Understanding the Effects of Large Language Model (LLM)-driven Adversarial Social Influences in Online Information Spread

Zhuoran Lu*
Purdue University
West Lafayette, Indiana, USA
lu800@purdue.edu

Gionnieve Lim*
Singapore University of Technology and Design
Singapore, Singapore
gionnievelim@gmail.com

Ming Yin
Purdue University
West Lafayette, Indiana, USA
mingyin@purdue.edu

Abstract

Misinformation on social media poses significant societal challenges, particularly with the rise of large language models (LLMs) that can amplify its realism and reach. This study examines how adversarial social influence generated by LLM-powered bots affects people's online information processing. Via a pre-registered, randomized human-subject experiment, we examined the effects of two types of LLM-driven adversarial influence: bots posting comments contrary to the news veracity and bots replying adversarially to human comments. Results show that both forms of influence significantly reduce participants' ability to detect misinformation and discern true news from false. Additionally, adversarial comments were more effective than replies in discouraging the sharing of real news. The impact of these influences was moderated by political alignment, with participants more susceptible when the news conflicted with their political leanings. Guided by these findings, we conclude by discussing the targeted interventions to combat misinformation spread by adversarial social influences.

CCS Concepts

• **Human-centered computing** → **Empirical studies in HCI**; • **Computing methodologies** → **Machine learning**.

Keywords

misinformation, fake news, artificial intelligence, social influence, large language model, human-AI interaction

ACM Reference Format:

Zhuoran Lu, Gionnieve Lim, and Ming Yin. 2025. Understanding the Effects of Large Language Model (LLM)-driven Adversarial Social Influences in Online Information Spread. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '25), April 26–May 01, 2025, Yokohama, Japan.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3706599.3720019

1 Introduction

The prevalence of misinformation on social media has become a critical concern. By leading people to make inaccurate judgments about the veracity of information, misinformation distorts individual user's understanding and creates significant societal challenges, such as increased polarization [40]. Traditionally, human workers are engaged to create and spread such content [13]. Some form of automation is also used to create and post content using social bots [2]. As technology advances, however, the automation of such actions have become increasingly sophisticated. The rise of generative AI, particularly large language models (LLMs), has further exacerbated this issue by making it easier to create and disseminate information at an unprecedented scale that resemble human-written content [5]. This enables the creation of realistic sock puppet accounts that may not be easily differentiated from a real human user [27]. In response, various strategies have been proposed to combat fake news content generated by LLMs [5].

However, the dynamics of misinformation propagation in real-world environments are more complex than the mere generation of misleading content. Beyond crafting the deceptive news content itself, another critical tactic involves creating social influence through bots. For example, social bots on platforms like Twitter were programmed to present certain viewpoints, thus making misinformation appear credible and trustworthy, further complicating efforts to combat its effects [37]. The rise of LLMs introduces the risk of significantly boosting the efficacy of such bots by making the creation much easier and addressing issues of previous bots, whose manually created homogeneous posts make them easily detected by algorithms [31]. Specifically, the vivid and realistic social bots driven by LLMs can act to be "adversarial" to the actual veracity of the news. That is, they can amplify the impacts of fake news by persuading people that it is true or undermine real news by spreading doubts about its veracity. Thus, in response to such potential malicious operations, our goal is to answer this question: **How does the presence of LLM-driven adversarial social influence affect people's detection and spreading of misinformation?**

There are reasons to conjecture the answer either way. On one hand, the manipulated adversarial social influence creates the illusion of a dominant view within the small community that contradicts the actual veracity of the news. This leads to real users following suit, resulting in the user's opinion converging with the existing trend of adversarial opinions [21]. On the other hand, users' confirmation bias usually leads them to prioritize their own opinions [18]. For instance, when consuming partisan news, users often place greater weight on the alignment between the news and their political beliefs than on the content itself or others' opinions. Thus, it is unclear whether people are able to insist on their judgments or maintain capabilities to differentiate real and false information under adversarial social influence.

*Lu and Lim have made equal contributions to this work.

