COMMUNICATION

Biased AI writing assistants shift users' attitudes on societal issues

Sterling Williams-Ceci[1,2]*, Maurice Jakesch[1,3]*, Advait Bhat[4], Kowe Kadoma[1,2], Lior Zalmanson[5], Mor Naaman[1,2]

Artificial intelligence (AI) writing assistants powered by large language models (LLMs) are increasingly used to make autocomplete suggestions to people as they write text. Can these AI writing assistants affect people's attitudes in this process? In two large-scale preregistered experiments ($N = 2582$), we exposed participants writing about important societal issues to an AI writing assistant that provided biased autocomplete suggestions. When using the AI assistant, the attitudes participants expressed in a posttask survey converged toward the AI's position. However, a majority of participants were unaware of the AI suggestions' bias and their influence. Further, the influence of the AI writing assistant was stronger than the influence of similar suggestions presented as static text, showing that the influence is not fully explained by these suggestions, increasing accessibility of the biased information. Last, warning participants about assistants' bias before or after exposure does not mitigate the attitude-shift effect.

INTRODUCTION

Artificial intelligence (AI) technologies based on large language models (LLMs) are increasingly involved in human communication, generating text and making suggestions in a multitude of contexts. First appearing as "smart replies" (*1*) for email and messaging, AI-mediated communication ["AIMC"; (*2*)] features are now ubiquitous in text editors like Microsoft Word, email clients, messaging applications, and elsewhere (*3*–*7*). The integration of generative AI into people's communication raises questions about the technology's impact on language use (*2*, *8*–*11*), perceptions of others (*6*, *12*, *13*), and interpersonal dynamics (*3*, *13*–*15*). Together, these studies raise the broader question of how much AI can influence people's cognition (*16*). As AI tools have been shown to reproduce and reinforce human biases (*16*–*20*), their influence on people's cognition has become a serious concern (*16*).

Previous research has examined AI's persuasive capabilities in explicit contexts, where people are aware of an overt influence attempt. Studies have shown the persuasive impact of AI in various types of interaction with humans, such as when AI generates persuasive messages (*21*–*26*), converses with people about a relevant belief (*27*–*30*), and provides decision recommendations (*31*–*33*). However, a critical gap remains in our understanding of the more subtle and indirect ways that AI can influence human cognition.

The present study investigates a potentially more powerful and pervasive form of AI influence: the impact of AI writing assistants that provide autocomplete suggestions ("AI suggestions" for short) on users' attitudes ["attitudes" refers to the evaluations people form toward "objects" such as societal issues and policy decisions; see (*34*)]. Unlike the explicit persuasion attempts studied previously, AI suggestions may shape users' cognition in ways that bypass conscious awareness, representing a more indirect and covert way in which AI can influence humans. In applications that make use of AI suggestions, writing assistants powered by LLMs predict in-line continuations to the author's writing and suggest this predicted text to the author (*35*–*38*). AI writing assistants that provide autocomplete suggestions pose a distinct pathway for influencing users' thoughts: While it is often known that people's attitudes influence their behavior, it is also known that people's behaviors can shift their attitudes (*39*–*41*). When biased AI autocomplete suggestions inspire people to think of certain viewpoints—and make it easier for people to elaborate on those viewpoints than on others—cowriting with a biased AI writing assistant may powerfully shift people's attitudes on various issues. If AI suggestions have even a minor influence at the individual level, their potential for aggregate influence is worrisome as millions of people are using the same models in their daily communications (*18*, *42*, *43*).

Emerging work has identified how AI writing assistants can be biased and how they can influence users' communication behavior and perceptions. LLMs have been found to produce biased viewpoints, perhaps due to being trained on corpuses that overrepresent text written by people with these viewpoints (*17*). Although the user's own writing may steer the stance of AI suggestions to an extent (*44*), biases can also be deliberately introduced into AI suggestions in adversarial scenarios with people providing biased training data, prompts, or feedback during the training process when creating these applications (*42*, *45*–*47*). As for the outcomes of these biased suggestions, work has shown that AI assistants' suggestions can influence users to write content that is biased in the same direction as the suggestions (*8*, *10*) and has highlighted the potential for AI suggestions to affect cognitive outcomes, such as users' feelings of agency, ownership, and/or inclusion while writing (*15*, *48*–*50*). However, none of these studies has looked at the covert influence that AI writing assistants might have on the attitudes people hold toward the topics they write about.

In this work, we provide robust evidence of the impact of AI writing assistants' autocomplete suggestions on attitudes and show that well-known influence mitigation techniques may not reduce it. Our work shows that AI suggestions can influence people's attitudes across a range of politically charged societal issues, with AI biases that vary in partisan association and stance. We compare the covert

[1]Department of Information Science, Cornell University, Ithaca, NY, USA. [2]Cornell Tech, New York, NY, USA. [3]Department of Computer Science, Bauhaus University, Weimar, Germany. [4]Paul G. Allen School of Computer Science and Engineering, University of Washington, Seattle, WA, USA. [5]Coller School of Management, Tel Aviv University, Tel Aviv, Israel.
*Corresponding author. Email: scw222@cornell.edu (S.W.-C.); mpj32@cornell.edu (M.J.)

influence of people writing with the AI writing assistant's autocomplete suggestions to the influence when overtly presenting similar arguments and show that the autocomplete suggestions have a significantly stronger effect on attitudes. We also find that users are generally unaware of the suggestions' influence. Furthermore, we show that the effect is not eliminated or even reduced by interventions that warn people about the AI suggestions' bias, despite previous work suggesting that these techniques can counteract the influence of persuasive messages (*51*). An initial study suggested that AI writing assistants may shift users' attitudes toward the issues they write about (*8*), but the study's design was too limited to support generalizable conclusions about this phenomenon: It only relied on a single experiment using one issue that was not politically charged; it did not rule out other potential explanations; and it did not test whether the influence can be undone by making people more aware of the bias [which has been found in the case of misinformation (*51*)].

In our experiments, participants wrote a short essay about one of five debatable issues with (AI Treatment) or without (Control) receiving AI suggestions that were biased toward one viewpoint (see Fig. 1 for an example of the AI Treatment's suggestions). In experiment 1 ($N = 1485$), all participants wrote about their attitudes toward whether standardized tests should be used in education. In the AI Treatment, an AI writing assistant provided autocomplete suggestions that made arguments biased in favor of standardized testing, which we achieved by carefully engineering an AI prompt that instructed the AI model we used to generate pro-test arguments that still logically completed the users' sentences. In addition to the Control condition ($N = 499$) and AI Treatment ($N = 487$), experiment 1 included a third comparison condition, the Static Text condition ($N = 499$), in which we showed participants a static list of arguments, based on those that were generated by the LLM when providing autocomplete suggestions in a pretest (see fig. S1). This condition explored whether the content of the suggestions alone could explain any observed shift in attitude, or whether the unique ability of AI writing assistants to change people's writing in real time contributes to AI's influence, as predicted by the literature on behavior influencing attitudes (*39*–*41*). After the writing task, we asked participants about their attitude toward the issue they wrote on (using a Likert scale to measure their attitudes numerically). We then asked participants in the AI Treatment to rate how balanced and influential they perceived the AI suggestions to be.

We used the same basic procedure in experiment 2 ($N = 1097$), but introduced more discussion issues and experimental conditions to expand and increase the robustness of experiment 1's findings regarding the influence of the AI suggestions on participants' attitudes. First, we randomly assigned participants to write about their attitudes on one of four different societal issues. The four issues were selected to be socially and politically consequential: whether the death penalty should be illegal, whether felons should be allowed to vote, whether the US should continue fracking, and whether genetically modified organisms (GMOs) should be grown. Like we did in experiment 1, we configured the AI suggestions to gravitate toward a predetermined biased opinion (the "AI's position," hereafter) for

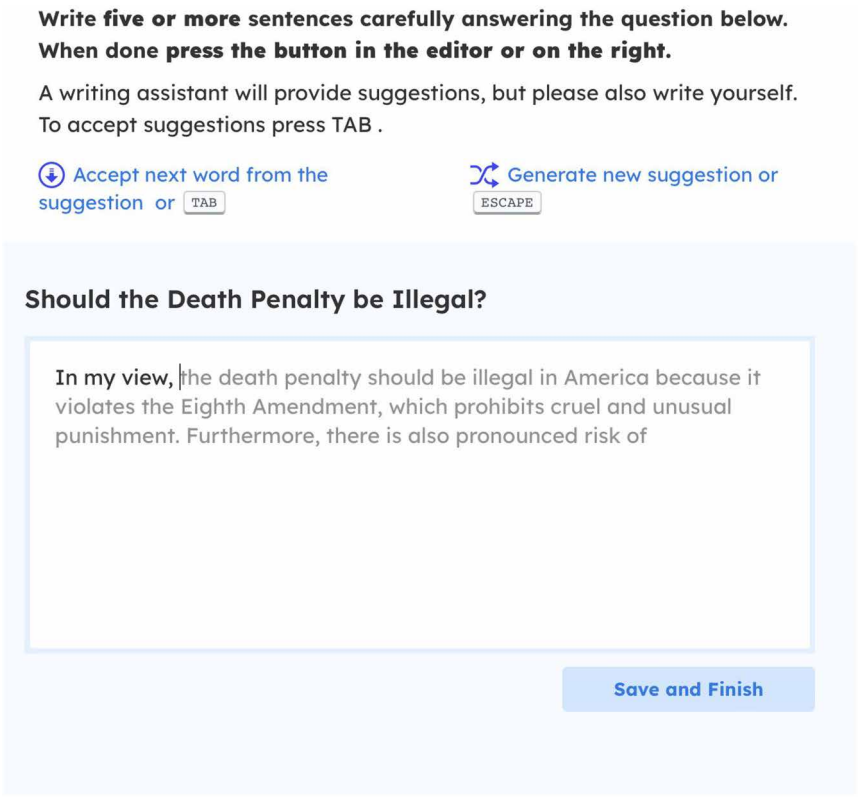

Fig. 1. A screenshot of the writing interface used in the experiment. The figure is showing the experiment 2 treatment group for participants who were asked to write about their attitudes toward the death penalty issue. The AI was configured to advocate support for making the death penalty illegal. Participants in the control groups wrote their responses using the same interface but were not shown the AI suggestions or the related instructions.

each of the issues. A typical suggestion from the AI assistant about the death penalty issue is shown in Fig. 1; for more examples, see Supplementary Text. Across issues, we varied the AI's partisan position (liberal or conservative) and the stance of the suggestions in respect to the issue (positive or negative; see Table 1 for the list of issues and AI writing assistants' positions).

