AI Writing Assistants Influence Topic Choice in Self-Presentation

Ritika Poddar*
Cornell Tech
New York, New York, USA

Rashmi Sinha*
Cornell Tech
New York, New York, USA

Mor Naaman
Cornell Tech
New York, New York, USA

Maurice Jakesch
Cornell University
Ithaca, New York, USA

ABSTRACT

AI language technologies increasingly assist and expand human communication. While AI-mediated communication reduces human effort, its societal consequences are poorly understood. In this study, we investigate whether using an AI writing assistant in personal self-presentation changes how people talk about themselves. In an online experiment, we asked participants (N=200) to introduce themselves to others. An AI language assistant supported their writing by suggesting sentence completions. The language model generating suggestions was fine-tuned to preferably suggest either interest, work, or hospitality topics. We evaluate how the topic preference of a language model affected users' topic choice by analyzing the topics participants discussed in their self-presentations. Our results suggest that AI language technologies may change the topics their users talk about. We discuss the need for a careful debate and evaluation of the topic priors built into AI language technologies.

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in collaborative and social computing**; **Interaction design theory, concepts and paradigms**.

KEYWORDS

Co-writing, GPT-3, risks of large language models

ACM Reference Format:
Ritika Poddar, Rashmi Sinha, Mor Naaman, and Maurice Jakesch. 2023. AI Writing Assistants Influence Topic Choice in Self-Presentation. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems (CHI EA '23), April 23–28, 2023, Hamburg, Germany.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3544549.3585893

1 INTRODUCTION

AI language technologies are increasingly used to augment or enhance human communication [18]. From single-word shortcuts and suggestions to sentence completion and translation by services [26], AI language technologies enable new types of interactions that reduce human effort. With the recent generation of large transformer-based generative language models like GPT-3 [8], coherent human-like language can be generated automatically at scale [24]. However, infusing human communication with AI-generated language suggestions can have unintentional side effects with far-reaching consequences that may be hard to foresee. For example, when language models generate certain topics more often than others, the communication tools they power may change conversation topics and entire societal discourses.

The current study investigates whether a language-model-powered AI writing assistant that preferably suggests certain topics shifts how its users present themselves to others. We explore this research question in the context of self-presentation of the type prevalent in online profiles, e.g., on online hospitality platforms [20, 38]. Previous work on online self-presentation [12, 39] has shown that impression formation based on self-descriptions helps establish trust in technology-mediated environments [14, 29]. When AI language technologies change how people present themselves to others, they may not only change how people perceive themselves and are perceived by others [12, 14] but may undermine the trust required for social interactions such as sharing [23].

In a quantitative experimental study, we asked participants (N=200) to introduce themselves to potential guests on an online hospitality platform. Each participant was assisted by an AI writing assistant that suggested completions for their sentences. The writing assistant was powered by a customized version of OpenAI's GPT-2, a transformer-based large language model [34]. We fine-tuned three different versions of the model such that they preferably generated suggestions related to people's (1) interests and hobbies, (2) work and education, or (3) hospitality and guests. We randomly assigned one of these three topic-biased language models to power the AI writing assistant that supported participants in writing their self-presentations. We analyzed how participants interacted with the writing assistant and how the composition of topics in their self-presentations differed across treatment groups.

To preview our results, participants were significantly more likely to talk about the topics that the AI writing assistant suggested – even if we account for suggestions they directly accepted from the model. We discuss the implications of our findings in the context of the increasing deployment of AI language technologies into our communication [19] and argue that there is an urgent need for careful monitoring and evaluation of their topic preferences.

*Authors contributed equally.

CHI EA '23, April 23–28, 2023, Hamburg, Germany

ACM ISBN 978-1-4503-9422-2/23/04.
https://doi.org/10.1145/3544549.3585893

2 RELATED WORK

Our work builds on previous research on people's interactions with writing assistants and the societal risk of large language models.

