Co-Writing with Opinionated Language Models Affects Users' Views

Maurice Jakesch
Cornell University
Ithaca, New York, USA
mpj32@cornell.edu

Advait Bhat
Microsoft Research
Bengaluru, India

Daniel Buschek
University of Bayreuth
Bayreuth, Germany

Lior Zalmanson
Tel Aviv University
Tel Aviv, Israel

Mor Naaman
Cornell Tech
New York, New York, USA

ABSTRACT

If large language models like GPT-3 preferably produce a particular point of view, they may influence people's opinions on an unknown scale. This study investigates whether a language-model-powered writing assistant that generates some opinions more often than others impacts what users write – and what they think. In an online experiment, we asked participants (N=1,506) to write a post discussing whether social media is good for society. Treatment group participants used a language-model-powered writing assistant configured to argue that social media is good or bad for society. Participants then completed a social media attitude survey, and independent judges (N=500) evaluated the opinions expressed in their writing. Using the opinionated language model affected the opinions expressed in participants' writing and shifted their opinions in the subsequent attitude survey. We discuss the wider implications of our results and argue that the opinions built into AI language technologies need to be monitored and engineered more carefully.

CCS CONCEPTS

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

KEYWORDS

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

ACM Reference Format:
Maurice Jakesch, Advait Bhat, Daniel Buschek, Lior Zalmanson, and Mor Naaman. 2023. Co-Writing with Opinionated Language Models Affects Users' Views. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23), April 23–28, 2023, Hamburg, Germany.* ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3544548.3581196

1 INTRODUCTION

Large language models like GPT-3 [21, 97, 103] are increasingly becoming part of human communication. Enabled by developments in computer hardware and software architecture [97], large language models produce human-like language [56] by iteratively predicting likely next words based on the sequence of preceding words. Applications like writing assistants [38], grammar support [66], and machine translation [45] inject the models' output into what people write and read [51].

Using large language models in our daily communication may change how we form opinions and influence each other. In conventional forms of persuasion, a persuader crafts a compelling message and delivers it to recipients – either face-to-face or mediated through contemporary technology [94]. More recently, user researchers and behavioral economists have shown that technical choice architectures, such as the order of options presented affect people's behavior as well [42, 72]. With the emergence of large language models that produce human-like language [25, 56], interactions with technology may influence not only behavior but also opinions: when language models produce some views more often than others, they may persuade their users. We call this new paradigm of influence *latent persuasion* by language models, illustrated in Figure 1.

Latent persuasion by language models extends the insight that choice defaults affect people's behavior [42, 72] to the field of language and persuasion. Where *nudges* change behavior by making some choices more convenient than others, AI language technologies may shift opinions by making it easy to express certain views but not others. Such influence could be *latent* and hard to pinpoint: choice architectures are visible, but opinion preferences built into language models may be opaque to users, policymakers, and even system developers. While in traditional persuasion, a central designer intentionally creates a message to convince a specific audience, a language model may be opinionated by accident and its opinions may vary by user, product and context.

Prior research on the risks of generative language models has focused on conventional persuasion scenarios, where a human persuader uses language models to automate and optimize the production of content for advertising [39, 61] or misinformation [25, 67, 106]. Initial audits also highlight that language models reproduce stereotypes and biases [23, 54, 83] and support certain cultural values more than others [57]. While emerging research on co-writing with large language models suggests that models become increasingly active partners in people's writing [70, 104, 105], little is known about how the opinions produced by language models affect users' views. Work by Arnold et al. [3] and Bhat et al.

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

ACM ISBN 978-1-4503-9421-5/23/04.
https://doi.org/10.1145/3544548.3581196

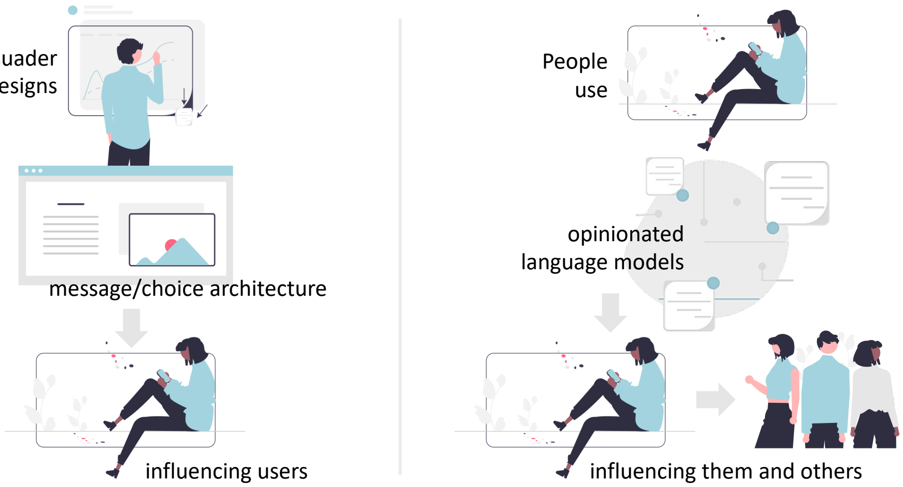

Persuader designs
message/choice architecture
influencing users
People use
opinionated language models
influencing them and others

Figure 1: Conventional technology-mediated persuasion (left) compared to *latent persuasion* by language models (right). In conventional influence campaigns, a central persuader designs an influential message or choice architecture distributed to recipients. In *latent persuasion*, language models produce some opinions more often than others, influencing what their users write, which is, in turn, read by others.

[16, 17] shows that a biased writing assistant may affect movie or restaurant reviews, but whether co-writing with large language models affect users' opinions on public issues remains an open and urgent question.

This study investigates whether large language models that generate certain opinions more often than others affect what their users write and think. In an online experiment (N=1,506), participants wrote a short statement discussing whether social media is good or bad for society. Treatment group participants were shown suggested text generated by a large language model. The model, GPT-3 [103], was configured to either generate text that argued that social media is good for society or text that argued the opposite. Following the writing task, we asked participants to assess social media's societal impact in a survey. A separate sample of human judges (N=500) evaluated the opinions expressed in participants' writing.

Our quantitative analysis tests whether the interactions with the opinionated language model shifted participants' writing and survey opinions. We explore how this opinion shift may have occurred in secondary analyses. We find that both participants' writing *and* their attitude towards social media in the survey were considerably affected by the model's preferred opinion. We conclude by discussing how researchers, AI practitioners, and policymakers can respond to the possibility of latent persuasion by AI language technologies.

2 RELATED WORK

Our study is informed by prior research on social influence and persuasion, interactions with writing assistants, and the societal risks of large language models.

2.1 Social influence and persuasion

Social influence is defined as a shift in an individual's thoughts, feelings, attitudes, or behaviors as a result of interaction with others [92]. While social influence is integral to human collaboration, landmark studies have shown that it can also lead to unreasonable or unethical behavior. On a personal level, people may conform to majority views against their better judgement [6] and obey an authority figure even if it means harming others [80]. On a societal level, researchers have shown that social influence drives speculative markets [93], affects voting patterns [69], and contributes to the spread of unhealthy behaviors such as smoking and obesity [30, 31].

Following the rise of social media, how online interactions affect people's opinions and decisions has been studied extensively. Research has shown that a variety of sources influences users' attitudes and behaviors, including friends, family, experts, and internet celebrities [48, 78]; the latter group was labeled *influencers* due to their influence on a large group of 'followers' [10]. Research has found that in online settings, users can be influenced by non-human entities such as brand pages, bots, and algorithms [41]. Studies have evaluated the influence that technical artifacts such as personalized recommendations, chatbots, and choice architectures have on users' decision-making [15, 35, 50, 72].

The influence that algorithmic entities have on people depends on how people perceive the algorithm, for example, whether they attribute trustworthiness to its recommendations [50, 76]. The influence of algorithms on individuals tends to increase as the environment becomes more uncertain and decisions become more difficult [20]. With the public's growing awareness of developments in artificial intelligence, people may regard *smart* algorithms as a source of authority [2, 60, 76]. There is recent evidence that people may accept algorithmic advice even in simple cases when it is clearly wrong [74]. In the related field of automation, such over-reliance on machine output has been referred to as *automation bias* [86, 87, 102].

2.2 Interaction with writing assistants

Historically, HCI research for text entry has predominantly focused on efficiency [68]. Typical text entry systems attend to language context at the word [18, 98] or sentence level [5, 27]. They suggest one to three subsequent words based on underlying likelihood distributions [40, 43, 49, 89]. More recent systems also provide multiple short reply suggestions [59] or a single long phrase suggestion [29]. More extensive suggestions have traditionally been avoided because the time taken to read and select them might exceed the time required to enter that text manually. Several studies indicate that features such as auto-correction and word suggestions can negatively impact typing performance and user experience [11, 26, 37, 85].

However, with the emergence of larger and more powerful language models [21, 97, 103], there has been a growing interest in design goals beyond efficiency. Studies have investigated interface design factors and interactions with writing assistants that directly or indirectly support inspiration [17, 70, 95, 105], language proficiency [27], story writing [95, 105], text revision [36, 107] or creative writing [33, 46]. Here, language models are framed as *active writing partners* or *co-authors* [70, 104, 105], rather than tools for prediction or correction. There is evidence that a system that suggests phrases rather than words [5] is more likely to be perceived as a collaborator and content contributor by users.