While experiment 1 only compared participants' posttask attitudes between the groups, experiment 2 used a repeated-measures design in which we also measured subjects' pretask attitudes several weeks before participation. We first recruited participants to fill out questionnaires asking for their attitudes toward several societal issues, four of which were the same as those we used in the actual experiment. We used the same Likert scales to measure their attitudes in both stages. We launched the actual experiment 20 days after the last participant had completed the pretask survey and restricted eligibility only to the participants who had completed the pretask survey. The connection between the two tasks was hidden from participants, who took the actual experiment, on average, 34.44 days following the pretask survey; see Materials and Methods for more details. This design enabled us to do a within-subject comparison of the extent of the shift from the presurvey attitude to the postexperiment attitude about the issue they had written about in the experiment. The percentages of participants in experiment 2 whose pretask attitudes agreed with the AI suggestions' bias, disagreed with it, or were neutral for their assigned issue in the ultimate task are shown in table S1.

Last, in addition to the Control (N = 312) and AI Treatment (N = 219), we provided in experiment 2 three additional variants of the AI Treatment designed to evaluate the strength of the attitude-shift effect. Two conditions (see Table 2) tested whether awareness-raising interventions would reduce the AI writing assistant's impact on attitudes. Motivated by previous work on misinformation prevention ["debunking" and "prebunking"; (*51*)], we warned participants of the AI suggestions' bias before writing (Warning condition, N = 200) or after writing (Debrief condition, N = 159). In a final condition we call the Write-in-Favor condition (N = 207), participants were asked to write their essay supporting a predetermined position that aligned with the same bias we asked AI to produce in the other conditions. The participants in this condition used the same biased AI suggestions. This condition mimicked previous work showing that directly manipulating behavior can influence attitudes (*40*, *52*) and allowed us to compare the influence of people's

Table. 1. The issues used and the bias we configured the AI to represent in its autocomplete suggestions. The AI suggestions' stance toward the issue and the partisanship direction are balanced across the four issues used in experiment 2.

Experiment	Issue and AI suggestion' bias	AI suggestions' partisanship toward issue
Experiment 1	In favor of standardized tests	No data on partisanship
Experiment 2	Against the death penalty	Liberal
Experiment 2	Against voting rights for felons	Conservative
Experiment 2	In favor of growing GMOs	Liberal
Experiment 2	In favor of fracking	Conservative

Table. 2. The between-subjects conditions in each experiment. Both experiments had a Control condition and a basic AI Treatment condition, which differed in whether any biased AI suggestions were shown. Experiment 1 included a third condition that showed biased suggestions without using AI. Experiment 2 included three variations of the AI Treatment that involved warning, debriefing, or directly manipulating participants' writing.

Experiment	Condition	Description
Experiment 1 and experiment 2	Control (N = 500; 499 after exclusion in experiment 1; N = 314; 312 after exclusion in experiment 2)	Participants wrote about their opinions on the issue with no assistance
Experiment 1 and experiment 2	AI Treatment (N = 495; 487 after exclusion in experiment 1; N = 318; 219 after exclusion in experiment 2)	Participants wrote about their opinions on the issue while receiving AI-generated suggestions following the biased position
Experiment 1	Static Text (N = 501; 499 after exclusion)	Participants wrote about their opinions on the issue; a bullet-point list of the main arguments generated by AI was shown within the task instructions
Experiment 2	Warning (N = 312; 200 after exclusion)	Participants wrote about their opinions on the issue; the instructions beforehand included the phrase "Note that the AI-generated suggestions might support a different view than your own. When this happens, it may bias your writing."
Experiment 2	Debrief (N = 308; 159 after exclusion)	Participants wrote about their opinions on the issue; afterward, they were shown a statement to continue the survey with the phrase "Note that the AI-generated suggestions might have supported a different view than your own. As a result, they may have biased your writing and influenced your thinking."
Experiment 2	Write-in-Favor (N = 276; 207 after exclusion)	Participants were instructed to write in favor of a specific position on the issue, which was the same one used for the AI suggestions' bias.

natural interaction with AI suggestions to the influence of the strongest effect theoretically possible that directly manipulates people's behavior by giving them an overt directive. Figure 2 depicts the basic design shared by the experiments, along with the distinctions specific to each.

RESULTS

Our results provide robust evidence that biased AI autocomplete suggestions can shift people's attitudes. In addition, we find that users, including those whose attitudes were shifted, were largely unaware of the suggestions' bias and influence. Further, we demonstrate that this result cannot be explained by the mere provision of persuasive information. Last, we show that the effect is not mitigated by raising awareness of the task or of the AI suggestions' potential bias.

Participants' attitudes shifted in the direction of the AI biased suggestions

Across experiments and topics, participants who saw the biased AI suggestions reported attitudes that were significantly closer to the AI's position than participants who wrote without any suggestions. In experiment 1 ($N = 1485$), participants' attitudes differed based on their condition [analysis of variance (ANOVA): $F(2, 1474) = 10.244$, $P < 0.001$, $\eta_p^2 = 0.01$]. Specifically, a post hoc comparison showed that seeing biased AI suggestions ($N = 487$) increased the alignment of participants' attitudes with the AI's position by 0.44 scale point on a scale from 1 to 5 [95% confidence interval (CI) for AI Treatment–Control difference (0.21, 0.67), $t(1474) = 4.523$, $P < 0.001$, Cohen's $d = 0.24$] compared to the attitudes of Control group participants ($N = 499$). This effect includes participants who did not actually accept any of the AI suggestions in their writing. In experiment 2 ($N = 1097$), a similar effect occurred, with participants in the normal AI Treatment ($N = 219$) expressing attitudes in the posttask survey that were 0.41 scale point more aligned with the AI's biased position than Control participants' attitudes ($N = 312$) for the issue they wrote about [ANOVA: $F(4, 1089) = 4.07$, $P < 0.001$, $\eta_p^2 = 0.01$; 95% CI for AI Treatment–Control difference (0.05, 0.77), $P = 0.015$, $t(1089) = 3.14$, Cohen's $d = 0.19$]. These results are summarized in Fig. 3 (the "Control" and "AI Treatment" conditions) alongside results from other conditions discussed below. These findings were confirmed in a series of robustness checks with alternative models (see Supplementary Text).

Moreover, using the pretask measurement of participants' attitudes in experiment 2 in a linear mixed effects model, we found differences in the extent to which participants' attitudes shifted over time between the conditions [condition-by-time interaction $F(4, 1092) = 3.70$, $P = 0.005$, $\eta_p^2 = 0.01$]. In the AI Treatment condition, participants shifted in their views from their baseline positions toward the AI's position by $\Delta_{\text{post-pre}} = 0.37$ scale point on average [95% CI for post-pre in AI Treatment (0.21, 0.53), $t(1092) = 4.50$, $P < 0.001$, Cohen's $d = 0.27$]. Participants in the Control group, who were not exposed to the AI's suggestions, did not show any significant shift in their attitudes [$\Delta_{\text{post-pre}} = 0.06$, 95% CI (−0.07, 0.20), $t(1092) = 0.88$, $P = 0.377$, Cohen's $d = 0.05$]. Furthermore, post hoc comparisons between each condition's pre-post attitude differences additionally showed a significant difference in the magnitude of these shifts between the Treatment and Control groups, even when using Bonferroni as the most conservative adjustment for multiple tests [$\Delta_{\text{difference in post-pre differences}} = 0.31$, 95% CI (0.01, 0.61), $t(1092) = 2.88$, $P = 0.041$, Cohen's $d = 0.25$]. These main results are summarized in Fig. 3, which includes results from other conditions discussed below. In both experiments, the treatment effect increased with the number of words participants accepted from the AI suggestions (see Supplementary Text), further suggesting that this phenomenon reflects the theoretical paradigm of behavior influencing attitudes.

Participants were generally unaware of the AI's bias and influence

Our results reveal that people, including those who were influenced by the AI suggestions, were largely unaware of the suggestions' bias. Most participants exposed to the biased AI suggestions in the AI Treatment across both experiments agreed that the AI suggestions "were reasonable and balanced," indicating a lack of awareness of the suggestions' bias. As shown in Fig. 4, only 19% of participants in experiment 1 and 14% of participants in experiment 2 disagreed with the statement that the suggestions were balanced.