2.1 Interaction with writing assistants

Word suggestions and sentence completions have become popular features in commercial products, such as mobile phones and email clients [9, 10]. While basic assistants predict single words based on frequency distributions [13, 17], more sophisticated assistants suggest sentence completions or entire paragraphs using the generative power of large language models [9]. Prior research has primarily focused on the impact of suggestions on writing efficiency [6, 15]. More recent work with advanced models [8, 34] has explored different outcomes of interest, such as enhanced creativity. Studies have been conducted using systems to aid in writing slogans [11], creative stories[36], science fiction [31], and metaphors [16]. There is relatively little prior work on the types of unwanted side effects that AI language technologies may have on people's writing. Initial studies found that predictive text systems could introduce biases into reviews [1, 5] or image caption [2] based on the predominant sentiment of the suggestions. Similarly, experimental work has suggested that opinionated language models may change their users' views and attitudes [22]. In the current experiment, we evaluate how co-writing with large language may affect the topics that users write about.

2.2 Societal risks of large language models

Large generative language models [7, 41, 42] have received much attention for the new types of product and interactions they enable [7]. However, they have also raised concerns about societal risks associated with the use and capabilities of these models. A major concern is that biases prevalent in the language generated can lead to unintended negative impacts on system users, for example minoritized groups may be more susceptible to discrimination and exclusion [8, 21, 32]. In addition, the technology may contribute to new forms of misinformation [27, 28, 35, 43] and cause other environmental [40] and socioeconomic harms [4]. Comparatively little work has considered how the use of AI language technologies may change our communication topics and behaviors [19]. First audits of widely used models seem to suggest a western bias, e.g., GPT-3's output aligns more with reported dominant US values than those upheld in other cultures [25] and may reinforce the respective values when widely used. We contribute to this debate by evaluating how AI language models may shift how people present themselves.

3 METHODS

We conducted a randomized controlled experiment where we asked participants (N=200) to write a self-presentation or personal introduction for an online hospitality platform. In their writing, they were assisted by an AI language assistant that suggested possible sentence completions. The language model powering the assistant was biased to overly suggest selected topics. Our primary hypothesis was that the text suggestions generated by the language model–if they were biased towards a certain topic–would lead participants to write more extensively about that topic in their self-presentation. We test this hypothesis by comparing the number of words participants wrote about each topic across treatment groups.

3.1 Experiment design

Our experiment took places in an interactive text editor with an AI writing assistant powered by a customized version of GPT-2 [34] in the backend. The writing assistant generated sentence continuations for participants as they typed, giving the option of either accepting the suggestions, requesting a new suggestion, or ignoring it. Once participants were satisfied with their self-presentation, they could submit their text, which we saved in a cloud storage bucket.

We used a between-subjects design with block randomization and a single independent variable: the topic bias of the model that powered the writing assistant. Participants were unaware of which model they were using. We used GPT-2 [34] as our baseline model and finetuned different model versions that preferably generated suggestions for one of the following topics: (1) interests and hobbies, (2) work and education, or (3) hospitality and guests. To create these models, we draw on a previous study of Airbnb host profiles [29]. The study's authors collected 1,200 Airbnb host profiles. They manually labeled their sentences into topic categories and made the coded data publicly available. In a subsequent study[30], they used the coding scheme scheme and extended it to 4,180 Airbnb host profiles by developing a computational classifier.

Using these datasets, we extracted the sentences that were labeled as mentioning only interest, work, or hospitality topics. We obtained 885 sentences related to work and education, 1,085 examples of interest-related sentences, and 908 self-presentation text samples related to guests and hospitality. We applied further preprocessing to parts of the data to reduce noise, such as collapsing multiple white spaces. We split each sample set into a train and test set with a train-to-test ratio of 0.9. We finetuned a version of GPT-2 on each topic dataset for five epochs with a learning rate of 0.005 and 100 warm-up steps.

We built a topic bias detection pipeline to evaluate the topic bias of each finetuned model. We first used nine minimal prompts ('I', 'We', 'As a', 'My', etc.) to generate sample continuations with each of the three fine-tuned models. We then trained a BERTopic, BagOfWords, and BertForSequenceClassification classifier on an extended version of the training data discussed above to predict the topic of text sequences. Since BertForSequenceClassification offered the best performance on the test set (F1=0.86), we used it for the evaluation of the topic bias of our fine-tuned models. The results indicate the continuations generated by the finetuned model contained the intended treatment topic in about 70% of sentences for work and education, 75% of sentences for hospitality and guests, and 83% of sentences for interests and hobbies.

