal, Elias Benussi, Filippo Volpin, Frederic A. Dreyer, Aleksandar Shtedritski, and Yuki M. Asano. 2021. How true is gpt-2? an empirical analysis of intersectional occupational biases.

Austin C. Kozlowski, Matt Taddy, and James A. Evans. 2019. The geometry of culture: Analyzing the meanings of class through word embeddings. *American Sociological Review*, 84(5):905–949.

Eve Kraicer and Andrew Piper. 2018. Social characters: The hierarchy of gender in contemporary English-language fiction. *Cultural Analytics*.

Max Kreminski, Melanie Dickinson, Michael Mateas, and Noah Wardrip-Fruin. 2020. Why are we like this?: The AI architecture of a co-creative storytelling game. In *International Conference on the Foundations of Digital Games*, FDG '20, New York, NY, USA. Association for Computing Machinery.

Brian Larson. 2017. Gender as a variable in natural-language processing: Ethical considerations. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 1–11, Valencia, Spain. Association for Computational Linguistics.

Marilee Long, Jocelyn Steinke, Brooks Applegate, Maria Knight Lapinski, Marne J. Johnson, and Sayani Ghosh. 2010. Portrayals of male and female scientists in television programs popular among middle school-age children. *Science Communication*, 32(3):356–382.

Li Lucy, Dorottya Demszky, Patricia Bromley, and Dan Jurafsky. 2020. Content analysis of textbooks via natural language processing: Findings on gender, race, and ethnicity in Texas U.S. history textbooks. *AERA Open*, 6(3):2332858420940312.

Tomas Mikolov, Greg Corrado, Kai Chen, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *Proceedings of the International Conference on Learning Representations (ICLR)*, pages 1–12.

Corinne A. Moss-Racusin, John F. Dovidio, Victoria L. Brescoll, Mark J. Graham, and Jo Handelsman. 2012. Science faculty's subtle gender biases favor male students. *Proceedings of the National Academy of Sciences*, 109(41):16474–16479.

Nikola Mrkšić, Diarmuid Ó Séaghdha, Blaise Thomson, Milica Gašić, Lina M. Rojas-Barahona, Pei-Hao Su, David Vandyke, Tsung-Hsien Wen, and Steve Young. 2016. Counter-fitting word vectors to linguistic constraints. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 142–148, San Diego, California. Association for Computational Linguistics.

Brian A Nosek, Mahzarin R Banaji, and Anthony G Greenwald. 2002. Harvesting implicit group attitudes and beliefs from a demonstration web site. *Group Dynamics: Theory, Research, and Practice*, 6(1):101.

Rita Raley and Minh Hua. 2020. Playing with unicorns: AI dungeon and citizen NLP. *Digital Humanities Quarterly*, 14(4).

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*,

pages 8–14, New Orleans, Louisiana. Association for Computational Linguistics.

Maarten Sap, Marcella Cindy Prasettio, Ari Holtzman, Hannah Rashkin, and Yejin Choi. 2017. Connotation frames of power and agency in modern films. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2329–2334, Copenhagen, Denmark. Association for Computational Linguistics.

Kevin Scott. 2020. Microsoft teams up with OpenAI to exclusively license GPT-3 language model. The Official Microsoft Blog.

Emily Sheng, Kai-Wei Chang, Prem Natarajan, and Nanyun Peng. 2020. Towards Controllable Biases in Language Generation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3239–3254, Online. Association for Computational Linguistics.

Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2019. The woman worked as a babysitter: On biases in language generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3407–3412, Hong Kong, China. Association for Computational Linguistics.

Vered Shwartz, Rachel Rudinger, and Oyvind Tafjord. 2020. "you are grounded!": Latent name artifacts in pre-trained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6850–6861, Online. Association for Computational Linguistics.

Stacy L Smith, Marc Choueiti, Ashley Prescott, and Katherine Pieper. 2012. Gender roles & occupations: A look at character attributes and job-related aspirations in film and television. *Geena Davis Institute on Gender in Media*, pages 1–46.

Social Security Administration. 2020. Popular baby names: Beyond the top 1000 names. National Data.

Peter D. Turney and Michael L. Littman. 2003. Measuring praise and criticism: Inference of semantic orientation from association. *ACM Trans. Inf. Syst.*, 21(4):315–346.

William E Underwood, David Bamman, and Sabrina Lee. 2018. The transformation of gender in English-language fiction. *Journal of Cultural Analytics*, 1(1).

Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. 2019. Universal adversarial triggers for attacking and analyzing NLP. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2153–2162, Hong Kong, China. Association for Computational Linguistics.