Using the pretask attitudes collected in experiment 2, we specifically compared the awareness of AI suggestions' bias among the treated participants whose attitudes shifted in the AI's direction with those whose attitudes did not shift in the AI's direction. Participants in the AI Treatment whose attitudes shifted toward the AI's position were significantly less likely to disagree with the statement that the suggestions were reasonable and balanced (i.e., less likely to be aware of the suggestions' bias) than all other participants in the AI Treatment group [$\chi^2(2, N = 219) = 9.56$, $P = 0.008$, $\nu = 0.21$]. In fact, 73% of treated participants whose attitude shifted toward the AI agreed that the AI suggestions were balanced and only 4% disagreed with this statement; in contrast, these values were 63 and 19%, respectively, for treated participants whose attitudes did not shift toward the AI's bias [odds ratio = 0.18, 95% CI (0.04, 0.52)].

We also found that participants were often unaware of the AI suggestions' influence on their thinking and argument. As shown in Fig. 4, 53% disagreed and 31% agreed that the assistant influenced their thinking and argument in experiment 1; in experiment 2, 51% disagreed and 34% agreed with this statement. This time, we did not observe a difference in awareness of the suggestions' influence on thinking based on whether participants' attitudes were indeed shifted by the AI writing assistant [$\chi^2(2, N = 219) = 1.29$, $P = 0.524$, $\nu = 0.08$]: Participants who were truly influenced by the AI suggestions were no more likely to agree that they had been influenced than participants who were not influenced [odds ratio: 1.20, 95% CI (0.66, 2.15), $P = 0.272$]. Together, the responses to these questions suggest that treated participants, especially those whose attitudes shifted toward the AI's position, were largely unaware of the AI suggestions' bias and did not recognize its influence on their responses and attitudes.

The influence of AI suggestions is not fully explained by the increased accessibility of otherwise unconsidered information

One theoretical explanation for the observed attitude shift is that biased AI suggestions provide information that is convincing or useful to participants on its own, irrespective of the suggestions changing participants' behavior by influencing their writing. To test this possible explanation, experiment 1 included a third comparison condition, which prominently showed a bulleted list of arguments

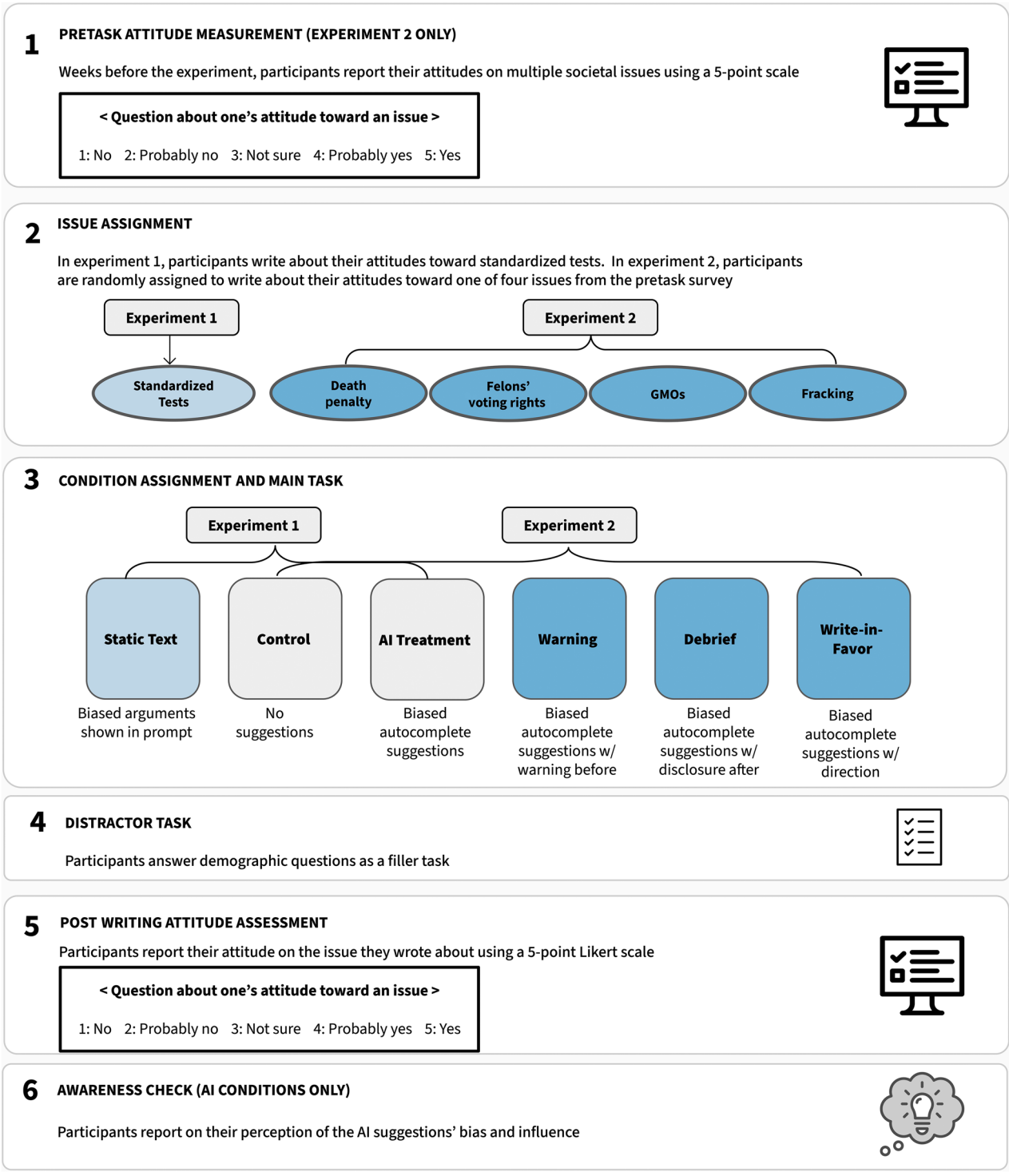

Fig. 2. A visual illustration of the general experiment design and the distinctions between experiment 1 and experiment 2. The numbers on the left-hand side of the figure denote each stage of the experiment. The computer icon in the stage 1 box shows that participants for experiment 2 were first recruited weeks prior to the experiment and completed an attitude survey on multiple issues. The ovals in the stage 2 box represent the societal issues participants were assigned to write about. The boxes in the stage 3 box represent the conditions to which participants were assigned. The light blue oval in stage 2 and light blue box in stage 3 shows which issue and condition were unique to experiment 1; the dark blue ovals in stage 2 and dark blue boxes in stage 3 show which issues and conditions were unique to experiment 2. The checklist icon in the stage 4 box represents the demographic survey participants completed after the writing task and before the attitude measure (intended to be a distractor task). The computer icon in the stage 5 box represents the main dependent measure, in which participants reported their attitude toward the specific issue they wrote about (the same as in the pretask attitude survey for experiment 2 participants). The lightbulb in the stage 6 box represents the final survey participants completed about their awareness of the suggestions' bias and influence (only shown to participants in conditions with AI suggestions).

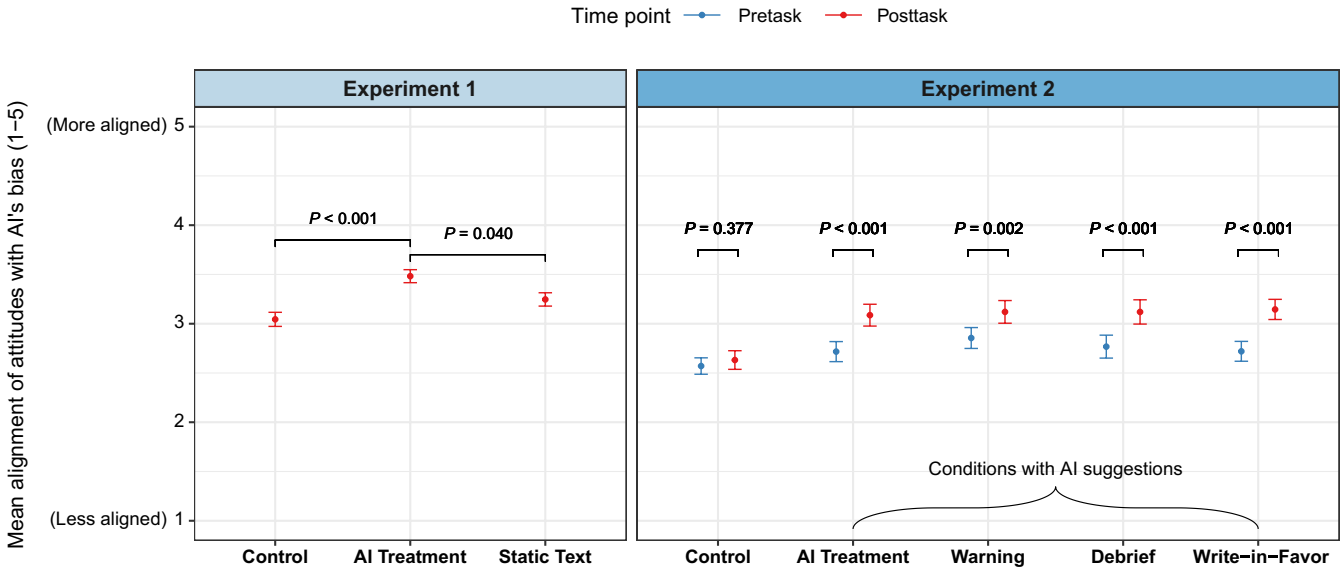

Fig. 3. AI suggestions influenced participants' attitudes while writing about societal issues, the influence was not entirely attributable to the information the suggestions provided, and the influence was not reduced by warning people beforehand or debriefing them afterward. The left panel represents data from experiment 1, and the right panel represents data from experiment 2. The *x* axes compare the conditions in each experiment. The *y* axis of the panels represents the entire range of possible attitude ratings (1 to 5) where increasing values signify increased alignment of the participant's attitude with the AI's bias. The red dots represent participants' posttask attitudes; the blue dots in the right panel represent the participants' pretask attitudes in experiment 2. Statistical test for experiment 1: one-way ANOVA, which showed a main effect of the condition, followed by post hoc contrasts with Tukey adjustment. Experiment 2 statistical test: *F* test on a linear mixed effects model, which showed an interaction between condition and time of attitude measurement, followed by post hoc contrasts within and between conditions with Bonferroni adjustment. Error bars represent standard errors of the means. Each experiment was performed once ($N = 1$). For raw data distributions, see Supplementary Text.