3.2 Outcome measures and covariates

Two researchers independently coded the topics participants mentioned in the self-presentations they had written with the AI writing assistant. They labeled the data on a sentence level, attributing each sentence to a topics: (1) interests and hobbies, (2) work and education, or (3) hospitality and guests. Sentences that referred to neither were coded as (4) Other. Sentences that referred to multiple

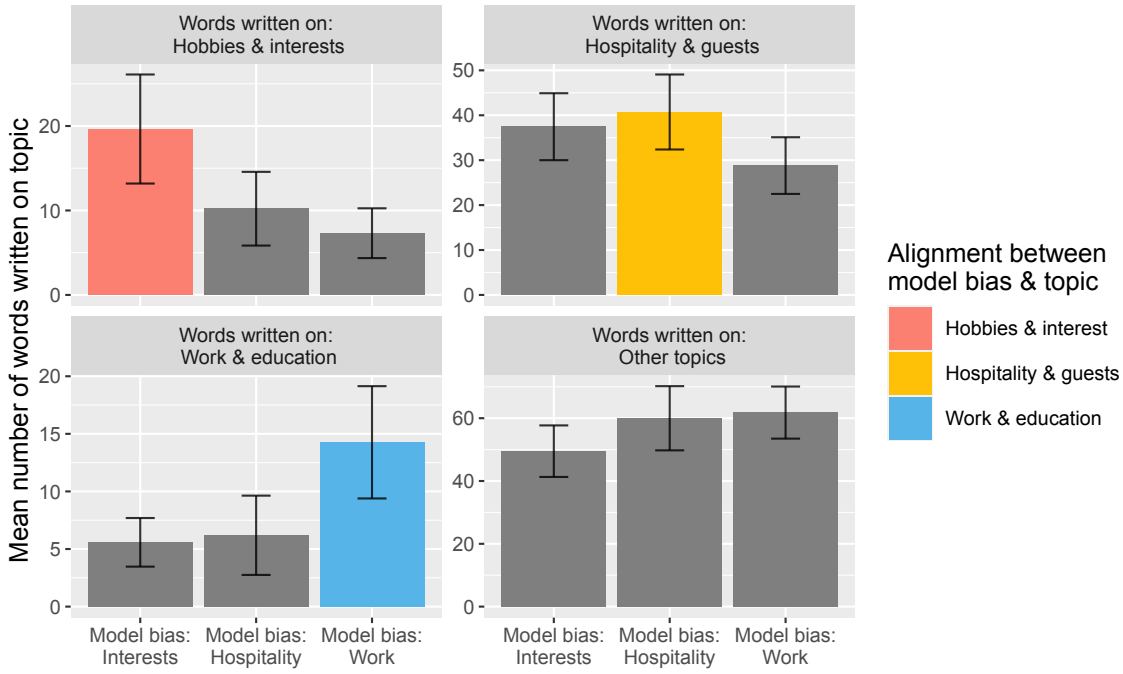

Figure 1: Participants assisted by a model that preferably suggested a certain topic were more likely to write about the topic in their self-descriptions. N_s=21,843 words written by N_p=192 participants. Error bars represent 95% confidence intervals. The y-axis indicates the average number of words participants wrote on the topic shown in the panel title. Word counts are split by experimental treatment based on whether participants co-wrote with a language model that preferably suggested writing about their interests, work, or hospitality. Coloured bars indicate alignment between the topic measure and the model bias.

topics were split into sentence fragments that referred to a specific topic or represented the share of the topic in a sentence according to the annotators' judgment. The researchers were aware of the experimental group participants were assigned to. To account for the different lengths of the fragments, we used the sum of fragment word counts–rather than sentence counts–as our main outcome measure. Through this coding process, we achieved an interrater reliability of 87.8%. Disagreements in codes were resolved in a collective discussion of cases.

In addition to collecting the final essays, we kept track of participants' interactions with the writing assistant and collected their demographics. We measured how many words they accepted from the models' suggestions. We use a covariate to differentiate between direct topic influence by accepting a suggestion and indirect influence by merely reading and being inspired by suggestions. We also recorded the time participants took to write their essays and asked them about their gender, age, and ethnicity.

3.3 Experiment procedure and participant recruitment

We recruited 200 participants through Prolific [33]. We first elicited participants' consent, then conducted three attention checks and provided them with a link to the web app with the writing task. The study was open to participants 18+ years old and proficient in English. 68% of participants were between 18 and 24 years of age, 24% were between 25 and 34, 5% were between 35 and 44, and 3% were older than 45 years. 52.4% of our participants were male, 44.2% were female, and 3.4% reported themselves as other or preferred not to say. 49.0% of participants self-identified as White, 7.8% as Black or African American, 2.9% as Asian, and 1.0% as American Indian or Alaska Native. About 39.3% selected *other* or preferred not to say.