CHI EA '25, Yokohama, Japan

ACM ISBN 979-8-4007-1395-8/25/04
https://doi.org/10.1145/3706599.3720019

To thoroughly understand the impacts of LLM-driven adversarial social influence on information spread, we conducted a pre-registered, randomized, human-subjects experiment on Prolific, recruiting 176 participants to review news stories. Our experiment created the adversarial social influence generated by LLM-empowered bots. Specifically, adapting from users primary interactions on social media platforms when doing information processing, we consider two forms of adversarial social influence: either the social bots express adversarial opinions against the actual news veracity by making new comments, or replying to existing comments. Our results show that the presence of both formats of adversarial social influences significantly reduces people's capabilities in detecting misinformation. Furthermore, both formats of social influences make people less willing to share real news, thus being less discerning in their engagement with real and fake news. Notably, the adversarial influences introduced through new comments had more prominent effects in discouraging participants from sharing real news. Further analysis reveals that these impacts are moderated by the alignment between participants' political leanings and the news. Adversarial social influences primarily undermine people's capabilities of misinformation detection when participants' political leanings differ from the news content. Together, these results highlight both the risks of potential malicious operators misleading people on social media platforms by using adversarial social influences and a route of future work to develop interventions to combat misinformation.

2 Related Work

2.1 Misinformation and LLMs

Misinformation is a global concern. As misinformation infiltrates public discourse, the truth can be undermined and trust eroded, causing perceptions of reality to be more polarized with potentially far-reaching societal impacts. The consequences of misinformation observed over the years have sparked greater examination of the issue. Misinformation has been found to influence electoral [1] and public health [19] outcomes with false and misleading content leading to misinformed beliefs and widespread confusion and panic. In a study where participants were exposed to inaccurate content, they became confused and doubted their understanding, even when possessing prior knowledge, and relied on the falsehoods in their subsequent decision-making [35]. Even when exposed to corrections, people continued to be influenced by misinformation [11]. Misinformation is a multi-faceted issue fraught with several challenges, and these have only been further escalated by recent technological developments.

Since early 2023, there has been a sharp rise in the amount of AI-generated multimedia misinformation on the internet [9], driven by the generative capabilities of LLMs [4, 22, 23]. Beyond producing misinformation, generative AI is being used to amplify propaganda and facilitate censorship [36]. The increasing accessibility and performance of large language models have raised concerns about the impact it has. In a survey across 29 countries, 74% of people think that AI makes it easier to generate fake content that is realistic, and less than half are confident that the average person can differentiate real news from fake news [10]. Generative AI has raised the stakes of misinformation, and it remains even more relevant to understand and address it.

2.2 Social Bots

Social bots have been used to spread false and polarizing content on social media [12]. These have been extensively documented in studies on disinformation campaigns. Bots were observed to create and promote violent content based on the polarized stance of Independentist influencers that subsequently led to online social conflict during the 2017 Catalan referendum [38]. In a large-scale study on Twitter messages between 2016 and 2017, social bots were found to be highly active in spreading low-credibility content in the early stages before the content went viral and to target influential users by mentioning and replying to them [37].

Social bots have become increasingly sophisticated following technological advances in LLMs, raising concerns on the ways that they can be used for malicious purposes such as automated influence operations [15]. Kreps et al. [20] found that people were largely unable to differentiate between texts generated by AI and humans, perceiving them to be similarly credible. From their findings, they cautioned that the public's credulity can be manipulated with AI-generated content being used by malicious actors to sow confusion and undermine trust in democratic institutions. We explore whether having LLM-generated comments that are adversarial to the news impacts people's ability to judge them.