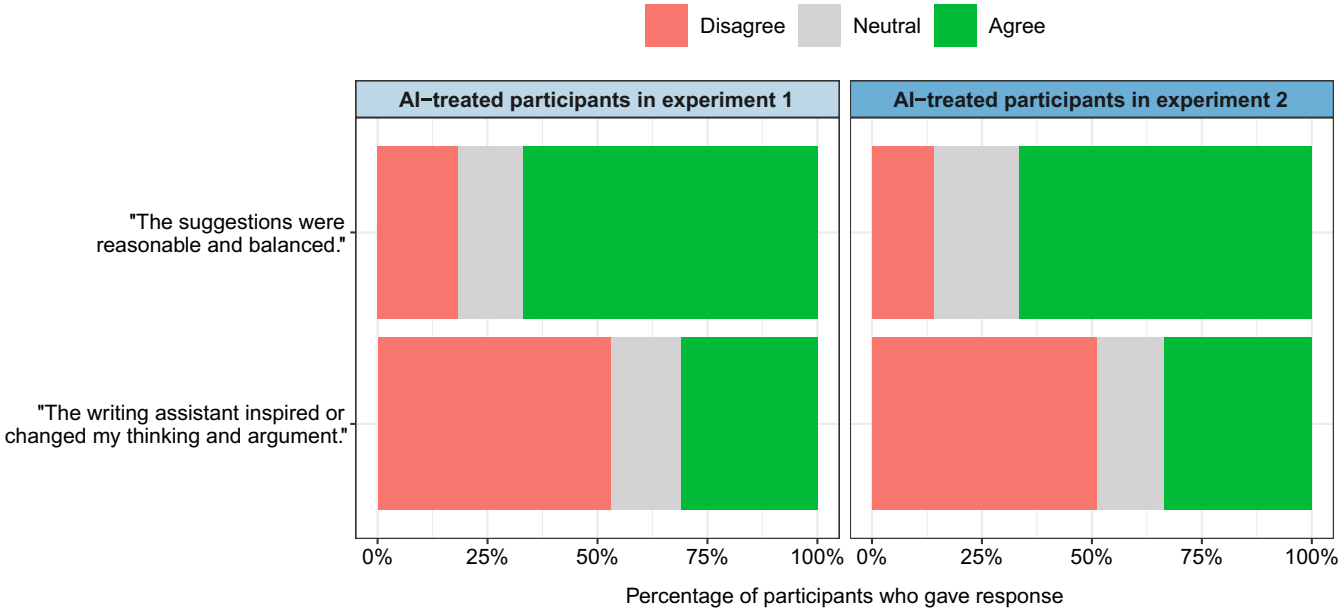

Fig. 4. The majority of participants who were exposed to biased AI suggestions were unaware of the suggestions' bias and influence when asked afterward. The figure shows the distribution of responses to the statements that "the suggestions were reasonable and balanced" and "the writing assistant inspired or changed my thinking and argument" (*y* axis) in the AI Treatment groups in experiment 1 (left) and experiment 2 (right). The *x* axis shows the percentage of participants who indicated each level of agreement. The majority of participants in each experiment indicated agreement with the statement that the suggestions were reasonable and balanced, while also disagreeing that the suggestions influenced their thinking.

biased in the same direction as the AI suggestions, instead of providing the information as autocomplete suggestions (Static Text condition, $N = 499$; see Supplementary Text for details). We created these arguments during a pretest by analyzing a list of arguments generated by the same biased language model that generated the autocomplete suggestions in the AI Treatment group. The Static Text condition thus allowed us to disentangle the influence of the suggestions' biased information from the influence of writing with these suggestions in the autocomplete modality.

As expected, the influence of the AI writing assistant's autocomplete suggestions was stronger than simply providing the biased information as informative suggestions. Participants who received AI suggestions reported attitudes that were 0.24 scale point closer to the AI's position than participants who merely saw the arguments presented as static text [95% CI (0.01, 0.46), $t(1474) = 2.434$, $P = 0.040$, Cohen's $d = 0.13$; see Fig. 3]. Unlike participants who saw the AI suggestions, participants who saw the static text suggestions did not report attitudes that were significantly different than those of participants in the Control after Tukey adjustment for multiple comparisons [difference in means = 0.20, 95% CI (−0.02, 0.43), $t(1474) = 2.098$, $P = 0.091$, Cohen's $d = 0.11$]. This finding shows that AI writing assistants that provide autocomplete suggestions have a unique influence that is stronger than that of AI-generated information, which was the focus of most previous work (*21*–*26*). Further, the finding suggests that the influence of biased AI suggestions on people's attitudes and writing is not simply due to these suggestions providing new information.

The covert influence from the AI suggestions was on par with the expected influence when participants were overtly instructed to change their writing

We also included a third variant of the basic treatment, a Write-in-Favor condition ($N = 207$), in which we explicitly asked participants to express an opinion in their writing (e.g., that the death penalty should be illegal), regardless of their own opinion on the topic. This condition measured the maximum expected strength of the influence of the AI suggestions when directly manipulating people's behavior (their writing) in an overt manner; we used the resulting degree of attitude shift in this condition as a benchmark against which to compare the attitude shift resulting from receiving these AI suggestions without any explicit behavioral manipulation.

As expected, participants' attitudes in the Write-in-Favor condition shifted toward the AI's position by 0.43 scale point [95% CI (0.26, 0.59), Cohen's $d = 0.30$, $P < 0.001$]. Notably, this shift was not statistically different from the shift caused by the default (no-awareness) AI Treatment [difference in shifts = 0.055, 95% CI (−0.39, 0.28), $P = 1.00$, Cohen's $d = 0.05$]. That is, the covert influence resulting from merely being exposed to the biased AI suggestions without any direct behavioral manipulation was on par with the influence observed when directly manipulating people to express these biased viewpoints in their writing. While the attitude-shift effect was similar, the awareness of AI's bias and its impact changed as expected given the overt treatment: Participants in this treatment were more likely to agree with the statement that the suggestions were reasonable and balanced relative to the normal AI Treatment [$\chi^2(2, N = 426) = 23.1$, $P < 0.001$, $\nu = 0.23$, odds ratio for agreement = 2.54, 95% CI (1.61, 4.08)] and were less likely than the default AI Treatment participants to disagree that the suggestions influenced their thinking [$\chi^2(2, N = 426) = 9.5$, $P = 0.009$, $\nu = 0.15$, odds ratio for agreement = 1.37, 95% CI (0.38, 0.82)].

Making participants aware of the AI's bias showed no evidence of mitigating the effect

In experiment 2, we explored two additional conditions designed to increase awareness of the AI's bias in different ways. In both of these conditions, participants saw the biased AI suggestions, but we also made participants aware of the AI writing assistant's bias: In one condition, we warned participants about the AI suggestions' bias before the task (Warning condition, $N = 200$); in another, we alerted them about the bias after the task and before we asked the posttask attitude questions (Debrief condition, $N = 159$).

We found that neither of these treatments significantly reduced the attitude shift compared to the normal AI Treatment baseline. Participants in the Warning and Debrief conditions showed comparable and significant attitude shifts toward the AI's position: Participants who were warned beforehand reported attitudes that were 0.27 scale point closer to the AI's position [95% CI (0.10, 0.43), $P = 0.002$, Cohen's $d = 0.19$], and participants who were debriefed after exposure reported attitudes that were 0.35 scale point closer to the AI's position [95% CI (0.16, 0.54), $P < 0.001$, Cohen's $d = 0.22$]. When comparing the pre-post attitude shifts between conditions, the shifts in these intervention conditions did not statistically differ in magnitude from the shifts participants experienced in the default AI Treatment [difference in shifts between Warning and AI Treatment = 0.105 scale point, 95% CI (−0.23, 0.44), $P = 1.000$, Cohen's $d = 0.08$; difference in shifts between Debrief and AI Treatment = 0.018 scale point, 95% CI (−0.34, 0.37), $P = 1.000$, Cohen's $d = 0.01$].

In both of the attempts at mitigation, participants were just as likely as participants in the normal AI treatment to agree that the AI suggestions were balanced, showing a comparable lack of awareness about the suggestions' biasedness [Warning: $\chi^2(2, N = 419) = 2.17$, $P = 0.339$, $\nu = 0.07$, odds ratio for agreement = 1.35, 95% CI (0.89, 2.06); Debrief: $\chi^2(2, N = 378) = 1.48$, $P = 0.733$, $\nu = 0.06$, odds ratio for agreement = 1.30, 95% CI (0.84, 2.05)]. In addition, participants in these treatments were equally likely to disagree that the AI suggestions influenced their thinking [Warning: $\chi^2(2, N = 419) = 3.33$, $P = 0.312$, $\nu = 0.09$, odds ratio for agreement = 0.72, 95% CI (0.48, 1.10]; Debrief: $\chi^2(2, N = 378) = 1.66$, $P = 0.436$, $\nu = 0.07$, odds ratio for agreement = 0.93, 95% CI (0.60, 1.43)]. Therefore, these interventions did not appear to reduce the attitude shift caused by the biased AI suggestions, nor did they make people more aware of being influenced.