Participants received about 10$ per hour for their time. We planned the sample size based on the effect we observed in a pilot with 30 participants, with an alpha of 0.05 and a power of 0.80. We rejected 12 submissions for writing very short self-presentations of less than 55 words and reopened the study to replace these rejected submissions. We also had to drop eight submissions during the analysis where the writing assistant had generated no suggestions due to system issues. Our experimental protocols were approved by the Cornell University Institutional Review Board.

4 RESULTS

Figure 1 shows the mean number of words participants wrote about interests, work, and hospitality in their self-descriptions. The top left panel counts how many words in participants' self-descriptions were related to describing their hobbies and interests. The word counts are further disaggregated based where the AI

Table 1: Linear regression predicting the number of words written on interests, hospitality, and work & education based on the bias of the mode powering the writing assistant. The model baseline (constant) is the group using a model train to preferably suggest work & education topics.

	Dependent variable:		
	Interest word count	Hospitality word count	Work & edu. word count
	(1)	(2)	(3)
Model bias: Interests	9.909*	9.459	−9.898**
Model bias: Hospitality	0.214	12.117*	−5.545
Words accepted from model	−0.201	0.239	0.195
Model: Interests * accepted words	0.201	−0.726	−0.001
Model: Hospitality * accepted words	0.275	0.194	−0.378
Writing time	0.008**	−0.004	0.001
Participant age	−0.083	0.079	−0.092
Constant	3.123	28.398**	14.867**
Observations	192	192	192
R^2	0.101	0.060	0.111
Adjusted R^2	0.066	0.022	0.076
Residual Std. Error	19.338	30.143	15.554
F Statistic	2.848**	1.603	3.155**
Note:			*p<0.05; **p<0.01; ***p<0.001

assistant preferably suggested interest (left bars), hospitality (center), or work topics (right). On average, participants using a model that preferably generated suggestions related to personal interests wrote 19.64 words on their hobbies and interests (top left in red). In comparison, participants who used the hospitality- or work-biased model only wrote 10.21 and 7.30 words on their interests, respectively. Similarly, as shown in the bottom left panel, participants using a model that preferably generated suggestions related to work wrote 14.26 words about their work and education (blue), while participants using an interest- or hospitality-biased model wrote 5.58 and 6.19 words about their work and education respectively. The outcome difference across treatment groups is less pronounced for participants' writing on hospitality, shown in yellow in the top right panel. Here, participants using a hospitality-biased model wrote 40.72 words on hospitality topics (yellow), while participants with interest- and work-biased models wrote 37.45 and 28.79 words, respectively.

We continue our analysis with fitting a regression model to the data to confirm whether the differences in outcomes are statistically significant across treatment groups. We also explore whether the differences observed may be due to participants conveniently accepting the model's suggestions. The three models shown in Table 1 predict the words participants wrote on each topic based on the language model they used. The regression baseline is the group using a language model preferably suggesting work & education topics. In addition to the main treatment variables (rows one and two), the regression models include a covariate for the number of words participants accepted from the models' suggestion (row three) and an interaction term between the model bias and the number of accepted words (rows four and five).

The left column shows the fitted coefficients predicting the number of words participants wrote on their interests and hobbies. Participants using an interest-biased model wrote significantly more words about their interests than participants using a work-biased model (the baseline of the regression), as indicated by the significant coefficient in the first row ($B = 9.9, p < 0.05$). Similarly, participants who took more time to write their self-descriptions wrote more words about their interests ($B = 0.0087, p < 0.01$). While participants who used an interest-biased model and accepted a larger number of accepted words from the model (interaction term in the fourth row) wrote even more words about their interests ($B = 0.2$, not statistically significant), the number of words accepted did not explain the difference in words counts between treatment groups (ie., the main effect coefficient in the first row remains significant). Demographic factors like age were not significantly correlated with topic prevalence.

The right column shows the coefficients predicting the word count related to work and education. Here, since the work-biased model constitutes the model baseline, the coefficients for the other treatment groups are negative. For example, participants using an interests-biased model wrote fewer words on work and education than participants using a work-biased model ($B = -9.9, p < 0.01$). The statistical results for the hospitality topic (center column) and work topic (right column) are similar to those of the interest topic discussed above.

