# Mitigating Societal Harms in Large Language Models

**Sachin Kumar**[*,♣] **Vidhisha Balachandran**[*,♣] **Lucille Njoo**[♡]
**Antonios Anastasopoulos**[♢] **Yulia Tsvetkov**[♡]
[♣]Language Technologies Institute, Carnegie Mellon University
[♢]Department of Computer Science, George Mason University
[♡]Paul G. Allen School of Computer Science & Engineering, University of Washington
{sachink,vbalacha}@cs.cmu.edu, lnjoo@cs.washington.edu
antonis@gmu.edu, yuliats@cs.washington.edu

## Abstract

Numerous recent studies have highlighted societal harms that can be caused by language technologies deployed in the wild. While several surveys, tutorials, and workshops have discussed the risks of harms in specific contexts—e.g., detecting and mitigating gender bias in NLP models—no prior work has developed a unified typology of technical approaches for mitigating harms of language generation models. Our tutorial is based on a survey we recently wrote that proposes such a typology. We will provide an overview of potential social issues in language generation, including toxicity, social biases, misinformation, factual inconsistency, and privacy violations. Our primary focus will be on how to systematically identify risks, and how eliminate them at various stages of model development, from data collection, to model development, to inference/language generation. Through this tutorial, we aim to equip NLP researchers and engineers with a suite of practical tools for mitigating safety risks from pretrained language generation models.

## 1 Motivation

With the widespread success and increasing adoption on natural language processing (NLP) technologies in user-facing products including machine translation (Vaswani et al., 2017; Lewis et al., 2020), dialogue systems (Andreas et al., 2020; Gangadharaiah and Narayanaswamy, 2020) and recommendation systems (Jannach et al., 2020) the NLP community is becoming increasingly aware that we have a responsibility to evaluate the effects of our research and mitigate harmful outcomes (Bender et al., 2021). Indeed, models have been shown to introduce vulnerabilities and threats, both inadvertent and malicious, to individual users, social groups, and content integrity. Without social context and content control, deployed language generators have quickly derailed to racist, homophobic, hateful comments (Hunt, 2016; Jang, 2021; Wolf et al., 2017; Vincent, 2022), compromised user privacy (Carlini et al., 2021), spread disinformation (Shao et al., 2018), and even encouraged suicide (Daws, 2020). Prior works have outlined these risks (Maynez et al., 2020; Sheng et al., 2021; Weidinger et al., 2021), proposed taxonomies (Weidinger et al., 2022), discussed their points of origin, and advocated for research on ethical development of LMs (Bender et al., 2021; Solaiman et al., 2019).

However, there is little work that summarizes **actionable approaches and technical solutions** to preventing or mitigating these harms. This is the purpose of our tutorial, which is based on a survey we have recently conducted (Kumar et al., 2022). In this tutorial, we aim to provide a **comprehensive, unified taxonomy** of relevant **mitigation strategies** proposed in prior literature, specifically focusing on **language generation models**.

## 2 Tutorial Content and Relevance

**What are language models? A brief background:** To build a common ground for discussing the risk mitigation strategies, this tutorial will begin with a brief overview of recent trends in language modeling and pretraining. We will cover both causal (Radford et al., 2019; Brown et al., 2020) and non-causal language models (Devlin et al., 2019) highlighting their differences and their impact on NLP research. We will briefly discuss how pretrained models can be adapted to different tasks covering model finetuning (both complete and adapter based) as well as prompt-based formulation to solve NLP tasks. We will also focus on their scale both in terms of model parameters as well as training data size.