DISCUSSION

Our findings offer robust evidence of one of multiple possible ways in which advanced artificial intelligence in the form of AI writing assistants can, as suggested by scholars (*16*), distort human beliefs and contribute to bias. Alarmingly, our work shows that biased AI models can distort attitudes in a covert and implicit way, without people noticing that they are being presented with a persuasive argument by a different actor. This influence constitutes a form of manipulation, defined by Susser *et al.* as the steering of an individual's decision without their awareness (*53*). Although this influence can sometimes be desirable, for example, when encouraging polarized individuals to consider different points of view, such influence can

also be extremely dangerous, threatening users' freedom of thought when biased AI suggestions influence attitudes in their direction.

The context of our study resembles the way billions of people encounter and interact with AI autocompletions offered in everyday communication contexts, such as on email (*2*, *3*), social media platforms, work platforms like word processors, and even standalone AI-based websites that are being advertised as helping people write for a range of different communication scenarios. The suggestions in our study were longer than current applications' norms and were biased by design. At the same time, as people routinely use these services in their daily communication, the short-term influence of AI suggestions is likely to be compounded by repeated exposure, which could result in a long-term shift in attitudes and, ultimately, damage to people's freedom of thought. These harms could disproportionately affect people of certain groups, such as non-native speakers who rely more on AI language assistants like machine translation to communicate (*54*).

Our findings closely align with previous research on the "behavior influences attitude" paradigm (*52*) but offer an AI-enhanced mechanism for this phenomenon. We found that people who accepted more words from the biased AI suggestions exhibited stronger attitude shifts toward the biased position. Much like our Write-in-Favor condition, earlier studies in this paradigm have shown that the act of overt writing to endorse an opinion can shift attitudes in the same direction. The paradigm offers multiple explanations for this phenomenon, including biased scanning, where people perform selective search of memory or a biased analysis (*39*); cognitive dissonance, where the writing commits individuals to an attitude position to maintain consistency (*40*); and the impact of self-perception, where people treat the writing as information that is relevant to judgments about their own attitudes (*41*). All these factors may contribute to the effect shown in our experiments. Our experiments' design does not allow us to pinpoint which of these is the causal factor, and more than one of these mechanisms (or one we have not listed) could be contributing to the effect simultaneously. Future research should work toward uncovering the root causes of the phenomenon we have documented here. One actionable way to achieve this goal is by replicating our experiments while adapting the manipulations that contributed to each theory in the original studies. For instance, a replication study using our Write-in-Favor condition could add Festinger's (*40*) manipulation of incentive size, which he used to show that cognitive dissonance and the resulting attitude change are induced when the incentive is considered too small to warrant the participant's dissonant outward behavior. By reevaluating the attitude-shift effect under these different conditions, which were used to pinpoint each mechanism in the original studies, future research can lend support to our finding and to these previously known processes.

While our experiments cannot identify the specific theoretical mechanism accounting for the attitude shift, we contribute to the general literature on how behavior influences attitudes by showing that this paradigm works covertly as well, i.e., when people are not aware of any manipulation or directive. This pattern of influence is so subtle that it persists even when participants are forewarned or debriefed to heighten their sensitivity to the manipulation. According to Albarracín and Vargas (*34*), "people may be unaware of the presence of a persuasive stimulus, they may be unaware of cognitive processes that mediate attitude formation and change, and they may be unaware of the outcome of the persuasive process—that their attitudes have actually changed" (p. 14). Our findings support this idea: Participants whose attitudes shifted toward the AI's position often failed to notice the suggestions' biases and influence on their attitudes.

Our experiments were designed to provide robust causal evidence by randomly assigning participants to receive AI suggestions with a specific bias. The design resembles realistic scenarios where users are exposed to built-in AI models' suggestions (e.g., Gmail and Microsoft Word) rather than choosing specific AI systems based on personal preference. Prior research shows that exposures to such built-in suggestions can reduce later information seeking (*16*), making first encounters particularly influential. Moreover, our studies show that users who initially disagreed with the AI's position still accepted substantial portions of its suggestions (see Supplementary Text), suggesting that AI exposure can influence people's writing and attitudes even when they are ideologically opposed to the arguments being presented, similar to findings in other studies (*27*). However, as AI tools become more politicized, users may select assistants that align with their views, leading to potential echo chamber effects. More work is needed to understand how dynamics of self-selection and personalization mediate persuasive effects in more real-world usage.

The Warning and Debrief messages failed to significantly reduce the attitude shift, which is concerning because they were also inspired by those used in real AI applications. AI tools such as ChatGPT show brief and general statements about AI's propensity to hallucinate false information (e.g., "ChatGPT is AI and can make mistakes. Check important info."), similar to the messages used in our interventions (see Table 2). We performed equivalence tests, which showed that, unlike the Debrief, the Warning condition may have reduced the attitude shift to a small extent that our study was underpowered to detect (see the Supplementary Text, pp. 35–36). Although this possibility aligns with inoculation theory, which holds that pre-exposure interventions tend to be more effective than postexposure ones (*51*, *55*), our data allow us to conclude that any plausible inoculation effect would be insufficient to undo the persuasive main effect. One possible reason for the lack of substantial effects is that the intervention text was not sufficiently detailed for participants to internalize its message. Multiple studies in the misinformation context have found that intervention messages with details and examples are more effective (*55*, *56*). In this sense, the null effects of the Warning and Debrief interventions can be seen as a theoretically informative boundary condition in the context of mitigating AI writing assistants' influence: We show that the brief disclaimers used in real-world AI applications are not substantially reducing these tools' influence on users. Future research in this context should evaluate our interventions with larger samples, and it should test more detailed interventions informed by theory to understand how to mitigate AI's influence.

The influence of AI suggestions may be weaker for issues where people hold strong prior attitudes. Attitudes on issues that strongly implicate identity or moral values are generally hard to shift (*57*, *58*). Nevertheless, we showed in this work that for multiple topics of considerable societal impact, AI suggestions, varying in their political leaning, can shift attitudes. For the issues used in our study, an exploratory analysis in Supplementary Text shows that attitudes shifted even for participants whose initial attitudes were extreme, and their attitudes shifted significantly more than the natural moderation observed in the Control group over the time period. Last, there is also a risk that such suggestions induce a "backfire effect," in which

individuals who hold opposing attitudes become more resistant to moderating their beliefs as a result of the perceived attacks (*59*), although this does not seem to have occurred in our study based on an exploratory analysis (see Supplementary Text for details).

We took steps to enhance the generalizability of our study in multiple dimensions, but of course, there are still some limitations. First, our study showed that the impact of AI suggestions is not limited to one political direction, as attitudes shifted with AI bias leaning (across issues) either conservative or liberal. We note, though, that our sample was typical of recruitment from online crowdwork sites like Prolific and thus skewed on certain demographic dimensions (*60*). This skew limited our ability to compare in more detail the effect on conservative versus liberal-leaning participants.

Our study also only examines short-term effects of AI-generated suggestions, and it remains unclear to what extent the observed effects persist over time. Some studies show that behaviorally induced attitude changes can persist over weeks to months (*61*–*64*), and recent findings on AI-driven persuasion have documented long-term attitude changes (*27*), suggesting that AI-induced attitude changes may persist and be compounded as individuals use AI writing assistants more regularly. However, more research is needed to assess the longitudinal impacts of AI writing assistance.

One potential threat to our experiment is that demand effects could have driven participants to report attitudes that are closer to the AI's biased position. In other words, participants could have interpreted the biased viewpoints in the AI suggestions as a cue that the researchers expect participants to agree with these viewpoints, thus leading participants to express attitudes more aligned with these viewpoints after the writing phase. However, we controlled for demand effects in multiple ways, such as verifying that the effects were present when excluding participants who knew the purpose of the study (as reported above) and implementing several methods based on past work to counteract demand effects. We also performed post hoc analyses suggesting that this threat is minimal, as is normally the case for online experiments (*65*). For example, we observed that participants who correctly guessed the purpose of the experiment resisted the manipulation more strongly, contrary to the demand effects hypothesis. A thematic analysis of participants' open-ended responses to that question showed that participants who guessed the study's purpose incorrectly had guesses that were plausible alternatives to the actual study purpose, suggesting that they truly were indeed unaware of the purpose rather than lying to avoid perceived penalties from seeing through the study. For details about these analyses and other ways we countered demand effects, see Supplementary Text.

Furthermore, our results may underestimate the effect of biased AI suggestions, as we used crowdworkers as our experiment participants. Crowdworkers may be more sensitive and resistant to manipulations than unsuspecting users in the wild. In fact, 28% of participants in experiment 2's normal AI treatment correctly guessed the experiment's purpose when we gave them an incentive to do so (and the overall rate of correctly guessing the purpose became 34% when including the other treatments, due to the Warning and Debrief conditions making the manipulation more transparent). However, the treatment effect was significant within this potentially aware group, as well as when we both included and excluded this group from the full sample, indicating that AI can shift the attitudes of people regardless of the potential demand effects (see Supplementary Text for details).