5 DISCUSSION

Our results indicate that when people are writing with an AI writing assistant that preferably suggests a certain topic, they write more extensively about the topic preferred by the AI system. In

the context of this study–personal self-presentation [20, 38]–this meant that people described themselves depending on what type of AI assistant they were using. The difference in topic distributions across treatment group persisted even when we controlled for the words that were directly accepted from the model in a regression analysis, suggesting that the AI assistant influenced topic choice not only through accepted suggestions, but also indirectly by merely displaying possible continuations.

We note that for the hospitality and guests topic, the effect the model bias had on participants' topic choice was less pronounced. We interpret this smaller effect size as a ceiling effect, as participants wrote a large number of words on hospitality to topics independent of the treatment group due to the hospitality context of our experimental task.

Our findings align with recent research finding that using AI language technologies in human communication does not only increase efficiency but changes the content of what is being said. Previous studies have shown that using an AI writing assistant affects the sentiment and valence of what people write [1, 5] and may lead to shorter or more generic text being written [2]. Another study even showed that large language models that preferably generate certain opinions can shift not only the opinions users express in writing, but also the opinions people hold after interacting with the language model [22].

Changing the topics people write (or talk about) should not been seen as a minor side-effect of AI language technologies. In the context of our study, personal self-presentation, the topics people choose to talk about have downstream effects on how others trust them [23, 29, 30]. A change in how people describe themselves may not only affect how others perceive them but, according to self-perception theory [3], even change how they think about themselves. If our findings generalize to topic choice in other social contexts than self-presentation, AI language technologies could influence larger societal discourses by agenda-setting, and the priming, and framing of topics [37] through their suggestions. Here, a more careful discussion and monitoring of topic priors in AI language technologies seems necessary in both academia and industry.

Our early study has several limitations: We only investigated the effects of topic biases through large language models in a specific social context–online self-presentation–and with specific language model (GPT-2). A more extensive replication study can solidify our findings and test how they generalize to different social contexts. Using a different language model–such as GPT-3–in a different AI language product–such as smart replies–may lead to different outcomes in topic influence. Further investigations are also required to understand the significance of our findings. We still need to understand the mechanisms of the effects we observed–to what extent are they, for example, due to convenience or perceived authority of the AI model? Future studies will also need to explore how large the topic shifts are compared to other influences on topics choice, and how transient or lasting the observed treatment effects are.

ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under Grant No. CHS 1901151/1901329 and the German National Academic Foundation.

REFERENCES

[1] Kenneth C. Arnold, Krysta Chauncey, and Krzysztof Z. Gajos. 2018. Sentiment Bias in Predictive Text Recommendations Results in Biased Writing. In *Proceedings of the 44th Graphics Interface Conference* (Toronto, Canada) *(GI '18)*. Canadian Human-Computer Communications Society, Waterloo, CAN, 42–49. https://doi.org/10.20380/GI2018.07

[2] Kenneth C. Arnold, Krysta Chauncey, and Krzysztof Z. Gajos. 2020. Predictive Text Encourages Predictable Writing. In *Proceedings of the 25th International Conference on Intelligent User Interfaces* (Cagliari, Italy) *(IUI '20)*. Association for Computing Machinery, New York, NY, USA, 128–138. https://doi.org/10.1145/3377325.3377523

[3] Daryl J Bem. 1972. Self-perception theory. In *Advances in experimental social psychology*. Vol. 6. Elsevier, 1–62.

[4] 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*. 610–623.

[5] Advait Bhat, Saaket Agashe, Niharika Mohile, Parth Oberoi, Ravi Jangir, and Anirudha Joshi. 2022. Studying writer-suggestion interaction: A qualitative study to understand writer interaction with aligned/misaligned next-phrase suggestion. (2022). Publisher: arXiv.

[6] Xiaojun Bi, Tom Ouyang, and Shumin Zhai. 2014. Both Complete and Correct? Multi-Objective Optimization of Touchscreen Keyboard. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Toronto, Ontario, Canada) *(CHI '14)*. Association for Computing Machinery, New York, NY, USA, 2297–2306. https://doi.org/10.1145/2556288.2557414

[7] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258* (2021).