The more writing assistants become *active writing partners* rather than mere tools for text entry, the more the writing process and output may be affected by their "co-authorship". Bhat et al. [17] discuss how writers evaluate the suggestions provided and integrate them into different cognitive writing processes. Work by Singh et al. [95] suggests that writers make active efforts or 'leaps' to integrate generated language into their writing. Buschek et al. [27] conceptualized nine behavior patterns that indicate varying degrees of engagement with suggestions, from ignoring them to chaining multiple ones in a row. Writing with suggestions correlates with shorter and more predictable texts being written [4], along with increased use of standard phrases when writing with a language model [17, 27]. Furthermore, the sentiment of the suggested text may affect the sentiment of the written text [3, 52].

2.3 Societal risks of large language models

Technical advances have given rise to a generation of language models [21] that produces language so natural that humans can barely distinguish it from real human language [56]. Enabled by improvements in computer hardware and the transformer architecture [97], models like GPT-3 [23, 90] have attracted attention for their potential to impact a range of beneficial real-world applications [21]. However, more cautious voices have also warned about the ethical and social risks of harm from large language models [100, 101], ranging from discrimination and exclusion [23, 54, 83] to misinformation [67, 75, 91, 106] and environmental [96] and socioeconomic harms [14].

Comparatively little research has considered widespread shifts in opinion, attitude, and culture that may result from a comprehensive deployment of generative language models. It is known that language models work and perform better for the languages and contexts they are trained in (primarily English or Mandarin Chinese) [23, 91, 103]. Small-n audits have also suggested that the values embedded in models like GPT-3 were more aligned with reported dominant US values than those upheld in other cultures [57]. Work by Jakesch et al. [55] has highlighted that the values held by those developing AI systems differ from those of the broader populations interacting with the systems. The adjacent question of AI alignment – how to build AI systems that act in line with their operators' goals and values – has received comparatively more attention [7], but primarily from a control and safety angle.

A related topic, the political repercussions of social media and recommender systems [108], has received extensive research attention, however. After initial excitement about social media's democratic potential [62], it became evident that technologies that affect public opinion will be subject to powerful political and commercial interests [22]. Rather than mere technical platforms, algorithms become constitutive features of public life [47] that may change the political landscape [1]. Even without being designed to shift opinions, it has been found that algorithms may contribute to political polarization by reinforcing divisive opinions [9, 24, 32].

3 METHODS

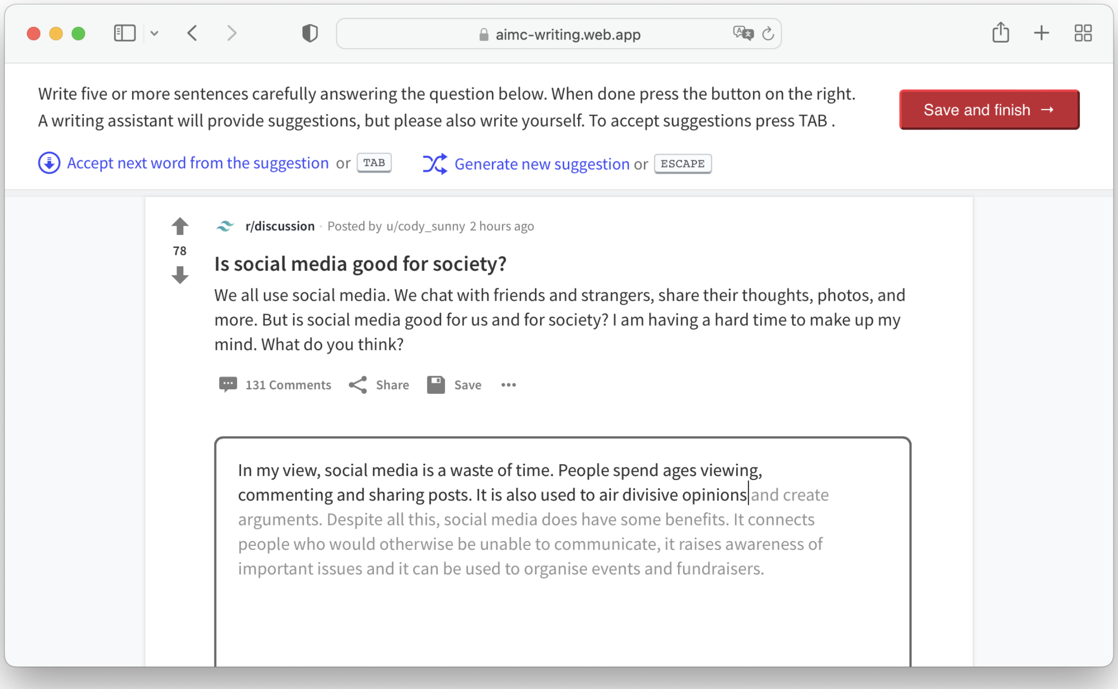
To investigate whether interacting with opinionated language models shifts people's writing and affects people's views, we conducted an online experiment asking participants (N=1,506) to respond to a social media post in a simulated online discussion using a writing assistant. The language model powering this writing assistant was configured to generate text supporting one or the other side of the argument. We compared the essays and opinions of participants to a control group that wrote their social media posts without writing assistance.

3.1 Experiment design

To study interactions between model opinion and participant opinion in a possibly realistic and relevant setting, we created the scenario of an opinionated discussion on social media platforms like Reddit. Such discussions have a large readership [79], pertain to political controversies, and are plausible application settings for writing assistants and language models. We searched ProCon.org[1], an online resource for research on controversial issues, to identify a discussion topic. We selected "Is Social Media Good for Society?" as a discussion topic. We chose this topic because it is an easily accessible discussion topic that is politically relevant but not considered so controversial that entrenched views may limit constructive debate.

To run the experiment, we created a custom experimental platform combining a mock-up of a social media discussion page, a rich-text editor, and a writing assistant. The assistant was powered by a language generation server and included comprehensive logging tools. To provide a realistic-looking social media mock-up, we copied the design of a Reddit discussion page and drafted a

[1] https://www.procon.org/


aimc-writing.web.app
Save and finish →
Accept next word from the suggestion or TAB
Generate new suggestion or ESCAPE
78
r/discussion · Posted by u/cody_sunny 2 hours ago
131 Comments
Share
Save

Write five or more sentences carefully answering the question below. When done press the button on the right. A writing assistant will provide suggestions, but please also write yourself. To accept suggestions press TAB .

Is social media good for society?

We all use social media. We chat with friends and strangers, share their thoughts, photos, and more. But is social media good for us and for society? I am having a hard time to make up my mind. What do you think?

In my view, social media is a waste of time. People spend ages viewing, commenting and sharing posts. It is also used to air divisive opinions and create arguments. Despite all this, social media does have some benefits. It connects people who would otherwise be unable to communicate, it raises awareness of important issues and it can be used to organise events and fundraisers.

Figure 2: Screenshot of the writing task. The task is shown on the top of the page, followed by usage instructions for the writing assistant. Below, participants read a Reddit-style discussion post to which they were asked to reply. The writing assistant displayed writing suggestions (shown in grey) extending participants' text. The participant in the screenshot wrote an argument critical of social media, but the model is configured to argue that social media is *good* for society.

question based on the ProCon.org discussion topic. Figure 2 shows a screenshot of the experiment. We asked participants to write at least five sentences expressing their take on social media's societal impact. We randomly assigned participants to three different treatment groups:

(1) *Control group:* participants wrote their answers without a writing assistant.
(2) *Techno-optimist language model treatment:* participants were shown suggestions from a language model configured to argue that social media is good for society.
(3) *Techno-pessimist language model treatment:* participants received suggestions from a language model configured to argue that social media is bad for society.

3.2 Building the writing assistant

Similar to Google's *Smart Compose* [29] and Microsoft's predictive text in Outlook, the writing assistant in the treatment groups suggested possible continuations (sometimes called "completions") to text that participants had entered. We integrated the suggestions into a customized version of the rich-text editor Quill.js[2]. The client sent a generation request to the server whenever a participant paused their writing for a certain amount of time (750ms). Including round-trip and generation time, a suggestion appeared on participants' screens about 1.5 seconds after they paused their writing.

When the editor client received a text suggestion from the server, it revealed the suggestion letter by letter with random delays calibrated to resemble a co-writing process (cf. [71]). Once the end of a suggested sentence was reached, the editor would pause and request from the server an extended generation until at least two sentences had been suggested. Participants could accept each suggested word by pressing the tab key or clicking an accept button on the interface. In addition, they could reset the generation, requesting a new suggestion by pressing a button or key.

We hosted the required cloud functions, files, and interaction logs on Google's Firebase platform.

3.3 Configuring an opinionated language model

In this study, we experimented with language models that *strongly* favored one view over another. We chose a strong manipulation as we wanted to explore the *potential* of language models to affect users' opinions and understand whether they could be used or abused to shift people's views [8].

We used GPT-3 [23] with manually designed prompts to generate text suggestions for the experiment in real-time. Specifically, we accessed OpenAI's most potent 175B parameter model ("text-davinci-002"). We used temperature sampling, a method for choosing a specific next token from the set of likely next tokens inspired by statistical thermodynamics. We set the sampling temperature (randomness parameter) to 0.85 to generate suggestions that are varied and creative. We set the frequency and presence penalty

[2] https://quilljs.com/

parameters to 1 to reduce the chance that the model suggestions would become repetitive. We also prevented the model from producing new lines, placeholders, and list by setting logit bias parameters that reduced the likelihood of the respective tokens being selected.