**How can language models cause societal harm?** After presenting the background on language models, we will then give a formal definition of harms based on taxonomy defined in prior work (Barocas et al., 2017) and focus on *representational harms*



*Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts*, pages 26–33
December 6-10, 2023 

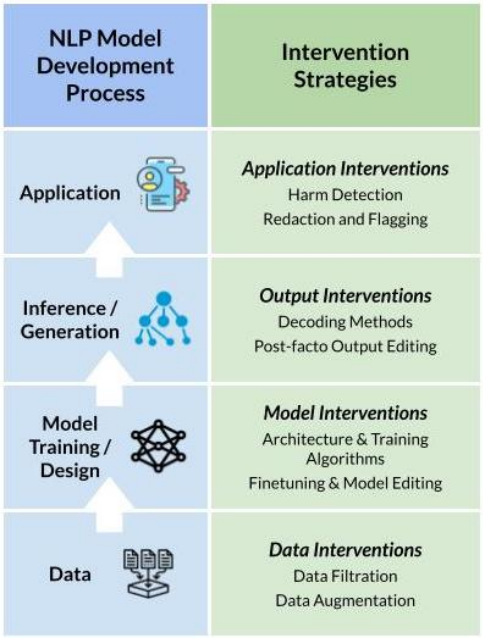


Figure 1: Overview of Intervention Strategies. Our survey presents a taxonomy of intervention strategies organized around the different phases where they can be applied.

in this tutorial. Highlighting the impact of heedlessly using web data which is usually population-imbalanced (Bender et al., 2021) and contains biased language against towards specific populations, we will discuss how language models tend to reinforce and amplify bias against sub-populations based on different personal and social attributes such as gender (Stanovsky et al., 2019; de Vassimon Manela et al., 2021), race (Liang et al., 2021; Field et al., 2021), region (Huang et al., 2020), demographics (Huang et al., 2020), age (Nangia et al., 2020) among others. We will also discuss, that by not being grounded in real world knowledge, they pickup on spurious statistical correlations in data and generate (in other words, hallucinate) factually incorrect content which can potentially be used to spread misinformation (Zellers et al., 2020; Kryscinski et al., 2020). Major content of this section is borrowed from the course on Ethics in NLP developed at Carnegie Mellon University and the University the Washington by organizer Yulia Tsvetkov.

**Can we reduce or mitigate such harms?** Finally, in this part, we will focus on work on mitigating harmful effects of language generation systems. While still a nascent field of research, several solutions in this space have been proposed which we categorize into four categories, visualized in Fig. 1. We organize and discuss in detail intervention strategies based on where they fit in different stages of LM development: **in data collection, modeling, post-factum decoding, and application**. Within each of these categories, our taxonomy brings together prior works that have been treated as disjoint areas targeting different types of harms (toxic/biased language and misinformation).

Since LMs learn and amplify biases present in the training data, we will first discuss data level interventions which focus on either (1) filtering the pretraining corpora to create more balanced datasets (Jia et al., 2020), or (2) finetuning trained LMs on sanitized data (Gehman et al., 2020a). Second, we will review model level interventions where we consider approaches which modify either the architecture or training objectives to induce or remove desired biases (Nan et al., 2021; Cao and Wang, 2021). Third, we will present methods to modify model outputs post generation using decoding and editing methods to demote or remove harmful content (Yang and Klein, 2021; Kumar et al., 2021; Cao et al., 2020). These techniques are especially useful for cases where it is impossible to modify data or models or even decoding strategies such as in case of GPT3 (Brown et al., 2020) which are only available through an API. Finally, we will end with application level interventions where we show how methods to flag and redact harmful content allow applications to shield such content from reaching users (Vaidya et al., 2020; Sun et al., 2019).

Throughout the tutorial, we will highlight both detection and mitigation approaches, as well as their specific limitations and shortcomings. By the end of the tutorial, participants will be better informed where to focus future research efforts.

Due to the vast range of societal harms and their mitigation strategies, we do not plan an exhaustive treatment of this material. One central goal is to raise awareness for participants of the relevant issues, so that when they return to their research they will be more able to notice ways in which their research based on large language models might impact different variety of users. To achieve this goal, we will aim for a "T-shape" in terms of breadth and depth: to briefly mention a number of core questions and then to drill down into a few particular case studies to see how these issues play out in real research settings.