[8] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[9] Daniel Buschek, Martin Zürn, and Malin Eiband. 2021. The Impact of Multiple Parallel Phrase Suggestions on Email Input and Composition Behaviour of Native and Non-Native English Writers. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 732, 13 pages. https://doi.org/10.1145/3411764.3445372

[10] Mia Xu Chen, Benjamin N. Lee, Gagan Bansal, Yuan Cao, Shuyuan Zhang, Justin Lu, Jackie Tsay, Yinan Wang, Andrew M. Dai, Zhifeng Chen, Timothy Sohn, and Yonghui Wu. 2019. Gmail Smart Compose: Real-Time Assisted Writing. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (Anchorage, AK, USA) *(KDD '19)*. Association for Computing Machinery, New York, NY, USA, 2287–2295. https://doi.org/10.1145/3292500.3330723

[11] 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* (Tokyo, Japan) *(IUI '18)*. Association for Computing Machinery, New York, NY, USA, 329–340. https://doi.org/10.1145/3172944.3172983

[12] M. A. DeVito, J. Birnholtz, J. T. Hancock, and "Platforms. 2017. *people.* and perception: Using affordances to understand self-presentation on social media" in Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing. 740–754 pages.

[13] Mark Dunlop and John Levine. 2012. Multidimensional Pareto Optimization of Touchscreen Keyboards for Speed, Familiarity and Improved Spell Checking. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Austin, Texas, USA) *(CHI '12)*. Association for Computing Machinery, New York, NY, USA, 2669–2678. https://doi.org/10.1145/2207676.2208659

[14] E. Ert, A. Fleischer, and N. Magen. 2016. Trust and reputation in the sharing economy: The role of personal photos in Airbnb. *Tour. Manag* 55 (2016).

[15] Andrew Fowler, Kurt Partridge, Ciprian Chelba, Xiaojun Bi, Tom Ouyang, and Shumin Zhai. 2015. Effects of Language Modeling and Its Personalization on Touchscreen Typing Performance. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul, Republic of Korea) *(CHI '15)*. Association for Computing Machinery, New York, NY, USA, 649–658. https://doi.org/10.1145/2702123.2702503

[16] K. Gero and Lydia B. Chilton. 2019. Metaphoria: An Algorithmic Companion for Metaphor Creation. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (2019).

[17] Mitchell Gordon, Tom Ouyang, and Shumin Zhai. 2016. WatchWriter: Tap and Gesture Typing on a Smartwatch Miniature Keyboard with Statistical Decoding. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 3817–3821. https://doi.org/10.1145/2858036.2858242

[18] Jeffrey T Hancock, Mor Naaman, and Karen Levy. 2020. AI-mediated communication: Definition, research agenda, and ethical considerations. *Journal of Computer-Mediated Communication* 25, 1 (2020), 89–100.

[19] Jeffrey T Hancock, Mor Naaman, and Karen Levy. 2020. AI-Mediated Communication: Definition, Research Agenda, and Ethical Considerations. *Journal of Computer-Mediated Communication* 25, 1 (01 2020), 89–100. https://doi.org/10.1093/jcmc/zmz022 arXiv:https://academic.oup.com/jcmc/article-pdf/25/1/89/32961176/zmz022.pdf

[20] B. Van Der Heide, J. D. D'Angelo, and E. M. Schumaker. 2012. The effects of verbal versus photographic self-presentation on impression formation in Facebook. *J. Commun* 62 (2012).

[21] Po-Sen Huang, Huan Zhang, Ray Jiang, Robert Stanforth, Johannes Welbl, Jack Rae, Vishal Maini, Dani Yogatama, and Pushmeet Kohli. 2019. Reducing sentiment bias in language models via counterfactual evaluation. *arXiv preprint arXiv:1911.03064* (2019).

[22] Maurice Jakesch, Advait Bhat, Daniel Buschek, Lior Zalmanson, and Mor Naaman. 2022. Co-Writing with Opinionated Language Models Affects Users' Views. (2022).

[23] Maurice Jakesch, Megan French, Xiao Ma, Jeffrey T. Hancock, and Mor Naaman. 2019. AI-Mediated Communication: How the Perception That Profile Text Was Written by AI Affects Trustworthiness. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3290605.3300469

[24] Maurice Jakesch, Jeffrey Hancock, and Mor Naaman. 2022. Human Heuristics for AI-Generated Language Are Flawed. *arXiv preprint arXiv:2206.07271* (2022).