We evaluated different techniques to create an opinionated model, i.e., a model that *likely supports a certain side of the debate* when generating a suggestion. We used prompt design [73], a technique for guiding frozen language models to perform a specific task. Rather than updating the weights of the underlying model, we concatenated an engineered prompt to the input text to increase the chance that the model generates a certain opinion. Specifically we inserted the prefix *"Is social media good for society? Explain why social media is good/bad for society:"* before participants' written texts when generating continuation suggestions. The engineered prompt was not visible to participants in their editor UI; it was inserted in the backend before generation and removed from the generated text before showing it to participants.

Initial experimentation and validation suggested that the prompt produced the desired opinion in the generated text, but when participants strongly argued for another opinion in their writing, the model's continuations would follow their opinion. In addition to the prefix prompt, we thus developed an infix prompt that would be inserted throughout participants' writing to reinforce the desired opinion. We inserted the snippet (*"One sentence continuing the essay explaining why social media is good/bad:"*) right before the last sentence that participants had written. This additional prompt guided the model's continuation towards the target opinion even if participants had articulated a different opinion earlier in their writing. Validation of the model opinion configuration is provided in section 4.5. We also experimented with fine-tuning [53] to guide the models' opinion, but the fine-tuned models did not consistently produce the intended opinion.

3.4 Outcome measures and covariates

We collected different types of outcome measures to investigate interactions between participants' opinions and the model opinion:

Opinion expressed in the post: To evaluate expressed opinion, we split participants' written texts into sentences and asked crowd workers to evaluate the opinion expressed in each sentence. Each crowd worker assessed 25 sentences, indicating whether each argued that social media is good for society, bad, or both good and bad. A fourth label was offered for sentences that argued neither or were unrelated. For example, *"Social media also promotes cyber bullying which has led to an increase in suicides" (P#421)* was labeled as arguing that social media is bad for society, while *"Social media also helps to create a sense of community" (P#1169)* was labeled as *social media is good for society.* We collected one to two labels for each sentence participants wrote and collected labels for a sample of the writing assistant's suggestions. In sentences where we collected multiple labels, the labels provided by different raters agreed 84.1% of the time (Cohen's $\kappa = 0.76$).

Real-time writing interaction data: We gathered comprehensive interaction logs at the key-stroke level of how participants interacted with the model's suggestions. We recorded which text the participant had written, what text the model had suggested, and what suggestions participants had accepted from the writing assistant. We measured how long they paused to consider suggestions and how many suggestions they accepted.

Opinion survey (post-task): After finishing the writing task, participants completed an opinion survey. The central question, "Overall, would you say social media is good for society?" was designed to assess shifts in participants' attitude. This question was not shown immediately after the writing task to reduce demand effects. Secondary questions were asked to understand participants' opinions in more detail: "How does social media affect your relationships with friends and family?", "Does social media usage lead to mental health problems or addiction?", "Does social media contribute to the spread of false information and hate?", "Do you support or oppose government regulation of social media companies?" The questions were partially adapted from Morning Consults' National Tracking Poll [34]; answers were given on typical 3- and 5-point Likert scales.

User experience survey (post-task): Participants in the treatment groups completed a survey about their experience with the writing assistant following the opinion survey. They were asked, "How useful was the writing assistant to you?", whether "The writing assistant understood what you wanted to say" and whether "The writing assistant was knowledgeable and had expertise." To explore participants' awareness of the writing assistant's opinion and its effect on their own views, we asked them whether "The writing assistant's suggestions were reasonable and balanced" and whether "The writing assistant inspired or changed my thinking and argument." Answers were given on a 5-point Likert scale from "strongly agree" to "strongly disagree." An open-ended question asked participants what they found most useful or frustrating about the writing assistant.

Covariates: We asked participants to self-report their age, gender, political leaning, and their highest level of education at the end of the study. We also constructed a "model alignment" covariate estimating whether the opinion the model supported was aligned with the participant's opinion. We did not ask participants to report their overall judgment before the writing task to avoid commitment effects. Instead, we asked them at the end of the study whether they believed social media was good for society before participating in the discussion. While imperfect, this provides a proxy for participants' pre-task opinions. It is biased by the treatment effect observed on this covariate, which amounts to 14% of its standard deviation.

3.5 Participant recruitment

We recruited 1,506 participants (post-exclusion) for the writing task, corresponding to 507, 508, and 491 individuals in the control, techno-optimist, and techno-pessimist treatment groups, respectively. The sample size was calculated based on effect sizes observed in the pilot studies' post-task question, "Overall, would you say social media is good for society?" at a power of 80%. The sample was recruited through Prolific [84]. The sample included US-based participants at least 18 years old (M= 37.7, SD= 14.2); 48.5% self-identified as female, and 48.6% identified as male. 38 participants identified as non-binary and eight preferred to self-describe or not disclose their gender identity. Six out of ten indicated liberal leanings; 57.1% had received at least a Bachelor's degree. Participants who failed the pre-task attention check (8%) were excluded. Six percent of participants admitted to the task did not finish it. We paid participants $1.50 for an average task time of 5.9 minutes based on an hourly compensation rate of $15. For the labeling task, we recruited a similar sample of 500 participants through Prolific. The experimental protocols were approved by the Cornell University Institutional Review Board.

3.6 Data sharing

The experiment materials, analysis code and data collected are publicly available through an Open Science repository (https://osf.io/upgqw/). A research assistant screened the data, and records with potentially privacy-sensitive information were removed before publication.

4 RESULTS

We first analyze the opinions participants expressed in their social media posts. We then examine whether participants may have accepted the models' suggestions out of mere convenience and whether the model influenced participants' opinions in a later survey. Finally, we present data on participants' perceptions of the model's opinion and influence. The reported statistics are based on a logistic regression model.

4.1 Did the interactions with the language model affect participants' writing?

Figure 3 shows how often participants in each of the treatment conditions (y-axis) argued that social media is good or bad for society (x-axis) in their writing. The social media posts written by participants in the control group (middle row) were slightly critical of social media: They argued that social media is bad for society in 38% and that social media is good in 28% of their sentences. In about 28% of their sentences, control group participants argued that social media is both good and bad, and 11% of their sentences argued neither or were unrelated.

Participants who received suggestions from a language model supportive of social media (top row of Figure 3) were 2.04 times more likely than control group participants (p<0.0001, 95% CI [1.83, 2.30]) to argue that social media is good. In contrast, participants who received suggestions from a language model that criticized social media (bottom row) were 2.0 times more likely (p<0.0001, 95% CI [1.79, 2.24] to argue that social media is bad than control group participants. We conclude that using an opinionated language model affected participants' writing such that the text they wrote was more likely to support the model's preferred view.

4.2 Did participants accept the model's suggestions out of mere convenience?

Participants may have accepted the models' suggestions out of convenience, even though the suggestions did not match what they would have wanted to say. Paid participants in online studies, in particular, may be motivated to accept suggestions to swiftly complete the task.

Our data shows that, across conditions and treatments, most participants did not blindly accept the model's suggestions but interacted with the model to co-write their social media posts. On average, participants wrote 63% of their sentences themselves without accepting suggestions from the model (compare Figure 5). About 25% of participants' sentences were written by both the participant and the model, which typically meant that the participant wrote some words and accepted the model's remaining sentence suggestion. Only 11.5% of sentences were fully accepted from the model. Participants whose personal views were likely aligned with the model were more likely to accept suggestions, while participants with opposing views accepted fewer suggestions. About one in four participants did not accept any model suggestion, and one in ten participants had more than 75% of their post written by the model.

4.2.1 Did conveniently accepted suggestions increase the observed differences in written opinion? The writing of participants who spent little time on the task was more affected by the model's opinion. We use the time participants took to write their posts to estimate to what extent they may have accepted suggestions without due consideration. For a concise statistical analysis, we treat the ordinal opinion scale as an interval scale. Since the opinion scale has comparable-size intervals and a zero point, continuous analysis is meaningful and justifiable [64]. We treat "social media is bad for society" as -1 and "social media is good for society" as 1. Sentences that argue both or neither are treated as zeros.

Figure 6 shows the mean opinion expressed in participants' social media posts depending on treatment group and writing time. The left panel shows participants' expressed opinions across times for reference, with a mean opinion difference of about 0.29 (p<0.001, 95% CI [0.25, 0.33], SD=0.58) between each treatment group and the control group (corresponding to a large effect size of d=0.5). Participants who took little time to write them (less than 160 seconds, left-most data in right panel) were more affected by the opinion of the language model (0.38, p<0.001, 95% CI [0.31, 0.45]). Our analysis shows that accepting suggestions out of convenience has contributed to the differences in the written opinion. However, even for participants who took four to six minutes to write their posts, we observed significant differences in opinions across treatment groups (0.20, p<0.001, 95% CI [0.13, 0.27], corresponding to a treatment effect of d=0.34).

4.3 Did the language model affect participants' opinions in the attitude survey?

The opinion differences in participants' writing may be due to shifts in participants' actual opinion caused by interacting with the opinionated model. We evaluate whether interactions with the language model affected participants' attitudes expressed in a post-task survey asking participants whether they thought social media was good for society. An overview of participants' answers is shown in Figure 4.