## 3 Tutorial Structure

We propose a **cutting-edge tutorial** on an emerging area that has not been previously covered in ACL/EMNLP/NAACL/COLING tutorials. This would be a discussion-style tutorial where the organizers will present material with structured time throughout for questions, and discussion amongst attendees. The duration of the tutorial will be 3 hours with 5 min breaks at the end of each hour. The following would be the outline of the talk:

**1.** Brief Introduction to Language models (10 mins) - We will provide a quick background on current state of NLP research with introduction to language models and their capabilities.

**2.** Possible Harms of Language Technologies (15 mins) - We will briefly cover examples of ethical concerns, societal harms and biases present in current NLP tools.

- Fairness/Bias - Research on human-like biases in NLP (Field et al., 2021; Caliskan et al., 2017; Field and Tsvetkov, 2020)
- Toxicity - Research on toxic text generated by NLP models (Gehman et al., 2020a) and biases propagated in efforts to correct them (Davidson et al., 2017).
- Misinformation, Factual Inconsistencies - factual errors in generated text (Cao et al., 2018; Buchanan et al., 2021; Zellers et al., 2020)
- Privacy - Models generating sensitive, identifying information like addresses, SSN, etc. (Carlini et al., 2020; Inan et al., 2021)

**3.** Application Level Interventions (30 mins) - Techniques to filter harmful content before presenting model outputs to users.

- Harm Detection - Research on Toxic text detection (Vaidya et al., 2020; Han and Tsvetkov, 2020), fact-checking (Zhou et al., 2021), hallucination detection (Kryscinski et al., 2020; Goyal and Durrett, 2020), bias-detection (Sun et al., 2019; Park et al., 2018).
- Redacting or Flagging Harmful Text - Research on application level warnings or redaction for harmful or inappropriate generated text (Xu et al., 2020).

**4.** Output Level Interventions (30 mins) - Techniques to modify outputs to remove harmful content.

- Decoding Techniques - Research on search and sampling algorithms for controllable generation by promoting or demoting specific properties in output text (Zhang et al., 2022; Krishna et al., 2022; King et al., 2022).
- Post-Factum Editing - Research to edit or revise generated text to remove harmful content (Pryzant et al., 2020; He et al., 2021; Balachandran et al., 2022).

**5.** Model Level Interventions (30 mins) - Techniques to modify or optimize model parameters to prevent risky generations.

- Architecture and Training - Research on objectives and model architectures to enforce safe and reliable text generation (Yu et al., 2022; Nan et al., 2021; Falke et al., 2019).
- Finetuning and Model Editing - Research on editing or finetuning model parameters to incorporate safety constraints, through with new objectives (Gururangan et al., 2020; Chan et al., 2021; Gehman et al., 2020b; Chronopoulou et al., 2020).

**6.** Data Level Interventions (30 mins) - Techniques to curate clean training data to prevent models from using harmful text.

- Data Filtration - Research on filtering/removing training data instances containing toxic or harmful content (Ngo et al., 2021; Brown et al., 2020).
- Data Augmentation - Research on adding safer examples to datasets to offset the effect of problematic data (Mathew et al., 2018; Dinan et al., 2020; Stafanovičs et al., 2020).

**7.** Open Problems and Future Research (20 mins)

The tutorial will be a series of presentations with a set of references to related research papers and external demos. The presentation will cover a wide array of research on the topics from across the field. We will share the slides with the participants in advance. We will additionally share an online repository of relevant research material and online links to available code and demos to help participants navigate and use relevant research for their work. No copyright issues are expected as we will use open-source material.

## 4 General Information

### 4.1 Organizers

**Sachin Kumar** is a sixth year PhD candidate at the Language Technologies Institute, School of Computer Science at CMU. Sachin's research tackles critical technical problems in core language generation with deep learning, such as open-vocabulary generation, detection and demotion of spurious confounders, and controllable generation.