[25] Rebecca L Johnson, Giada Pistilli, Natalia Menédez-González, Leslye Denisse Dias Duran, Enrico Panai, Julija Kalpokiene, and Donald Jay Bertulfo. 2022. The Ghost in the Machine has an American accent: value conflict in GPT-3. *arXiv preprint arXiv:2203.07785* (2022).

[26] Anjuli Kannan, Karol Kurach, Sujith Ravi, Tobias Kaufmann, Andrew Tomkins, Balint Miklos, Greg Corrado, Laszlo Lukacs, Marina Ganea, Peter Young, and Vivek Ramavajjala. 2016. Smart Reply: Automated Response Suggestion for Email. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (San Francisco, California, USA) *(KDD '16)*. Association for Computing Machinery, New York, NY, USA, 955–964. https://doi.org/10.1145/2939672.2939801

[27] Sarah Kreps, R Miles McCain, and Miles Brundage. 2022. All the news that's fit to fabricate: AI-generated text as a tool of media misinformation. *Journal of Experimental Political Science* 9, 1 (2022), 104–117.

[28] Stephanie Lin, Jacob Hilton, and Owain Evans. 2021. Truthfulqa: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958* (2021).

[29] Xiao Ma, Jeffery T. Hancock, Kenneth Lim Mingjie, and Mor Naaman. 2017. Self-Disclosure and Perceived Trustworthiness of Airbnb Host Profiles. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (Portland, Oregon, USA) *(CSCW '17)*. Association for Computing Machinery, New York, NY, USA, 2397–2409. https://doi.org/10.1145/2998181.2998269

[30] Xiao Ma, Trishala Neeraj, and Mor Naaman. 2017. A Computational Approach to Perceived Trustworthiness of Airbnb Host Profiles. *Proceedings of the International AAAI Conference on Web and Social Media* 11, 1 (May 2017). https://ojs.aaai.org/index.php/ICWSM/article/view/14937

[31] Enrique Manjavacas, Folgert Karsdorp, Ben Burtenshaw, and Mike Kestemont. 2017. Synthetic Literature: Writing Science Fiction in a Co-Creative Process. In *Proceedings of the Workshop on Computational Creativity in Natural Language Generation (CC-NLG 2017)*. Association for Computational Linguistics, Santiago de Compostela, Spain, 29–37. https://doi.org/10.18653/v1/W17-3904

[32] Debora Nozza, Federico Bianchi, and Dirk Hovy. 2021. HONEST: Measuring hurtful sentence completion in language models. In *The 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics.

[33] Stefan Palan and Christian Schitter. 2018. Prolific. ac—A subject pool for online experiments. *Journal of Behavioral and Experimental Finance* 17 (2018), 22–27.

[34] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1, 8 (2019), 9.

[35] Jack W Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, et al. 2021. Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446* (2021).

[36] Melissa Roemmele and Andrew S. Gordon. 2018. Automated Assistance for Creative Writing with an RNN Language Model. In *Proceedings of the 23rd International Conference on Intelligent User Interfaces Companion* (Tokyo, Japan) *(IUI '18 Companion)*. Association for Computing Machinery, New York, NY, USA, Article 21, 2 pages. https://doi.org/10.1145/3180308.3180329

[37] Dietram A Scheufele. 2000. Agenda-setting, priming, and framing revisited: Another look at cognitive effects of political communication. *Mass communication & society* 3, 2-3 (2000), 297–316.

[38] B. R. Schlenker. 2012. *Self-presentation in Handbook of Self and Identity*. 2nd Ed, (The Guilford Press. 542–570 pages.

[39] E. Schwämmlein and K. Wodzicki. 2012. What to Tell About Me? Self-Presentation in Online Communities. *J. Comput. -Mediat. Commun* 17 (2012).

[40] Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in NLP. *arXiv preprint arXiv:1906.02243* (2019).

[41] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[42] Genta Indra Winata, Andrea Madotto, Zhaojiang Lin, Rosanne Liu, Jason Yosinski, and Pascale Fung. 2021. Language models are few-shot multilingual learners. *arXiv preprint arXiv:2109.07684* (2021).

[43] Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019. Defending against neural fake news. *Advances in neural information processing systems* 32 (2019).