The figure shows the frequency of different survey answers (x-axis) for the participants in each condition (y-axis). Participants who did not interact with the opinionated models (middle row in Figure 4) were balanced in their evaluations of social media: 33% answered that social media is not good for society (middle, blue); 35% said social media is good for society. In comparison, 45% of

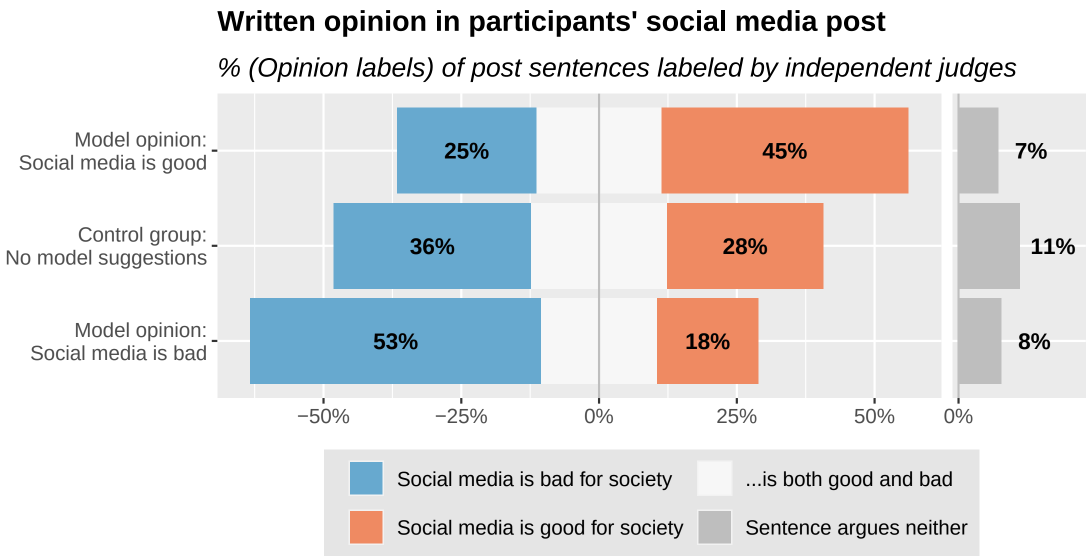


Figure 3: Participants assisted by a model supportive of social media were more likely to argue that social media is good for society in their posts (and vice versa). N_s=9,223 sentences written by N_p=1,506 participants evaluated by N_j=500 judges. The y-axis indicates whether participants wrote their social media posts with assistance from an opinionated language model that was supportive (top) or critical of social media (bottom). The x-axis shows how often participants argued that social media is bad for society (blue), good for society (orange), or both good and bad (white) in their writing.

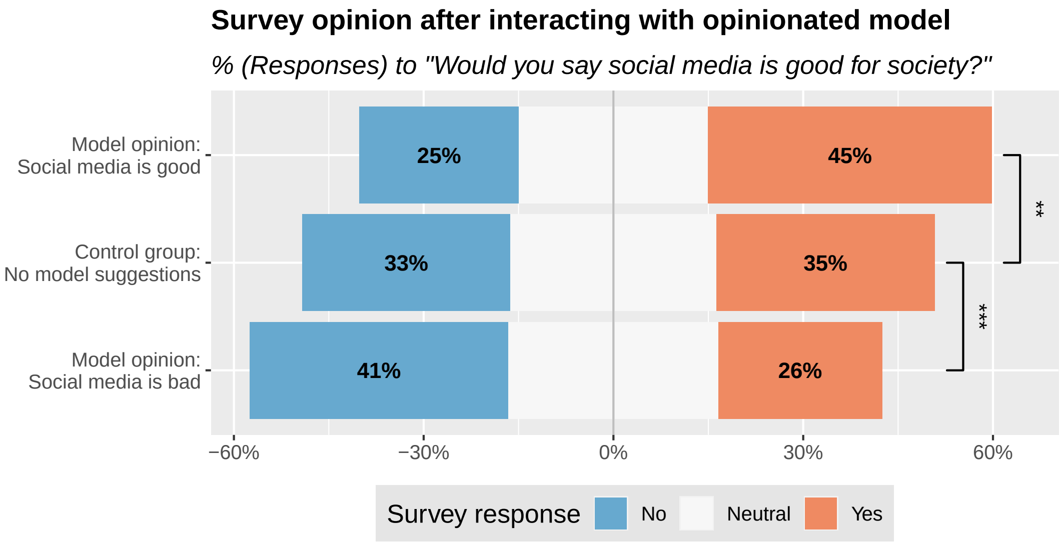


Figure 4: Participants interacting with a model supportive of social media were more likely to say that social media is good for society in a later survey (and vice versa). N_r=1,506 survey responses by N_r=1,506 participants. The y-axis indicates whether participants received suggestions from a model supportive or critical of social media during the writing task. The x-axis shows how often they said that social media was good for society (orange) or not (blue) in a subsequent attitude survey. Undecided participants are shown in white. Brackets indicate significant opinion differences at the $^{**}p<0.005$ and $^{***}p<0.001$ level.

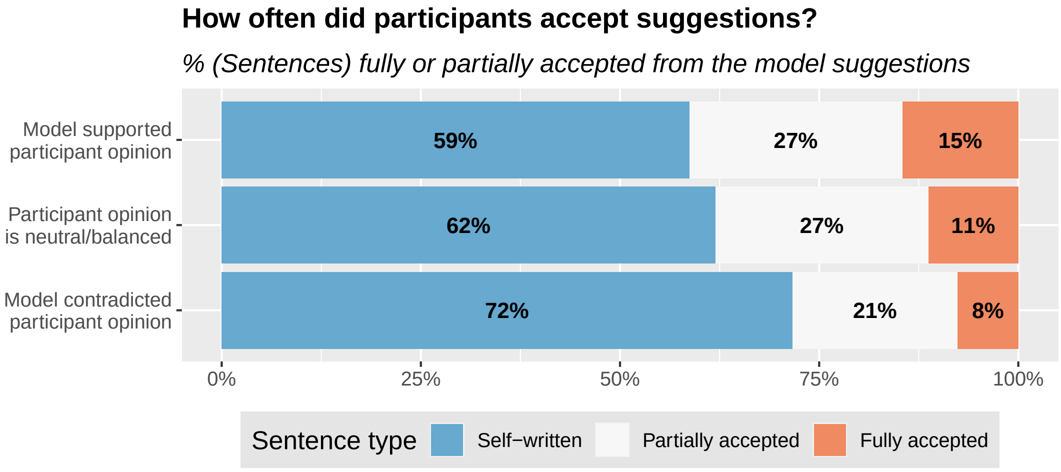


Figure 5: Participants were more likely to accept suggestions if the model's opinion aligned with their own views N_s=6,142 sentences by N_p=1,000 participants. The x-axis shows how many of the sentences participants had written themselves (blue), together with the model (white), or fully accepted from the model's suggestions (orange). The y-axis disaggregates the data based on whether the model suggestions were in line with participants' likely pre-task opinion.

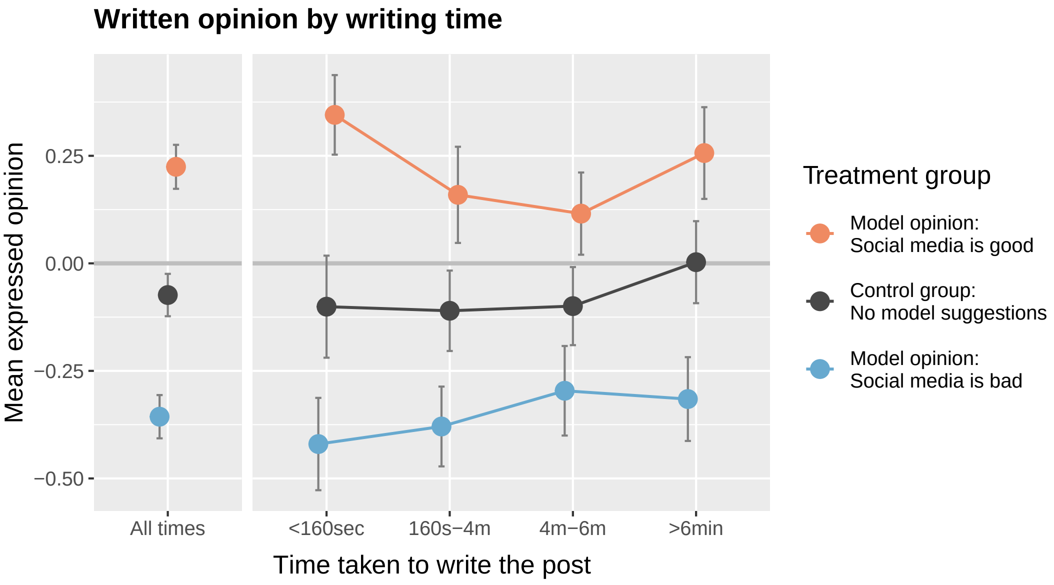


Figure 6: The opinion differences in participants' writing were larger when they finished the task quickly. N=1,506. The y-axis shows the mean opinion expressed in participants' social media posts based on aggregated sentence labels ranging from -1 for "social media is bad for society" to 1 for "social media is good for society". The x-axis indicates how much time participants took to write their posts. For reference, the left panel shows expressed opinions aggregated across writing times.

participants who interacted with a language model supportive of social media (top row) answered that social media is good for society. Converting participants' answers to an interval scale, this difference in opinion corresponds to an effect size of d=0.22 (p<0.001). Similarly, participants that had interacted with the language model critical of social media (bottom row) were more likely to say that social media was bad for society afterward (d=0.19, p<0.005).