**Vidhisha Balachandran** (she/her) is a fourth-year Ph.D. student at the Language Technologies Institute, School of Computer Science at CMU. Her current research focuses on building interpretable and reliable NLP models with a focus on summarization, factuality, and KB-based reasoning.
**Lucille Njoo** (she/her) is a second-year PhD student at the Paul G. Allen School of Computer Science and Engineering at the University of Washington. She works in the intersection of NLP, ethics, and computational social science, working on identifying societal harms in NLP models.
**Antonios Anastasopoulos** (he/him) is an Assistant Professor at the Department of Computer Science at George Mason University, USA. His research focuses on NLP for local and low-resource languages and varieties, cross-lingual learning and multilinguality, and cross-lingual fairness.
**Yulia Tsvetkov** (she/her) is an Assistant Professor at the Paul G. Allen School of Computer Science and Engineering at the University of Washington, USA. Her research focuses on computational ethics, multilingual NLP, and machine learning for NLP. She developed a course on Computational Ethics in NLP and is teaching it at both undergraduate and graduate levels since 2017, and she is a co-chair of the ACL Ethics Committee.

### 4.2 Audience and Pre-Requisites

We expect participants from a wide array of backgrounds, including researchers, engineers, and end users of NLP technologies. Based on prior iterations of the tutorial, we expect an audience size of 50-100. No prior experience with NLP/ML is required, but we believe that our tutorial will most benefit those who are currently using NLP or are intending to use NLP tools in the near future in their research/products. An optional list of papers is presented in our survey paper (Kumar et al., 2022).

### 4.3 Diversity

The content of this tutorial highlights the impact of LMs on diverse users and therefore we aim to reach wide and diverse audiences. We will advertise this tutorial to diverse groups of researchers (e.g., Masakane, LatinX, North Africans, disabled in AI, indigenous in AI, Khipu) to bring in participants from various backgrounds. A previous version of this tutorial attracted audience from diverse gender, race as well as professional backgrounds like researchers, beginners and industry practitioners. Accordingly, our content will be made accessible to such audiences. Our own team is also diverse across multiple demographic attributes as well as professional expertise.

## 5 Logistics

**Previous Editions** This is the second iteration of the tutorial. The first edition of the tutorial was presented at The Web Conference 2022. While the previous iteration was focused to a general CS audience with less NLP background, this iteration will be modified to be aligned more for NLP-focused audience. This would entail including deeper technical specification of the interventions, including data, models and objectives.

Our tutorial is related and complementary to prior ACL tutorials related to bias and fairness in NLP (Socially Responsible NLP at NAACL 2018, Bias and Fairness in NLP at EMNLP 2019, Integrating Ethics into the NLP Curriculum at ACL 2020). Complementary to the content of the above tutorials which highlight social harms in NLP and discuss their detection, primarily focusing on representation learning and text classification, our tutorial will focus on practical methods to identify and mitigate harms in large language models and language generation.

**Venue** We prefer EMNLP or ACL, but any venue would work for us.

**Technical Requirements** We will not require additional equipment other than presentation material: an LCD projector, a computer with PowerPoint and Acrobat Reader, and internet connection.

**Public Release** We will publicly release all tutorial materials, including prerecorded lectures as backup for the tutorial which will be uploaded prior to the tutorial. These will be hosted on an open-access platform and linked from our University websites.

## 6 Ethics Statement

Although the aim of this tutorial is to improve the safety and inclusivity of NLP technologies and equip practitioners with tools to do so, we are well aware that as a not perfectly-diverse group of researchers we might incorporate our own biases into tutorial stricture and its technical focus. We will acknowledge this limitation in our tutorial, as well as the fact that the field of computational ethics is developing rapidly, and thus the content of our tutorial is inherently incomplete.