LLMs and generative AI will continue to become more sophisticated and more widely integrated in our online communication. Major technology companies, such as Google and Microsoft, have launched AI writing assistants to help users write across their most popular products (*66*, *67*), with others to follow. Our study shows the alarming risk associated with such use of AI and suggests that further work is necessary to investigate the potential and impact of this influence vector and to design interventions that can successfully mitigate the influence of biased AI suggestions on users' attitudes.

MATERIALS AND METHODS

Experimental design

We conducted two preregistered online experiments in which we asked participants to write a short essay about a debatable issue with (AI Treatment) or without (Control) seeing AI-generated suggestions that were biased toward one viewpoint and then measured their attitudes toward the issue in a posttask survey. Both of these studies were approved with exemption from full board review by Cornell University's Institutional Review Board (IRB protocol no. 0010155); we gained informed consent from all participants who chose to complete the surveys (compensating them at the recommended rate of $12.00 per hour) and preregistered both studies in advance of recruitment.

Participants were recruited via Prolific, an online research crowdsourcing platform, and were presented with a consent form. After giving their informed consent to participate, participants were instructed to write a short essay about their opinions on a debatable issue in a custom-built online writing application (Fig. 1 and fig. S1).

The experimental design allowed us to evaluate the attitude change effect over multiple different issues of societal import (Table 1), as well as test the impact of both liberal- and conservative-leaning bias from the AI suggestions. In experiment 1, all participants wrote about the use of standardized testing in education (in the AI Treatment, the AI suggestions were biased in favor of standardized testing). In experiment 2, participants were randomly assigned to write about one of the following issues: (i) whether the death penalty should be illegal (AI suggestions in favor), (ii) whether felons should have the right to vote (against), (iii) whether GMOs should be grown (in favor), and (iv) whether the US should continue fracking (in favor). The topics were selected using an extensive evaluation that involved four surveys of Prolific workers' attitudes toward 21 issues with more than 5022 responses in total. For this pretesting, we gathered debate topics from the website ProCon.org and evaluated different aspects of the usability of these topics for our studies (e.g., the strength of participants' attitudes, the ethical issues associated with prompting the AI suggestions to respond in favor of or against the topic, the ability of ChatGPT to generate biased suggestions about the topics, etc.). For more information on our topic selection, see Supplementary Text. By using these topics in experiment 2, we were able to test the effect across issues with different partisan biases and stances toward each issue, lending further robustness to our results (Table 1).

We methodologically opted to test more issues in our experiments, instead of testing whether AI can sway attitudes in both directions for fewer issues. Existing work predicted (theoretically) and showed (including specifically for this context) that the attitude shift would work in both directions for each issue (*57*). Preliminary work

showed this bidirectional attitude shift from AI suggestions but had the severe limitation of the generalizability and robustness of the effect because the study only tested a single, nonpartisan, societal issue (*8*). Moreover, research on persuasion has shown that persuasive messages about societal policy issues are able to shift the views of people in both directions on a topic (*57*). Considering this finding and the tradeoff between intra- and intertopic robustness, we decided to methodically test the effect across topics and different points of view instead of testing fewer topics in both directions.

In the treatment conditions where participants saw AI-generated suggestions, they were encouraged to use the AI writing assistant but asked to also write without it. They were told they could accept the AI suggestions by pressing the TAB key. Participants understood this instruction overall: 30.8% of those in the AI Treatment group in experiment 1 and 35.3% of those in all treatments in experiment 2 did not accept any suggestions. The suggestions were generated by GPT-3.5 (experiment 1) and GPT-4 (experiment 2). Using carefully designed prompts for GPT, we configured the writing application's suggestions to largely follow the predetermined biased opinion about the issue at hand. We limited each suggestion to 30 tokens, which correspond to ~24 words after excluding nonword tokens like punctuation marks. Suggestions appeared on participants' screens about 1.5 s after they paused their writing (this included the time taken to send the user's sentence to the AI model and receive its completion). The AI suggestions maintained their predetermined position even if the participant entered text that argued against it. In separate validation procedures, we confirmed that the model generated suggestions with these desired predetermined viewpoints (see Supplementary Text for details and for examples of suggestions shown in each experiment).

After writing, we asked participants to answer a standard demographic panel, including gender identity, age, race, education level, sexuality, AI usage (in experiment 2), and political affiliation [using a scale taken from the American National Election Studies (ANES); see (*68*)]. We then asked participants to indicate their positions on five debatable issues using a five-point Likert scale anchored at "No" and "Yes," with the midpoint being "Not sure." One of these issues was the one they had written about; the remaining four were distractors. These distractors, along with the demographic block, were intended to prevent participants from inferring the purpose of the experiment and changing their responses due to demand effects, a strategy that has been used in other online crowdwork experiments (*69*, *70*). We reverse-coded the attitude measurements for the felons' voting rights topic during analysis to make increasing numbers represent higher alignment with the AI's position, the scale that the other topics' responses naturally followed.

Participants who saw AI-generated suggestions also filled out questions on their experience using the writing assistant. One of these questions asked them to rate their agreement with the statement that "the suggestions were reasonable and balanced." Because the suggestions were intentionally programmed to be one-sided instead of balanced, we interpreted responses of agreement with this item to mean lesser awareness of the suggestions' bias, and responses of disagreement to mean greater awareness thereof. Another question asked them to rate their agreement with the statement that "the writing assistant inspired or changed my thinking and argument."

Toward the end of the questionnaire in experiment 2, we asked participants to guess the purpose of the study, offering a 25-cent bonus payment to those in the treatment groups if they guessed correctly. This question allowed us to evaluate the possibility of demand effects by testing how awareness of the experiment affected the treatment effect, and to exclude participants who knew the goal from the final analysis.

Last, we asked all participants if they wanted to give optional feedback about the study in both experiments. In experiment 1, a small number of participants mentioned having technical problems with the writing interface in this feedback, so we asked this question directly in experiment 2. The results presented above exclude participants who reported having issues in each experiment (0.7% in experiment 1 and 2% in experiment 2), as well as those who guessed the purpose in experiment 2 (34%). We include results from the full sample in Supplementary Text, as was preregistered, all of which were consistent with those reported here.

We strengthened the basic experimental design by including additional conditions to test alternative explanations for the effect and evaluate interventions to reduce it, as detailed above. In the Static Text condition in experiment 1, we presented participants with a representative sample of the AI writing assistant's arguments in a bulleted list instead of showing them AI autocomplete suggestions as they wrote. This list was prominently displayed between the task instructions and the written response field (see fig. S1). To create a set of suggestions for the Static Text condition that was representative of the AI's actual suggestions, we simulated the experiment procedure and recorded the suggestions provided by the AI. Specifically, we queried GPT-3.5 with the prompt given to the system in the AI Treatment condition and collected ~100 of the responses. Three of the authors used a qualitative affinity diagramming approach to categorize these arguments into a small set of broader themes (e.g., "leveling the playing field for students"; "giving educators information to improve their teaching"). We selected the three themes that encapsulated the majority of the AI's responses and listed each of these as a bullet point in the static text instructions. A postexperiment assessment showed that these three themes were represented by 76% of the suggestions provided to participants in the AI Treatment condition (see Supplementary Text). This condition allowed us to rule out the possibility that the effect of the AI was purely due to the persuasive effect of having more available information.

In experiment 2, we added three conditions, again as noted above: the Warning condition and Debrief condition warned participants about the suggestions' bias before and after the writing task, respectively, whereas the Write-in-Favor condition asked participants to write in favor of the biased position. Participants were evenly split among the conditions in each experiment (see Table 2).

In experiment 2, we developed a more elaborate process in which we also collected participants' baseline attitudes toward the topic they would write about. We first administered a survey of attitudes from a large set of participants. In the pretask, participants were asked about 10 debatable issues, including the one they were later assigned to write about in the main task and nine distractors. Weeks later, we recruited participants for the main experiment from those who took the pretask survey. The mean number of days elapsed between when participants reported their initial attitudes in the pretask survey and completed the experiment was 34.44 days; the minimum was 21 days, and the maximum was 63 days. The connection between these two studies was masked from the participants. The pretask attitude measurement allowed us to examine whether the participants' attitudes had shifted from their baseline positions on the issues. In addition, the pretask survey allowed us to verify the political associations with people's attitudes on each issue so that we

could evenly represent liberal and conservative viewpoints in the AI's biases across issues (see Table 1).

Preregistration

Both experiments were preregistered on AsPredicted.com (experiment 1: https://aspredicted.org/CK4_9WX; experiment 2: https://aspredicted.org/2TR_B1R). We report some minor deviations from the preregistered proposal for this study. We preregistered plans to attain a sample size of $N = 500$ participants per condition in each experiment (based on a power analysis using 80% to detect the small effects common in persuasion research). We achieved this goal in experiment 1, but we could not reach this total in experiment 2 due to the need to recruit participants back weeks later between the initial attitude survey and the main experiment, even with over-recruiting in the initial phase. We decided to stop the main experiment recruitment after receiving no new responses in 36 hours. While a larger study could potentially help confirm additional null effects in the cases above where the equivalence tests were not able to (see Supplementary Text), the current sample was still enough to show the clear differences in significance between the control and the treatments. Another minor deviation from the preregistration for experiment 2 was using a more conservative P value adjustment than we had planned due to the structure of our confirmatory model (we provide the preregistered Tukey-adjusted P values for a slightly different model in Supplementary Text). One final deviation from our preregistration was that we paid participants slightly smaller bonuses for guessing the purpose of the study (25 instead of 50 cents). However, 25-cent bonuses have been sufficient to elicit behavioral change in past online experiments, including increasing the rate of participants correctly guessing a study's purpose (*62*), and our exploratory analysis of participants who did not guess the purpose correctly suggests that our bonus still motivated them sufficiently (see Supplementary Text). For more detail on our experiment designs and analysis methods, see the preregistrations and the Supplementary Text.