4.3.1 Did the opinionated model gradually convince the participant?

While we cannot ascertain the mechanism of persuasion, our results provide further insight into how this process might have occurred.

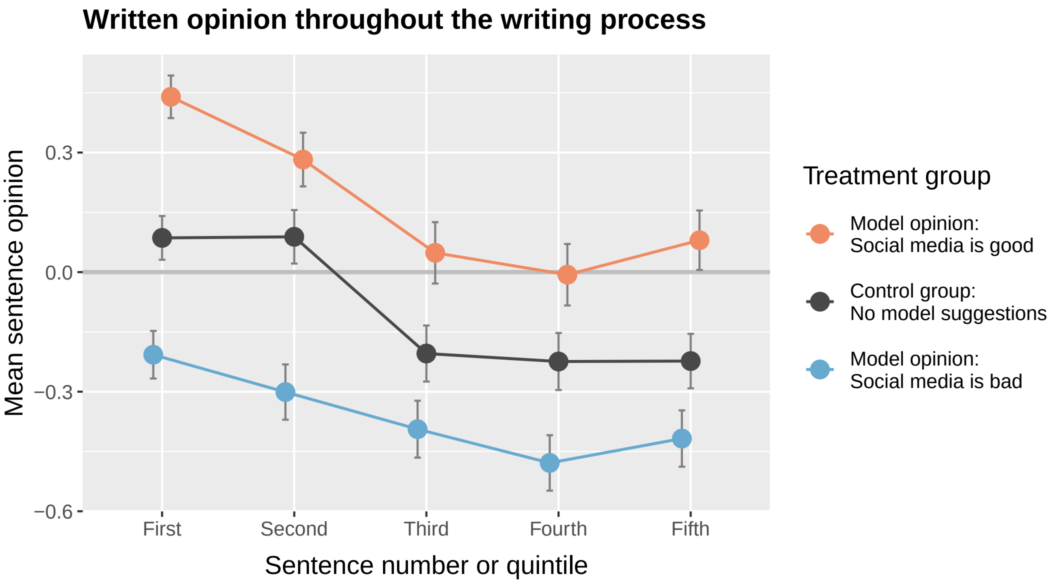


Figure 7: Participants' writing was affected by the model equally throughout the writing process. N_s=9,223 sentences by N_p=1,506 participants. The y-axis shows the mean opinion expressed in participants' sentences. The x-axis indicates whether the sentence was positioned earlier or later in participants' social media posts. Since most participants wrote five sentences as requested, the quintiles roughly correspond to sentence numbers.

Figure 7 shows how participants' written opinions evolved throughout their writing process. In the control group (shown in black), participants tended to start their posts with two positive statements, followed by two more critical statements and an overall critical conclusion. Participants interacting with a model that evaluated social media positively (orange) consistently evaluated social media more favorably throughout their entire statement. Participants interacting with a model critical of social media (blue) also wrote sentences that were more critical of social media, starting with their first sentence. Similar to the control group, they were more positive at the beginning and more critical towards the end of their post, showing that the writing assistant augmented rather than replaced their narrative.

4.4 Were participants aware of the model's opinion and influence?

After the writing task, we asked treatment group participants about their experience with the writing assistant. We use their answers to estimate to what extent they were aware of the model's opinion and influence.

The vast majority of participants thought the language model had expertise and was knowledgeable – even if it contradicted their personal views. As shown in Figure 8, 84% of participants said that the assistant was knowledgeable and had expertise when the language model supported their opinion. When the model contradicted their opinion, only 15% of participants said that it was not knowledgeable or lacked expertise.

While the language model was configured to support one specific side of the debate, the majority of participants said that the model's suggestions were balanced and reasonable. Figure 9 shows that, in the group of participants whose opinion was supported by the model, only 10% noticed that its suggestions were imbalanced (top row in blue). When the model contradicted participants' opinions, they were more likely (30%) to notice its skew, but still, more than half agreed that the model's suggestions were balanced and reasonable (bottom row in orange).

Figure 10 shows that the majority of participants were not aware of the model's effect on their writing. Participants using a model aligned with their view – and accepting suggestions more frequently – were slightly more aware of the model's effect (34%, top row in orange). In comparison, only about 20% of the participants who did not share the model's opinion believed that the model influenced them. Overall, we conclude that participants were often unaware of the model's opinion and influence.

4.5 Robustness and validation

We finally validate that the experimental manipulation worked as intended and address potential concerns about experimenter demand effects.

4.5.1 Did manipulating the models' opinion work as intended? To validate that the prompting technique led to model output opinionated as intended, we sampled a subset of all suggestions shown to participants and asked raters in the sentence labeling task to indicate the opinion expressed in each. We found that of the full sentences suggested by the model, 86% were labeled as supporting the intended view, and 8% were labeled as balanced. For partially suggested sentences, that is, suggestions where the participants

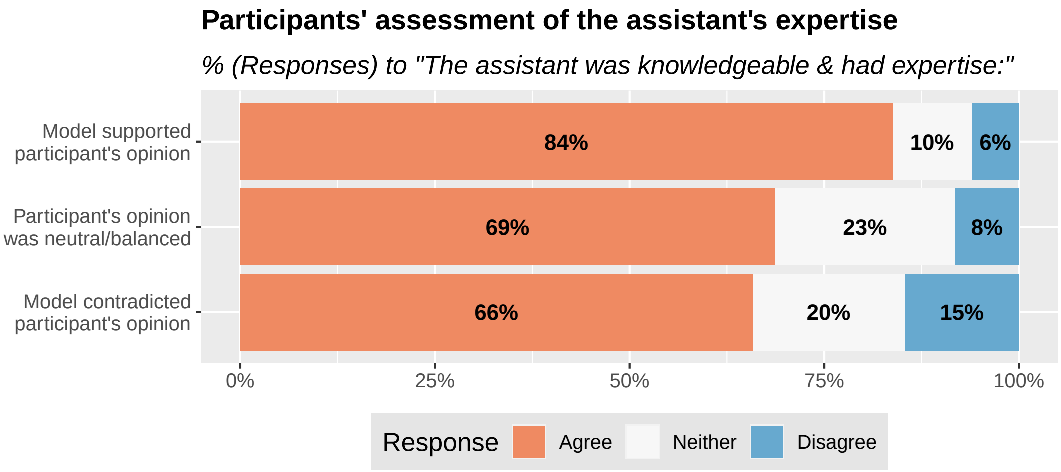


Figure 8: Participants viewed the model as knowledgeable – even if it did not share their opinion. N_p=1,000 treatment group participants. The x-axis indicates whether participants believed the language model had expertise. The y-axis indicates whether the model's opinion was aligned with participants' views.

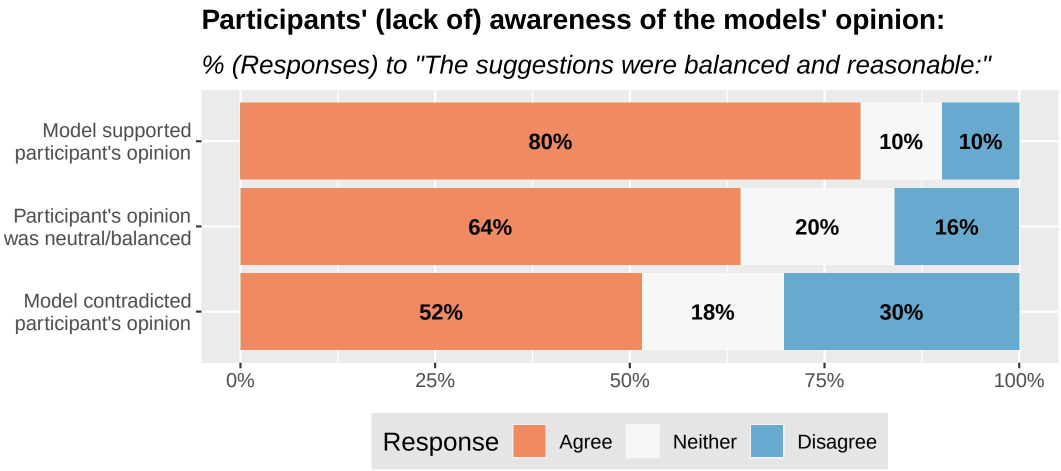


Figure 9: Participants were often unaware of the model's opinion. N_p=1,000 treatment group participants. The x-axis indicates whether participants found the model's suggestions balanced and reasonable. The y-axis indicates whether the model's opinion was aligned with participants' personal views.

had already begun a sentence and the model completed it, the ratio of suggestions that were opinionated as intended dropped to 62% (another 19% argued that social media is both good and bad). Overall, these numbers indicate that the prompting technique guided the model to generate the target opinion with a high likelihood.

4.5.2 Could participants have accepted the model suggestion and shifted their opinion to satisfy the experimenters? As in all subject-based research, there is a chance that participants adapted their behavior to fit their interpretation of the study's purpose. However, we have reason to believe that demand effects do not threaten the validity of our results. When participants were asked what they perceived as the purpose of the study, most thought we were studying what people think about social media or how they use writing assistants. Only about 14% mentioned that we might be studying the assistants' effect on people's opinions. Further, based on our post-task survey, most participants were not aware of the model's opinion and believed that the model did not affect their argument. These results suggest that participants did not adapt their views because they felt the research team expected them to.