## References


Jacob Andreas, John Bufe, David Burkett, Charles Chen, Josh Clausman, Jean Crawford, Kate Crim, Jordan DeLoach, Leah Dorner, Jason Eisner, Hao Fang, Alan Guo, David Hall, Kristin Hayes, Kellie Hill, Diana Ho, Wendy Iwaszuk, Smriti Jha, Dan Klein, Jayant Krishnamurthy, Theo Lanman, Percy Liang, Christopher H. Lin, Ilya Lintsbakh, Andy McGovern, Aleksandr Nisnevich, Adam Pauls, Dmitrij Petters, Brent Read, Dan Roth, Subhro Roy, Jesse Rusak, Beth Short, Div Slomin, Ben Snyder, Stephon Striplin, Yu Su, Zachary Tellman, Sam Thomson, Andrei Vorobev, Izabela Witoszko, Jason Wolfe, Abby Wray, Yuchen Zhang, and Alexander Zotov. 2020. Task-Oriented Dialogue as Dataflow Synthesis. *Transactions of the Association for Computational Linguistics*, 8:556–571.

Vidhisha Balachandran, Hannaneh Hajishirzi, William Cohen, and Yulia Tsvetkov. 2022. Correcting diverse factual errors in abstractive summarization via post-editing and language model infilling. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Solon Barocas, Kate Crawford, Aaron Shapiro, and Hanna Wallach. 2017. The problem with bias: From allocative to representational harms in machine learning. In *Proc. SIGCIS*.

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.

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.

Ben Buchanan, Andrew Lohn, Micah Musser, and Katerina Sedova. 2021. Truth, lies, and automation.

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

Meng Cao, Yue Dong, Jiapeng Wu, and Jackie Chi Kit Cheung. 2020. Factual error correction for abstractive summarization models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6251–6258, Online. Association for Computational Linguistics.

Shuyang Cao and Lu Wang. 2021. Cliff: Contrastive learning for improving faithfulness and factuality in abstractive summarization.

Ziqiang Cao, Furu Wei, Wenjie Li, and Sujian Li. 2018. Faithful to the original: Fact aware neural abstractive summarization. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 4784–4791. AAAI Press.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. 2020. Extracting training data from large language models. *arXiv preprint arXiv:2012.07805*.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. 2021. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2633–2650.

Alvin Chan, Yew-Soon Ong, Bill Pung, Aston Zhang, and Jie Fu. 2021. Cocon: A self-supervised approach for controlled text generation. In *Proc. ICLR*.

Alexandra Chronopoulou, Dario Stojanovski, and Alexander Fraser. 2020. Reusing a Pretrained Language Model on Languages with Limited Corpora for Unsupervised NMT. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2703–2711, Online. Association for Computational Linguistics.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 11.

Ryan Daws. 2020. Medical chatbot using OpenAI's GPT-3 told a fake patient to kill themselves.

Daniel de Vassimon Manela, David Errington, Thomas Fisher, Boris van Breugel, and Pasquale Minervini. 2021. Stereotype and skew: Quantifying gender bias in pre-trained and fine-tuned language models. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2232–2242, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of*

*the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Emily Dinan, Angela Fan, Adina Williams, Jack Urbanek, Douwe Kiela, and Jason Weston. 2020. Queens are powerful too: Mitigating gender bias in dialogue generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8173–8188, Online. Association for Computational Linguistics.

Tobias Falke, Leonardo F. R. Ribeiro, Prasetya Ajie Utama, Ido Dagan, and Iryna Gurevych. 2019. Ranking generated summaries by correctness: An interesting but challenging application for natural language inference. In *ACL (1)*, pages 2214–2220.

Anjalie Field, Su Lin Blodgett, Zeerak Waseem, and Yulia Tsvetkov. 2021. A survey of race, racism, and anti-racism in nlp. *arXiv preprint arXiv:2106.11410*.

Anjalie Field and Yulia Tsvetkov. 2020. Unsupervised discovery of implicit gender bias. *arXiv preprint arXiv:2004.08361*.