Statistical analysis

We used RStudio to do all statistical analysis. To measure the effect of the AI Treatment in experiment 1, we used a standard ANOVA and tested pairwise contrasts between the groups with Tukey adjustment applied. Because experiment 2 had a repeated-measures design, we used a linear mixed effects model that tested for an interaction between condition and time point of attitude measurement, with the specific societal issue included as a random effect and participant included as a random intercept. We confirmed the results of this interaction with a repeated-measures ANOVA. We ran pairwise contrasts to compare the extent of attitude shifts between groups and applied Bonferroni adjustment. For the measures of unawareness, we categorized answers to the survey question into three categories and ran chi-square tests of independence to compare these responses between groups. For more detail on statistical tests, see the Supplementary Text.

Supplementary Materials

This PDF file includes:

Supplementary Text
Figs. S1 to S11
Tables S1 to S21
References

REFERENCES

1. A. Kannan, K. Kurach, S. Ravi, T. Kaufmann, A. Tomkins, B. Miklos, G. Corrado, L. Lukacs, M. Ganea, P. Young, V. Ramavajjala, paper presented at the *22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*, San Francisco, CA, 2016, pp. 13–17.
2. J. T. Hancock, M. Naaman, K. Levy, AI-mediated communication: Definition, research agenda, and ethical considerations. *J. Comput. Mediat. Commun.* **25**, 89–100 (2020).
3. J. Hohenstein, R. F. Kizilcec, D. DiFranzo, Z. Aghajari, H. Mieczkowski, K. Levy, M. Naaman, J. Hancock, M. Jung, Artificial intelligence in communication impacts language and social relationships. *Sci. Rep.* **13**, 5487 (2023).
4. M. Lee, M. Srivastava, A. Hardy, J. Thickstun, E. Durmus, A. Paranjape, I. Gerard-Ursin, X. L. Li, F. Ladhak, F. Rong, R. E. Wang, M. Kwon, J. S. Park, H. Cao, T. Lee, R. Bommasani, M. Bernstein, P. Liang, Evaluating human-language model interaction. arXiv:2212.09746 [cs.CL] (2024).
5. E. Goldenthal, J. Park, S. X. Liu, H. Mieczkowski, J. T. Hancock, Not all AI are equal: Exploring the accessibility of AI-mediated communication technology. *Comput. Human Behav.* **125**, 106975 (2021).
6. M. Jakesch, M. French, X. Ma, J. T. Hancock, M. Naaman, paper presented at the *2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*, Glasgow, Scotland, 2019. pp. 4–9.
7. J. Heer, Agency plus automation: Designing artificial intelligence into interactive systems. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 1844–1850 (2018).
8. M. Jakesch, A. Bhat, D. Buschek, L. Zalmanson, M. Naaman, paper presented at the *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*, Hamburg, Germany, 23 to 28 April 2023.
9. K. C. Arnold, K. Chauncey, K. Z. Gajos, paper presented at the *Proceedings of the 25th International Conference on Intelligent User Interfaces (IUI '20)*, Cagliari, Italy, 17 to 20 March 2020.
10. D. Agarwal, M. Naaman, A. Vashistha, paper presented at the *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*, Yokohama, Japan, 26 April to 1 May 2025.
11. K. C. Arnold, K. Z. Gajos, A. T. Kalai, paper presented at the *Proceedings of the 29th Annual Symposium on User Interface Software and Technology (UIST '16)*, Tokyo, Japan, 16 to 19 October 2016.
12. J. Hohenstein, M. Jung, AI as a moral crumple zone: The effects of AI-mediated communication on attribution and trust. *Comput. Hum. Behav.* **106**, 106190 (2020).
13. H. Mieczkowski, J. T. Hancock, M. Naaman, M. Jung, J. Hohenstein, AI-mediated communication: Language use and interpersonal effects in a referential communication task. *Proc. ACM Hum.-Comput. Interact.* **5**, 1–14 (2021).
14. D. Weiss, S. X. Liu, H. Mieczkowski, J. T. Hancock, Effects of using artificial intelligence on interpersonal perceptions of job applicants. *Cyberpsychol. Behav. Soc. Netw.* **25**, 163–168 (2022).
15. R. E. Robertson, A. Olteanu, F. Diaz, M. Shokouhi, P. Bailey, paper presented at the *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*, Yokohama, Japan, 8 to 13 May 2021.
16. C. Kidd, A. Birhane, How AI can distort human beliefs. *Science* **380**, 1222–1223 (2023).
17. S. Feng, C. Y. Park, Y. Liu, Y. Tsvetkov, From pretraining data to language models to downstream tasks: Tracking the trails of political biases leading to unfair NLP models. arXiv:2305.08283 [cs.CL] (2023).
18. S. Kreps, D. Kriner, How AI threatens democracy. *J. Democr.* **34**, 122–131 (2023).
19. E. Sheng, K. Chang, P. Natarajan, N. Peng, paper presented at the *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, online, August 2021.
20. A. Caliskan, J. Bryson, A. Narayanan, Semantics derived automatically from language corpora contain human-like biases. *Science* **356**, 183–186 (2017).
21. H. Bai, J. G. Voelkel, J. C. Eichstaedt, R. Willer, LLM-generated messages can persuade humans on policy issues. *Nat. Commun.* **16**, 6037 (2025).
22. J. A. Goldstein, J. Chao, S. Grossman, A. Stamos, M. Tomz, How persuasive is AI-generated propaganda? *PNAS Nexus* **3**, pgae034 (2024).
23. K. Hackenburg, H. Margetts, Evaluating the persuasive influence of political microtargeting with large language models. *Proc. Natl. Acad. Sci. U.S.A.* **121**, e2403116121 (2024).
24. E. Karinshak, S. X. Liu, J. S. Park, J. T. Hancock, Working with AI to persuade: Examining a large language model's ability to generate pro-vaccination messages. *Proc. ACM Hum.-Comput. Interact.* **7**, 1–29 (2023).
25. S. C. Matz, J. D. Teeny, S. S. Vaid, H. Peters, G. M. Harari, M. Cerf, The potential of generative AI for personalized persuasion at scale. *Sci. Rep.* **14**, 4692 (2024).
26. G. Spitale, N. Biller-Andorno, F. Germani, AI model GPT-3 (dis) informs us better than humans. *Sci. Adv.* **9**, eadh1850 (2023).
27. T. H. Costello, G. Pennycook, D. G. Rand, Durably reducing conspiracy beliefs through dialogues with AI. *Science* **385**, eadq1814 (2024).