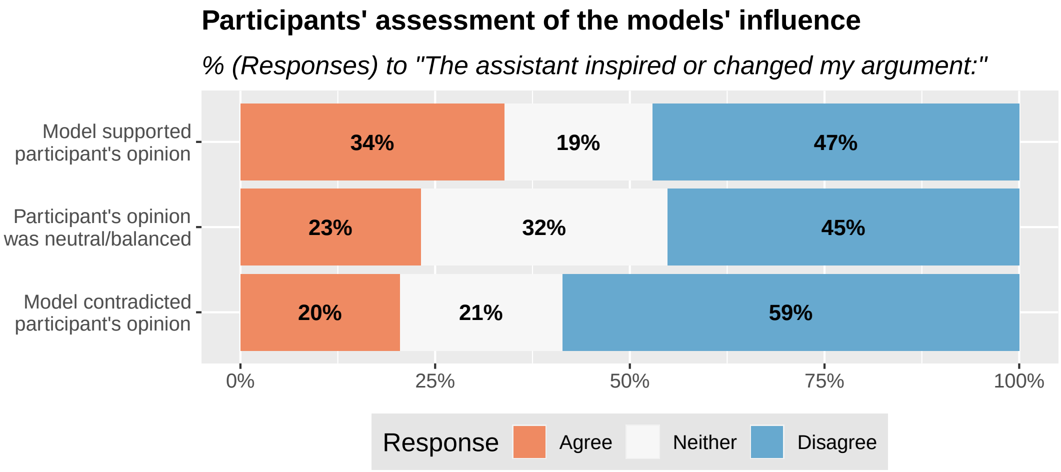


Figure 10: Participants interacting with a model that supported their opinion were more likely to indicate that the model affected their argument. N_p=1,000 treatment group participants. The x-axis indicates whether participants thought that the model affected their argument. The y-axis indicates whether the model's opinion was aligned with participants' personal views.

5 DISCUSSION

The findings show that opinionated AI language technologies can affect what users write and think. In our study, participants assisted by an opinionated language model were more likely to support the model's opinion in a simulated social media post than control group participants who did not interact with a language model. Even participants who took five minutes to write their post – ample time to write the five required sentences – were significantly affected by the model's preferred view, showing that conveniently accepted suggestions do not explain the model's influence. Most importantly, the interactions with the opinionated model also led to opinion differences in a later attitude survey. The opinion shifts in the survey suggest that the differences in written opinion were associated with a shift in personal attitudes. We attribute the shifts in written opinion and post-task attitude to a new form of technology-mediated influence that we call *latent persuasion* by language models.

5.1 Theoretical interpretation

The literature on social influence and persuasion [92] provides ample evidence that our thoughts, feelings, and attitudes shift due to interaction with others. Our results demonstrate that co-writing with an opinionated language model similarly shifted people's writing and attitudes. We discuss below how *latent persuasion* by AI language technologies extends and differs from traditional social influence and conventional forms of technology-mediated persuasion [94]. We consider how the model's influence can be explained by discussing two possible vectors of influence inspired by social influence theory [92]–informative and normative persuasion– and a third vector of influence extending the nudge paradigm [42, 72] to the realm of opinions.

5.1.1 Informational influence. The language model may have influenced participants' opinions by providing new information or compelling arguments, that is, through *informational influence* [81]. Some of the suggestions the language model provided may have made participants think about benefits or drawbacks of social media that they would not have considered otherwise, thus influencing their thinking. While the language model may have provided new information to writers in some cases, our secondary findings indicate that *informational influence* may not fully explain the observed shifts in opinion. First, the model influenced participants consistently throughout the writing process. Had the language models influenced participants' views through convincing arguments, one would expect a gradual or incremental change of opinion, as has been observed for human co-writers [63]. Further, our participants were largely unaware of the language model's skewed opinion and influence. The lack of awareness of the models' influence supports the idea that the model's influence was not only through conscious processing of new information but also through the subconscious [88] and intuitive processes [58].

5.1.2 Normative influence. The language model may have shifted participants' views through *normative influence* [81]. Under normative influence, people adapt their opinions and behaviors based on a desire to fulfill others' expectations and gain acceptance. This explanation aligns with the *computers are social actors* paradigm [82], where the writing assistant may have been perceived as an independent social actor. People may have felt the need to reciprocate the language model, applying the social heuristics they apply in interactions with other humans. The *normative influence* explanation is supported by the finding that participants in our experiment attributed a high degree of expertise to the assistant (see Figure 8). The wider literature similarly suggests that people may regard AI systems as authoritative sources [2, 60, 76]. However, our experimental design presented the language model as a support tool and did not personify the assistant. An ad-hoc analysis of participants' comments on the assistant suggested that they did not feel obliged to reciprocate or comply with the models' suggestions, indicating that the strength of normative influence may have been limited.

5.1.3 Behavioral influence. Large language models may affect people's views by changing behaviors related to opinion formation. The suggestions may have interrupted participants' thought processes and driven them to spend time evaluating the suggested argument [17, 27]. Similar to *nudges*, the suggestions changed participants' behavior, prompting participants to consider the models' view and even accept it in their writing. According to self-perception theory [13], such changes in behavior may lead to changes in opinion. People who do not have strongly formed attitudes may infer their opinion from their own behavior. Even participants with preformed opinions on the topic may have changed their attitudes by being encouraged to communicate a belief that runs counter to their own belief [12, 99]. The finding that the model strongly influenced participants who accepted the models' suggestions frequently corroborates that some of the opinion influence has been through behavioral routes. The *behavioral influence* route implies that the user interface and interaction design of AI language systems mediate the model's influence as they determine when, where, and how the generated opinions are presented.

We conclude that further research will be required to identify the mechanisms behind *latent persuasion* by language models. Our secondary findings suggest that the influence was at least partly subconscious and not simply due to the convenience and new information that the language model provided. Rather, co-writing with the language model may have changed participants' opinion formation process on a behavioral level.

5.2 Implications for research and industry

Our results caution that interactions with opinionated language models affect users' opinions, even if unintended. The results also show how simple it is to make models highly opinionated using accessible methods like prompt engineering. How can researchers, AI practitioners, and policymakers respond to this finding? We believe that our results imply that we must be more careful about the opinions we build into AI language technologies like GPT-3.

Prior work on the societal risks of large language models has warned that models learn stereotypes and biases from their training data [14, 28, 44] that may be amplified through widespread deployments [19]. Our work highlights the possibility that large language models reinforce not only stereotypes but all kinds of opinions – from whether social media is good to whether people should be vegetarians and who should be the next president. Initial tools have been developed for monitoring and mitigating generated text that is discriminating [23, 54, 83] or otherwise offensive [7]. We have no comparable tools for monitoring the opinions built into large language models and in the text they generate during use. A first exploration of the opinions built into GTP-3 by Johnson et al. [57] suggests that the model's preferred views align with dominant US public opinion. In addition, a version of GPT trained on 4chan data led to controversy about the ideologies that training data should not contain. We need theoretical advancements and a broader democratic discourse on what kind of opinions a well-designed model should ideally generate.

Beyond unintentional opinion shifts through carelessly calibrated models, our results raise concerns about new forms of targeted opinion influence. If large language models affect users' opinions, their influence could be used for beneficial social interventions, like reducing polarization in hostile debates or countering harmful false beliefs. However, the persuasive power of AI language technology may also be leveraged by commercial and political interest groups to amplify views of their choice, such as a favorable assessment of a policy or product. In our experiment, we have explored the scenario of influence through a language-model-based writing assistant in an online discussion, but opinionated language models could be embedded in other applications like predictive keyboards, smart replies, and voice assistants. Like search engine and social media network operators [65], operators of these applications may choose to monetize the persuasive power of their technology.

As researchers, we can advance an early understanding of the mechanisms and dangers of *latent persuasion* through AI language technologies. Studies that investigate how *latent persuasion* differs from other sorts of influence, how it is mediated by design factors and users' traits, and engineering work on how to measure and guide model opinions can support product teams in reducing the risk of misuse and legislators in drafting policies that preempt harmful forms of *latent persuasion*.

5.3 Limitations and generalizability

As appropriate for an early study, our experiment has several limitations: We only tested whether a language model affected participants' views on a single topic. We chose this topic as people had mixed views on it and were willing to deliberate. Whether our findings generalize to other topics, particularly where people hold strong entrenched opinions, needs to be explored in future studies. Further, we only looked at one specific implementation of a writing assistant powered by GPT-3. Interacting with different language models through other applications, such as a predictive keyboard that only suggests single words or an email assistant that handles entire correspondences, may lead to different influence outcomes.

Our results provide initial evidence that language models in writing assistance tasks affect users' views. How large is this influence compared to other types of influence, and to what extent effects persist over time, will need to be explored in future studies. For this first experiment, we created a *strongly opinionated* model. In most cases, model opinions in deployed applications will be less definite than in our study and subject to chance variation. However, our design also underestimates the opinion shifts that even weakly opinionated models could cause: In the experiment, participants only interacted with the model once. In contrast, people will regularly interact with deployed models over an extended period. Further, in real-world settings, people will not interact with models individually, but millions will interact with the same model, and what they write with the model will be read by others. Finally, when language models insert their preferred views into people's writing, they increase the prevalence of their opinion in future training data, leading to even more opinionated future models.