Rashmi Gangadharaiah and Balakrishnan Narayanaswamy. 2020. Recursive template-based frame generation for task oriented dialog. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2059–2064, Online. Association for Computational Linguistics.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A Smith. 2020a. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. *EMNLP*.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020b. RealToxicityPrompts: Evaluating neural toxic degeneration in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3356–3369, Online. Association for Computational Linguistics.

Tanya Goyal and Greg Durrett. 2020. Evaluating factuality in generation with dependency-level entailment. In *EMNLP (Findings)*, pages 3592–3603.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Xiaochuang Han and Yulia Tsvetkov. 2020. Fortifying toxic speech detectors against veiled toxicity. *arXiv preprint arXiv:2010.03154*.

Zexue He, Bodhisattwa Prasad Majumder, and Julian McAuley. 2021. Detect and perturb: Neutral rewriting of biased and sensitive text via gradient-based decoding. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4173–4181, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Po-Sen Huang, Huan Zhang, Ray Jiang, Robert Stanforth, Johannes Welbl, Jack Rae, Vishal Maini, Dani Yogatama, and Pushmeet Kohli. 2020. Reducing sentiment bias in language models via counterfactual evaluation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 65–83, Online. Association for Computational Linguistics.

Elle Hunt. 2016. Tay, Microsoft's AI chatbot, gets a crash course in racism from Twitter.

Huseyin A Inan, Osman Ramadan, Lukas Wutschitz, Daniel Jones, Victor Rühle, James Withers, and Robert Sim. 2021. Training data leakage analysis in language models. *arXiv preprint arXiv:2101.05405*.

Heesoo Jang. 2021. A South Korean chatbot shows just how sloppy tech companies can be with user data.

Dietmar Jannach, Ahtsham Manzoor, Wanling Cai, and Li Chen. 2020. A survey on conversational recommender systems. *arXiv preprint arXiv:2004.00646*.

Shengyu Jia, Tao Meng, Jieyu Zhao, and Kai-Wei Chang. 2020. Mitigating gender bias amplification in distribution by posterior regularization. In *ACL (short)*.

Daniel King, Zejiang Shen, Nishant Subramani, Daniel S Weld, Iz Beltagy, and Doug Downey. 2022. Don't say what you don't know: Improving the consistency of abstractive summarization by constraining beam search. *arXiv preprint arXiv:2203.08436*.

Kalpesh Krishna, Ya yin Chang, John Wieting, and Mohit Iyyer. 2022. Rankgen: Improving text generation with large ranking models. *ArXiv*, abs/2205.09726.

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346, Online. Association for Computational Linguistics.

Sachin Kumar, Vidhisha Balachandran, Lucille Njoo, Antonios Anastasopoulos, and Yulia Tsvetkov. 2022. Language generation models can cause harm: So what can we do about it? an actionable survey. *arXiv preprint arXiv:2210.07700*.

Sachin Kumar, Eric Malmi, Aliaksei Severyn, and Yulia Tsvetkov. 2021. Controlled text generation as continuous optimization with multiple constraints. In *Proc. NeurIPS*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Paul Pu Liang, Chiyu Wu, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2021. Towards understanding and mitigating social biases in language models. In *International Conference on Machine Learning*, pages 6565–6576. PMLR.

Binny Mathew, Punyajoy Saha, Hardik Tharad, Subham Rajgaria, Prajwal Singhania, Suman Kalyan Maity, Pawan Goyal, and Animesh Mukherje. 2018. Thou shalt not hate: Countering online hate speech.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

Feng Nan, Cicero Nogueira dos Santos, Henghui Zhu, Patrick Ng, Kathleen McKeown, Ramesh Nallapati, Dejiao Zhang, Zhiguo Wang, Andrew O. Arnold, and Bing Xiang. 2021. Improving factual consistency of abstractive summarization via question answering. 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 6881–6894, Online. Association for Computational Linguistics.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967, Online. Association for Computational Linguistics.