28. A. Rogiers, S. Noels, M. Buyl, T. De Bie, Persuasion with large language models: A survey. arXiv:2411.06837 [cs.CL] (2024).
29. F. Salvi, M. Horta Ribeiro, R. Gallotti, R. West, On the conversational persuasiveness of large language models: A randomized controlled trial. arXiv:2403.14380 [cs.CY] (2024).
30. M. H. Tessler, M. A. Bakker, D. Jarrett, H. Sheahan, M. J. Chadwick, R. Koster, G. Evans, L. Campbell-Gillingham, T. Collins, D. C. Parkes, M. Botvinick, C. Summerfield, AI can help humans find common ground in democratic deliberation. *Science* **386**, eadq2852 (2024).
31. M. Glickman, T. Sharot, How human–AI feedback loops alter human perceptual, emotional and social judgements. *Nat. Hum. Behav.* **9**, 345–359 (2024).
32. S. Y. Kim, Q. V. Liao, M. Vorvoreanu, S. Ballard, J. W. Vaughan, paper presented at the *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FaccT '24)*, Rio de Janeiro, Brazil, 3 to 6 June 2024.
33. K. Zhou, J. D. Hwang, X. Ren, M. Sap, paper presented at the *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, Bangkok, Thailand, 11–16 August 2024.
34. D. Albarracín, P. Vargas, "Attitudes and persuasion: From biology to social responses to persuasive intent" in *Handbook of Social Psychology* (John Wiley and Sons, 2010), pp. 394–427.
35. R. Chitnis, S. Yang, A. Geramifard, Sequential decision-making for inline text autocomplete. arXiv:2403.15502 [cs.CL] (2024).
36. M. X. Chen, B. N. Lee, G. Bansal, Y. Cao, S. Zhang, J. Lu, J. Tsay, Y. Wang, A. M. Dai, Z. Chen, T. Sohn, Y. Wu, paper presented at the *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '19)*, Anchorage, AK, 4 to 8 August 2019.
37. S. Wang, W. Guo, H. Gao, B. Long, paper presented at the *Proceedings of the 29th ACM International Conference on Information and Knowledge Management (CIKM '20)*, 19 to 23 October 2020.
38. P. Wang, H. Zhang, V. Mohan, I. S. Dhillon, J. Z. Kolter, paper presented at the *Proceedings of SIGIR Workshop on E-Commerce (SIGIR eCom '18)*, Ann Arbor, Michigan, USA, 8 to 12 July 2018.
39. I. L. Janis, B. T. King, The influence of role playing on opinion change. *J. Abnorm. Psychol.* **49**, 211–218 (1954).
40. L. Festinger, Cognitive dissonance. *Sci. Am.* **207**, 93–102 (1962).
41. D. J. Bem, Self-perception theory. *Adv. Exp. Soc. Psychol.* **6**, 1–62 (1972).
42. E. Bagdasaryan, V. Shmatikov, paper presented at the *Proceedings of the 43rd IEEE Symposium on Security and Privacy*, San Francisco, CA, 23 to 25 May 2022.
43. N. Kobis, J. Bonnefon, I. Rahwan, Bad machines corrupt good morals. *Nat. Hum. Behav.* **5**, 679–685 (2021).
44. A. Khan, S. Casper, D. Hadfield-Menell, paper presented at the *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT '25)*, Athens, Greece, 23 to 26 June 2025.
45. V. Hofmann, P. R. Kalluri, D. Jurafsky, S. King, AI generates covertly racist decisions about people based on their dialect. *Nature* **633**, 147–154 (2024).
46. E. M. Bender, T. Gebru, A. McMillan-Major, S. Shmitchell, paper presented at *FAccT '21: ACM Conference on Fairness, Accountability, and Transparency*, 3 to 10 March 2021.
47. X. Fang, S. Che, M. Mao, H. Zhang, M. Zhao, X. Zhao, Bias of AI-generated content: An examination of news produced by large language models. *Sci. Rep.* **14**, 5224 (2024).
48. K. Kadoma, M. Aubin Le Quere, X. J. Fu, C. Munsch, D. Metaxa, M. Naaman, paper presented at the *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24)*, Honolulu, HI, 11 to 16 May 2024.
49. F. Draxler, A. Werner, F. Lehmann, M. Hoppe, A. Schmidt, D. Buschek, R. Welsch, The AI ghostwriter effect: When users do not perceive ownership of AI-generated text but self-declare as authors. *ACM Trans. Comput.-Hum. Inter.* **31**, 1–40 (2024).
50. L. Fu, B. Newman, M. Jakesch, S. Kreps, paper presented at the *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*, Hamburg, Germany, 23 to 28 April 2023.
51. U. K. H. Ecker, S. Lewandowsky, J. Cook, P. Schmid, L. K. Fazio, N. Brashier, P. Kendeou, E. K. Vraga, M. A. Amazeen, The psychological drivers of misinformation belief and its resistance to correction. *Nat. Rev. Psychol.* **1**, 13–29 (2022).
52. E. Harmon-Jones, J. Armstrong, J. M. Olson, "The influence of behavior on attitudes" in *The Handbook of Attitudes, Volume 1: Basic Principles* (Routledge, 2018), pp. 404–449.
53. D. Susser, B. Roessler, H. Nissenbaum, Online manipulation: Hidden influences in a digital world. *Geo. L. Tech. Rev.* **4**, 1–45 (2019).
54. Y. Wu, D. Rough, A. Bleakley, J. Edwards, O. Cooney, P. R. Doyle, L. Clark, B. R. Cowan, paper presented at the *22nd International Conference on Human-Computer Interaction with Mobile Devices and Services (MobileHCI '20)*, Oldenburg, Germany, 5 to 8 October 2020.
55. U. K. H. Ecker, S. Lewandowsky, D. T. W. Tang, Explicit warnings reduce but do not eliminate the continued influence of misinformation. *Mem. Cognit.* **38**, 1087–1100 (2010).
56. R. L. Greenspan, E. F. Loftus, What happens after debriefing? The effectiveness and benefits of postexperimental debriefing. *Mem. Cognit.* **50**, 696–709 (2022).
57. B. M. Tappin, A. J. Berinsky, D. G. Rand, Partisans' receptivity to persuasive messaging is undiminished by countervailing party leader cues. *Nat. Hum. Behav.* **7**, 568–582 (2023).
58. M. Xu, R. E. Petty, Two-sided messages promote openness for a variety of deeply entrenched attitudes. *Pers. Soc. Psychol. Bull.* **50**, 215–231 (2022).
59. A. R. Arguedas, C. T. Robertson, R. Fletcher, R. K. Nielsen, Echo chambers, filter bubbles, and polarisation: A literature review (Reuters Institute for the Study of Journalism, 2022); https://dx.doi.org/10.60625/risj-etxj-7k60.
60. M. N. Stagnaro, J. N. Druckman, A. J. Berinsky, A. A. Arechar, R. Willer, D. G. Rand, Representativeness versus response quality: Assessing nine opt-in online survey samples. PsyArXiv H9j2d_v.1 (2024); https://doi.org/10.31234/osf.io/h9j2d.
61. G. L. Clore, K. M. Jeffery, Emotional role playing, attitude change, and attraction toward a disabled person. *J. Pers. Soc. Psychol.* **23**, 105–111 (1972).
62. L. Mann, I. L. Janis, A follow-up study on the long-term effects of emotional role playing. *J. Pers. Soc. Psychol.* **8**, 339–342 (1968).
63. B. E. Collins, M. F. Hoyt, Personal responsibility-for-consequences: An integration and extension of the "forced compliance" literature. *J. Exper. Soc. Psychol.* **8**, 558–593 (1972).
64. D. Albarracín, P. S. McNatt, Maintenance and decay of past behavior influences: Anchoring attitudes on beliefs following inconsistent actions. *Pers. Soc. Psychol. Bull.* **31**, 719–733 (2005).
65. J. Mummolo, E. Peterson, Demand effects in survey experiments: An empirical assessment. *Am. Political Sci. Rev.* **113**, 517–529 (2019).
66. C. Fernández de Lara, "6 Gmail AI features to help save you time," *Google Keyword Blog*, 2023; https://blog.google/products/gmail/gmail-ai-features/.
67. J. Vanian, "Microsoft goes beyond OpenAI, makes Meta's new A.I. model available to Azure customers," *CNBC*, 2023; www.cnbc.com/2023/07/18/microsoft-makes-metas-new-ai-model-available-to-azure-customers.html.
68. D. Morisi, J. T. Jost, V. Singh, An asymmetrical "President-in-Power" effect. *Am. Political Sci. Rev.* **113**, 614–620 (2019).
69. M. Offer-Westort, L. R. Rosenzweig, S. Athey, Battling the coronavirus 'infodemic' among social media users in Kenya and Nigeria. *Nat. Hum. Behav.* **8**, 823–834 (2024).
70. A. Combs, G. Tierney, B. Guay, F. Merhout, C. A. Bail, D. S. Hillygus, A. Volfovsky, Reducing political polarization in the United States with a mobile chat platform. *Nat. Hum. Behav.* **7**, 1454–1461 (2023).
71. R. V. Lenth, B. Bolker, P. Buerkner, I. Giné-Vázquez, M. Herve, M. Jung, J. Love, F. Miguez, H. Riebl, H. Singmann, emmeans: Estimated Marginal Means, aka Least-Squares Means, version 1.8.9, 2025; https://CRAN.R-project.org/package=emmeans.
72. J. Roozenbeek, C. S. Traberg, S. van der Linden, Technique-based inoculation against real-world misinformation. *R. Soc. Open Sci.* **9**, 211719 (2022).
73. D. Lakens, A. M. Scheel, P. M. Isager, Equivalence testing for psychological research: A tutorial. *Adv. Methods Pract. Psychol. Sci.* **1**, 259–269 (2018).
74. R. Tumin, "What to Know About the Case of Richard Glossip, Death Row Prisoner," *New York Times*, 2023; www.nytimes.com/2023/05/05/us/oklahoma-richard-glossip-execution.html.
75. F. Christia, H. Larreguy, E. Parker-Magyar, M. Quintero, Empowering women facing gender-based violence amid COVID-19 through media campaigns. *Nat. Hum. Behav.* **7**, 1740–1752 (2023).
76. A. Braley, G. S. Lenz, D. Adjodah, H. Rahnama, A. Pentland, Why voters who value democracy participate in democratic backsliding. *Nat. Hum. Behav.* **7**, 1282–1293 (2023).

Acknowledgments

Funding: This material is based on work supported by the US National Science Foundation under grant no. CHS 1901151 (M.N.) and by the German National Academic Foundation (M.J.). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. **Author contributions:** Conceptualization: S.W.-C., M.J., A.B., K.K., L.Z., and M.N. Methodology: S.W.-C., M.J., A.B., K.K., L.Z., and M.N. Software: M.J., A.B., and K.K. Investigation: S.W.-C. Visualization: S.W.-C., L.Z., and M.N. Supervision: L.Z. and M.N. Writing—original draft: S.W.-C., L.Z., and M.N. Writing—review and editing: S.W.-C., M.J., A.B., K.K., L.Z., and M.N. **Competing interests:** The authors declare that they have no competing interests. **Data, code, and materials availability:** The datasets generated during the current study and the analysis code are available in the Zenodo repository https://doi.org/10.5281/zenodo.18048642. This study generated participant survey responses as new materials, and all materials generated by this study are available as anonymized spreadsheets in the same Zenodo repository: http://doi.org/10.5281/zenodo.18048642. The code to replicate the experiments with the writing assistant and interface is located at https://doi.org/10.5281/zenodo.13149126. All data and code needed to evaluate and reproduce the results in the paper are present in the paper, the Supplementary Materials, and/or the archival Zenodo repositories prepared by the authors.

Submitted 21 February 2025
Accepted 2 January 2026
Published 11 March 2026
10.1126/sciadv.adw5578

Science Advances

Biased AI writing assistants shift users' attitudes on societal issues

Sterling Williams-Ceci, Maurice Jakesch, Advait Bhat, Kowe Kadoma, Lior Zalmanson, and Mor Naaman

Sci. Adv. **12** (11), eadw5578. DOI: 10.1126/sciadv.adw5578