5.4 Ethical considerations

The harm participants incurred through interacting with the writing assistant in our study was minimal. The opinion shift was likely transient, inconsequential, and not greater than shifts ordinarily encountered in advertising on the web and TV. Yet, given the weight of our research findings, we decided to share our results with all participants in a late educational debrief: In a private message, we invited crowdworkers who had participated in the experiment and pilot studies to a follow-up task explaining our findings. We reminded participants of the experiment, explained the experimental design, and presented our results in understandable language. We also provided them with a link to a website with a nonpartisan overview of the pros and cons of social media and asked them whether they had comments about the research. 1,469 participants completed the educational debrief in a median time of 109 seconds, for which they received a bonus payment of $0.50. We asked participants for open-ended feedback on our experiment so they could voice potential concerns. 839 participants provided open-ended comments on our experiment and results. Their feedback was exceptionally positive and is included in the Open Science Repository.

Considering the broader ethical implications of our results, we are concerned about misuse. On the one hand, we have shown how simple it is to create highly opinionated models. Our results might motivate some to develop technologies that exploit the persuasive power of AI language technology. In disclosing a new vector of influence, we face ethical tensions similar to cybersecurity researchers: On the one hand, publicizing a new vector of influence increases the chance that someone will exploit it; on the other hand, only through public awareness and discourse effective preventive measures can be taken at the policy and development level. While risky, decisions to share vulnerabilities have led to positive developments in computer safety [77]. We hope our results will contribute to an informed debate and early mitigation of the risks of opinionated AI language technologies.

ACKNOWLEDGMENTS

We thank Yujie Shao for her support in performing the pilot studies for this project. This material is based upon work supported by the National Science Foundation under Grant No. CHS 1901151/1901329 and the German National Academic Foundation. Part of the work is funded by the Bavarian State Ministry of Science and the Arts and coordinated by the Bavarian Research Institute for Digital Transformation (bidt).

REFERENCES

[1] Sinan Aral and Dean Eckles. 2019. Protecting elections from social media manipulation. *Science* 365, 6456 (2019), 858–861.

[2] Theo Araujo, Natali Helberger, Sanne Kruikemeier, and Claes H De Vreese. 2020. In AI we trust? Perceptions about automated decision-making by artificial intelligence. *AI & SOCIETY* 35, 3 (2020), 611–623.

[3] 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

[4] 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

[5] Kenneth C. Arnold, Krzysztof Z. Gajos, and Adam T. Kalai. 2016. On Suggesting Phrases vs. Predicting Words for Mobile Text Composition. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*. ACM, Tokyo Japan, 603–608. https://doi.org/10.1145/2984511.2984584

[6] Solomon E Asch. 1951. Effects of group pressure upon the modification and distortion of judgments. *Organizational influence processes* 58 (1951), 295–303.

[7] Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, et al. 2021. A general language assistant as a laboratory for alignment. *arXiv preprint arXiv:2112.00861* (2021).

[8] Eugene Bagdasaryan and Vitaly Shmatikov. 2021. Spinning Sequence-to-Sequence Models with Meta-Backdoors. *arXiv preprint arXiv:2107.10443* (2021).

[9] Christopher A Bail, Lisa P Argyle, Taylor W Brown, John P Bumpus, Haohan Chen, MB Fallin Hunzaker, Jaemin Lee, Marcus Mann, Friedolin Merhout, and Alexander Volfovsky. 2018. Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences* 115, 37 (2018), 9216–9221.

[10] Eytan Bakshy, Jake M Hofman, Winter A Mason, and Duncan J Watts. 2011. Everyone's an influencer: quantifying influence on twitter. In *Proceedings of the fourth ACM international conference on Web search and data mining*. 65–74.

[11] Nikola Banovic, Ticha Sethapakdi, Yasasvi Hari, Anind K. Dey, and Jennifer Mankoff. 2019. The Limits of Expert Text Entry Speed on Mobile Keyboards with Autocorrect. In *Proceedings of the 21st International Conference on Human-Computer Interaction with Mobile Devices and Services* (Taipei, Taiwan) *(MobileHCI '19)*. Association for Computing Machinery, New York, NY, USA, Article 15, 12 pages. https://doi.org/10.1145/3338286.3340126

[12] Carolyn Black Becker, Lisa M Smith, and Anna C Ciao. 2006. Peer-facilitated eating disorder prevention: A randomized effectiveness trial of cognitive dissonance and media advocacy. *Journal of Counseling Psychology* 53, 4 (2006), 550.

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

[14] 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.

[15] Shlomo Berkovsky, Jill Freyne, and Harri Oinas-Kukkonen. 2012. Influencing individually: fusing personalization and persuasion. , 8 pages.

[16] Advait Bhat, Saaket Agashe, and Anirudha Joshi. 2021. How do people interact with biased text prediction models while writing?. In *Proceedings of the First Workshop on Bridging Human–Computer Interaction and Natural Language Processing*. Association for Computational Linguistics, Online, 116–121. https://aclanthology.org/2021.hcinlp-1.18

[17] Advait Bhat, Saaket Agashe, Niharika Mohile, Parth Oberoi, Ravi Jangir, and Anirudha Joshi. 2022. Interacting with next-phrase suggestions: How suggestion systems aid and influence the cognitive processes of writing. (2022). https://doi.org/10.48550/ARXIV.2208.00636

[18] 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

[19] Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (Technology) is Power: A Critical Survey of Bias in NLP. *arXiv preprint arXiv:2005.14050* (2020).

[20] Eric Bogert, Aaron Schecter, and Richard T Watson. 2021. Humans rely more on algorithms than social influence as a task becomes more difficult. *Scientific reports* 11, 1 (2021), 1–9.

[21] 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).

[22] Samantha Bradshaw and Philip Howard. 2017. Troops, trolls and troublemakers: A global inventory of organized social media manipulation. (2017).

[23] 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.

[24] Axel Bruns. 2019. *Are filter bubbles real?* John Wiley & Sons.

[25] Ben Buchanan, Andrew Lohn, Micah Musser, and Katerina Sedova. 2021. Truth, Lies, and Automation. *Center for Security and Emerging Technology* (2021).

[26] Daniel Buschek, Benjamin Bisinger, and Florian Alt. 2018. ResearchIME: A Mobile Keyboard Application for Studying Free Typing Behaviour in the Wild. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, New York, NY, USA. https://doi.org/10.1145/3173574.3173829 event-place: Montreal, Quebec, CA.

[27] 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

[28] Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science* 356, 6334 (Apr 2017), 183–186. https://doi.org/10.1126/science.aal4230

[29] 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

[30] Nicholas A Christakis and James H Fowler. 2007. The spread of obesity in a large social network over 32 years. *New England journal of medicine* 357, 4 (2007), 370–379.

[31] Nicholas A Christakis and James H Fowler. 2008. The collective dynamics of smoking in a large social network. *New England journal of medicine* 358, 21 (2008), 2249–2258.

[32] Matteo Cinelli, Gianmarco De Francisci Morales, Alessandro Galeazzi, Walter Quattrociocchi, and Michele Starnini. 2021. The echo chamber effect on social media. *Proceedings of the National Academy of Sciences* 118, 9 (2021), e2023301118.

[33] 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

[34] Morning Consult. 2016. National Tracking Poll #2110047.

[35] Dan Cosley, Shyong K Lam, Istvan Albert, Joseph A Konstan, and John Riedl. 2003. Is seeing believing? How recommender system interfaces affect users' opinions. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 585–592.

[36] Wenzhe Cui, Suwen Zhu, Mingrui Ray Zhang, H. Andrew Schwartz, Jacob O. Wobbrock, and Xiaojun Bi. 2020. *JustCorrect: Intelligent Post Hoc Text Correction Techniques on Smartphones*. Association for Computing Machinery, New York, NY, USA, 487–499. https://doi.org/10.1145/3379337.3415857

[37] Girish Dalvi, Shashank Ahire, Nagraj Emmadi, Manjiri Joshi, Anirudha Joshi, Sanjay Ghosh, Prasad Ghone, and Narendra Parmar. 2016. Does Prediction Really Help in Marathi Text Input? Empirical Analysis of a Longitudinal Study. In *Proceedings of the 18th International Conference on Human-Computer Interaction with Mobile Devices and Services* (Florence, Italy) *(MobileHCI '16)*. Association for Computing Machinery, New York, NY, USA, 35–46. https://doi.org/10.1145/2935334.2935366

[38] Hai Dang, Karim Benharrak, Florian Lehmann, and Daniel Buschek. 2022. Beyond Text Generation: Supporting Writers with Continuous Automatic Text Summaries. *arXiv preprint arXiv:2208.09323* (2022).

[39] Sebastian Duerr and Peter A Gloor. 2021. Persuasive Natural Language Generation–A Literature Review. *arXiv preprint arXiv:2101.05786* (2021).

[40] 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

[41] Emilio Ferrara, Onur Varol, Clayton Davis, Filippo Menczer, and Alessandro Flammini. 2016. The rise of social bots. *Commun. ACM* 59, 7 (2016), 96–104.

[42] Brian J Fogg. 2002. Persuasive technology: using computers to change what we think and do. *Ubiquity* 2002, December (2002), 2.

[43] 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

[44] Ismael Garrido-Muñoz, Arturo Montejo-Ráez, Fernando Martínez-Santiago, and L Alfonso Ureña-López. 2021. A survey on bias in deep NLP. *Applied Sciences* 11, 7 (2021), 3184.

[45] Federico Gaspari, Antonio Toral, Sudip Kumar Naskar, Declan Groves, and Andy Way. 2014. Perception vs. reality: measuring machine translation post-editing productivity. In *Proceedings of the 11th Conference of the Association for Machine Translation in the Americas*. 60–72.