Helen Ngo, Cooper Raterink, João GM Araújo, Ivan Zhang, Carol Chen, Adrien Morisot, and Nicholas Frosst. 2021. Mitigating harm in language models with conditional-likelihood filtration. *arXiv preprint arXiv:2108.07790*.

Ji Ho Park, Jamin Shin, and Pascale Fung. 2018. Reducing gender bias in abusive language detection. *arXiv preprint arXiv:1808.07231*.

Reid Pryzant, Richard Diehl Martinez, Nathan Dass, Sadao Kurohashi, Dan Jurafsky, and Diyi Yang. 2020. Automatically neutralizing subjective bias in text. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(01):480–489.

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

Chengcheng Shao, Giovanni Luca Ciampaglia, Onur Varol, Kai-Cheng Yang, Alessandro Flammini, and Filippo Menczer. 2018. The spread of low-credibility content by social bots. *Nature communications*, 9(1):1–9.

Emily Sheng, Kai-Wei Chang, Prem Natarajan, and Nanyun Peng. 2021. Societal biases in language generation: Progress and challenges. 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 4275–4293, Online. Association for Computational Linguistics.

Irene Solaiman, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, Gretchen Krueger, Jong Wook Kim, Sarah Kreps, et al. 2019. Release strategies and the social impacts of language models. *arXiv preprint arXiv:1908.09203*.

Artūrs Stafanovičs, Toms Bergmanis, and Mārcis Pinnis. 2020. Mitigating gender bias in machine translation with target gender annotations. In *Proceedings of the Fifth Conference on Machine Translation*, pages 629–638, Online. Association for Computational Linguistics.

Gabriel Stanovsky, Noah A. Smith, and Luke Zettlemoyer. 2019. Evaluating gender bias in machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1679–1684, Florence, Italy. Association for Computational Linguistics.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating gender bias in natural language processing: Literature review. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1630–1640, Florence, Italy. Association for Computational Linguistics.

Ameya Vaidya, Feng Mai, and Yue Ning. 2020. Empirical analysis of multi-task learning for reducing identity bias in toxic comment detection. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 683–693.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

James Vincent. 2022. YouTuber trains AI bot on 4chan's pile o' bile with entirely predictable results.

Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, Zac Kenton, Sasha Brown, Will Hawkins, Tom Stepleton, Courtney Biles, Abeba Birhane, Julia Haas, Laura Rimell, Lisa Anne Hendricks, William S. Isaac, Sean Legassick, Geoffrey Irving, and Iason Gabriel. 2021. Ethical and social risks of harm from language models.

Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, et al. 2022. Taxonomy of risks posed by language models.

Marty J Wolf, Keith W Miller, and Frances S Grodzinsky. 2017. Why we should have seen that coming: comments on Microsoft's Tay "experiment," and wider implications. *The ORBIT Journal*, 1(2):1–12.

Jing Xu, Da Ju, Margaret Li, Y-Lan Boureau, Jason Weston, and Emily Dinan. 2020. Recipes for safety in open-domain chatbots. *arXiv preprint arXiv:2010.07079*.

Kevin Yang and Dan Klein. 2021. FUDGE: Controlled text generation with future discriminators. In *Proc. NAACL*.

Wenhao Yu, Chenguang Zhu, Zaitang Li, Zhiting Hu, Qingyun Wang, Heng Ji, and Meng Jiang. 2022. A survey of knowledge-enhanced text generation. *ACM Comput. Surv.* Just Accepted.

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2020. Defending against neural fake news. *Neurips*.

Hanqing Zhang, Haolin Song, Shaoyu Li, Ming Zhou, and Dawei Song. 2022. A survey of controllable text generation using transformer-based pre-trained language models.

Xuhui Zhou, Maarten Sap, Swabha Swayamdipta, Noah A Smith, and Yejin Choi. 2021. Challenges in automated debiasing for toxic language detection. *arXiv preprint arXiv:2102.00086*.