[46] 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).

[47] Tarleton Gillespie. 2014. The relevance of algorithms. *Media technologies: Essays on communication, materiality, and society* 167, 2014 (2014), 167.

[48] Sharad Goel, Duncan J Watts, and Daniel G Goldstein. 2012. The structure of online diffusion networks. In *Proceedings of the 13th ACM conference on electronic commerce*. 623–638.

[49] 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

[50] Junius Gunaratne, Lior Zalmanson, and Oded Nov. 2018. The persuasive power of algorithmic and crowdsourced advice. *Journal of Management Information Systems* 35, 4 (2018), 1092–1120.

[51] 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

[52] Jess Hohenstein and Malte Jung. 2020. AI as a moral crumple zone: The effects of AI-mediated communication on attribution and trust. *Computers in Human Behavior* 106 (2020), 106190. https://doi.org/10.1016/j.chb.2019.106190

[53] Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146* (2018).

[54] 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).

[55] Maurice Jakesch, Zana Buçinca, Saleema Amershi, and Alexandra Olteanu. 2022. How Different Groups Prioritize Ethical Values for Responsible AI. *arXiv preprint arXiv:2205.07722* (2022).

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

[57] 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).

[58] Daniel Kahneman. 2011. *Thinking, fast and slow*. Macmillan.

[59] 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

[60] Shivani Kapania, Oliver Siy, Gabe Clapper, Azhagu Meena SP, and Nithya Sambasivan. 2022. " Because AI is 100% right and safe": User Attitudes and Sources of AI Authority in India. In *CHI Conference on Human Factors in Computing Systems*. 1–18.

[61] Elise Karinshak, Sunny Liu, Joon Sung Park, and Jeff Hancock. 2022. Can AI Persuade? Examining A Large Language Model's Ability to Generate Pro-Vaccination Messages. *International Communication Association Annual Conference* (2022).

[62] Habibul Haque Khondker. 2011. Role of the new media in the Arab Spring. *Globalizations* 8, 5 (2011), 675–679.

[63] Joachim Kimmerle, Johannes Moskaliuk, Martina Bientzle, Ansgar Thiel, and Ulrike Cress. 2012. Using Controversies for Knowledge Construction: Thinking and Writing about Alternative Medicine. In *ICLS*.

[64] Thomas R Knapp. 1990. Treating ordinal scales as interval scales: an attempt to resolve the controversy. *Nursing research* 39, 2 (1990), 121–123.

[65] Johannes Knoll. 2016. Advertising in social media: a review of empirical evidence. *International journal of Advertising* 35, 2 (2016), 266–300.

[66] Svetlana Koltovskaia. 2020. Student engagement with automated written corrective feedback (AWCF) provided by Grammarly: A multiple case study. *Assessing Writing* 44 (2020), 100450.

[67] 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.

[68] Per Ola Kristensson and Keith Vertanen. 2014. The inviscid text entry rate and its application as a grand goal for mobile text entry. In *Proceedings of the 16th international conference on Human-computer interaction with mobile devices & services (MobileHCI '14)*. Association for Computing Machinery, Toronto, ON, Canada, 335–338. https://doi.org/10.1145/2628363.2628405

[69] Paul F Lazarsfeld, Bernard Berelson, and Hazel Gaudet. 1968. *The people's choice*. Columbia University Press.

[70] Mina Lee, Percy Liang, and Qian Yang. 2022. CoAuthor: Designing a Human-AI Collaborative Writing Dataset for Exploring Language Model Capabilities. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 388, 19 pages. https://doi.org/10.1145/3491102.3502030

[71] Florian Lehmann, Niklas Markert, Hai Dang, and Daniel Buschek. 2022. Suggestion Lists vs. Continuous Generation: Interaction Design for Writing with Generative Models on Mobile Devices Affect Text Length, Wording and Perceived Authorship. In *Mensch Und Computer 2022* (Darmstadt, Germany) *(MuC '22)*. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3543758.3543947

[72] Thomas C Leonard. 2008. Richard H. Thaler, Cass R. Sunstein, Nudge: Improving decisions about health, wealth, and happiness.

[73] Brian Lester and Noah Constant. 2022. Guiding frozen language models with learned soft prompts. https://ai.googleblog.com/2022/02/guiding-frozen-language-models-with.html

[74] Yotam Liel and Lior Zalmanson. 2020. What If an AI Told You That 2+ 2 Is 5? Conformity to Algorithmic Recommendations.. In *ICIS*.

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

[76] Jennifer M Logg, Julia A Minson, and Don A Moore. 2019. Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes* 151 (2019), 90–103.

[77] Kevin Macnish and Jeroen van der Ham. 2020. Ethics in cybersecurity research and practice. *Technology in society* 63 (2020), 101382.

[78] Alice E Marwick and Danah Boyd. 2011. I tweet honestly, I tweet passionately: Twitter users, context collapse, and the imagined audience. *New media & society* 13, 1 (2011), 114–133.

[79] Alexey N Medvedev, Renaud Lambiotte, and Jean-Charles Delvenne. 2017. The anatomy of Reddit: An overview of academic research. *Dynamics on and of Complex Networks* (2017), 183–204.

[80] Stanley Milgram. 1963. Behavioral study of obedience. *The Journal of abnormal and social psychology* 67, 4 (1963), 371.

[81] D.G. Myers. 2008. *Social Psychology.* McGraw-Hill.

[82] Clifford Nass, Jonathan Steuer, and Ellen R Tauber. 1994. Computers are social actors. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 72–78.

[83] 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.

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

[85] Kseniia Palin, Anna Maria Feit, Sunjun Kim, Per Ola Kristensson, and Antti Oulasvirta. 2019. How do People Type on Mobile Devices? Observations from a Study with 37,000 Volunteers. In *Proceedings of the 21st International Conference on Human-Computer Interaction with Mobile Devices and Services (MobileHCI '19)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3338286.3340120

[86] Raja Parasuraman and Dietrich H Manzey. 2010. Complacency and bias in human use of automation: An attentional integration. *Human factors* 52, 3 (2010), 381–410.

[87] Raja Parasuraman and Victor Riley. 1997. Humans and automation: Use, misuse, disuse, abuse. *Human factors* 39, 2 (1997), 230–253.

[88] Richard E Petty and John T Cacioppo. 1986. The elaboration likelihood model of persuasion. In *Communication and persuasion*. Springer, 1–24.

[89] Philip Quinn and Shumin Zhai. 2016. A Cost-Benefit Study of Text Entry Suggestion Interaction. 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, 83–88. https://doi.org/10.1145/2858036.2858305

[90] 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.

[91] 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).

[92] Lisa Rashotte. 2007. Social influence. *The Blackwell encyclopedia of sociology* (2007).

[93] Robert J Shiller. 2015. *Irrational exuberance.* Princeton university press.

[94] Herbert W Simons. 2011. *Persuasion in society.* Routledge.

[95] Nikhil Singh, Guillermo Bernal, Daria Savchenko, and Elena L. Glassman. 2022. Where to Hide a Stolen Elephant: Leaps in Creative Writing with Multimodal Machine Intelligence. *ACM Transactions on Computer-Human Interaction* (Feb. 2022), 3511599. https://doi.org/10.1145/3511599

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

[97] 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).

[98] Keith Vertanen, Haythem Memmi, Justin Emge, Shyam Reyal, and Per Ola Kristensson. 2015. VelociTap: Investigating Fast Mobile Text Entry using Sentence-Based Decoding of Touchscreen Keyboard Input. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. Association for Computing Machinery, Seoul, Republic of Korea, 659–668. https://doi.org/10.1145/2702123.2702135

[99] Chin-Sheng Wan and Wen-Bin Chiou. 2010. Inducing attitude change toward online gaming among adolescent players based on dissonance theory: The role of threats and justification of effort. *Computers & Education* 54, 1 (2010), 162–168. https://doi.org/10.1016/j.compedu.2009.07.016

[100] Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, et al. 2021. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359* (2021).

[101] 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. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. 214–229.

[102] Christopher D Wickens, Benjamin A Clegg, Alex Z Vieane, and Angelia L Sebok. 2015. Complacency and automation bias in the use of imperfect automation. *Human factors* 57, 5 (2015), 728–739.

[103] 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).

[104] Daijin Yang, Yanpeng Zhou, Zhiyuan Zhang, Toby Jia-Jun Li, and LC Ray. 2022. AI as an Active Writer: Interaction Strategies with Generated Text in Human-AI Collaborative Fiction Writing 56-65. In *IUI Workshops*.

[105] Ann Yuan, Andy Coenen, Emily Reif, and Daphne Ippolito. 2022. Wordcraft: Story Writing With Large Language Models. In *27th International Conference on Intelligent User Interfaces (IUI '22)*. Association for Computing Machinery, New York, NY, USA, 841–852. https://doi.org/10.1145/3490099.3511105

[106] 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).

[107] Mingrui Ray Zhang, He Wen, and Jacob O. Wobbrock. 2019. Type, Then Correct: Intelligent Text Correction Techniques for Mobile Text Entry Using Neural Networks. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology* (New Orleans, LA, USA) *(UIST '19)*. Association for Computing Machinery, New York, NY, USA, 843–855. https://doi.org/10.1145/3332165.3347924

[108] Ekaterina Zhuravskaya, Maria Petrova, and Ruben Enikolopov. 2020. Political effects of the internet and social media. *Annual Review of Economics* 12 (2020), 415–438.