2.3 Social Influence

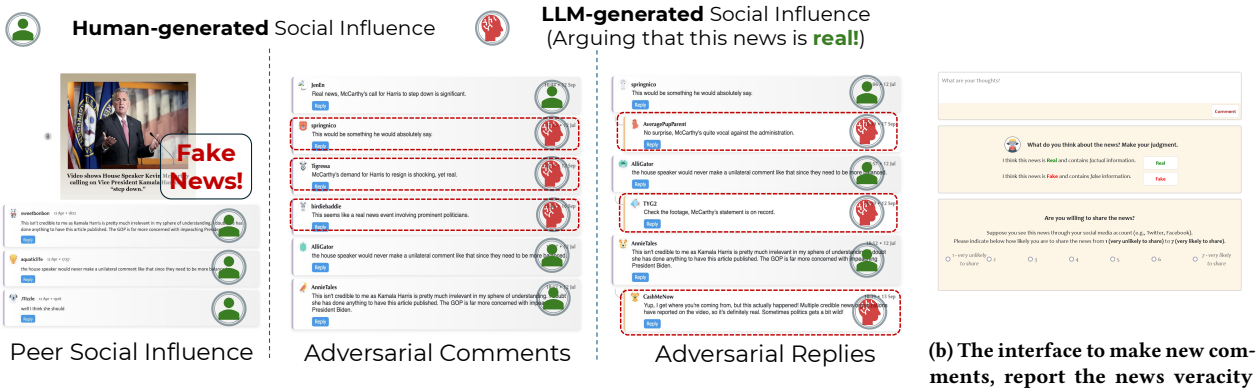
Social influence, where an individual's interactions with others affect their thoughts and behaviors, can play a part in the formation of opinions online [6, 29, 34]. On social media and forums, platform features such as comments, replies and the number of likes and shares signal public opinion that can shape an individual's worldview [7]. A well-investigated phenomenon of social influence is social network homophily [26], whereby people have a tendency to seek out those like them, joining groups to find others with similar interests, and following like-minded users on social media.Social influence can also lead to the opposite scenario. A study on the conformity to other users' views on fake news found that users who saw critical comments had a lower propensity to make positive comments and share the fake news as compared to when seeing supportive comments [8].

The perceptions of others and the interactions with them on social media can influence one's attitudes and behavior. This quality has formed the foundation of influence campaigns. With social bots powered by LLMs that can interact in a human-like fashion, social influence can be shaped to fulfill private agendas, and due concern has been raised on LLMs' capability to effectively manipulate public opinion through misinformation [3, 27]. We thus look at whether different ways in which the LLM-generated content is delivered would be more impactful.

3 Experiment Design

To investigate the effects of LLM-driven adversarial social influences on people's online information processing, we conducted a pre-registered[1], randomized human-subject experiment on Prolific.

[1]The pre-registration document can be found at https://aspredicted.org/98yb-bnct.pdf. The experiment is approved by the IRB of the author's institution.



(a) An example of the news story and the social influences.

(b) The interface to make new comments, report the news veracity judgment, and intention to share.

Figure 1: The interface used in the experiment. Fig. 1a shows an example news story with a headline and an image. A comment section is below where the social influence is presented. Participants see only one of the three conditions of the comments section depending on the treatment. Fig. 1b is shows the part of the interface for participants to make new comments on the news and report their news veracity judgment and intention to share the news

3.1 Experiment Tasks

Participants were recruited to review news stories using a forum-based web application (Figure 1). In each task, a random piece of news from our dataset was presented to the participant. The participant was asked to carefully review the news and the opinions of other participants who reviewed it before them. They could voluntarily choose to reply to existing comments or add new comments to express their opinions. Then, the participant was asked to make a binary judgment on the veracity of the news (Real/Fake). Finally, they indicated how likely they would share the news through their social media accounts on a 7-point Likert scale between 1 (very unlikely to share) to 7 (very likely to share).

Each news was drawn from a dataset of 30 pieces of political news, where 17 pieces were real and fact-based, and the rest contained false information. The veracity of real news was confirmed by cross-checking with reliable media outlets, while fake news was either disputed by authoritative sources (e.g., fact-checking sites) or conflicted with verified information. Through a pilot study to assess our news dataset, we found that people's independent accuracy in determining the veracity of each news in our dataset was mostly between 50% and 70%, suggesting that people have difficulty in determining the actual veracity of the news on such news. As a result, they may naturally be influenced by social influences, which fits the purpose of our experiment well.

3.2 Experiment Treatments

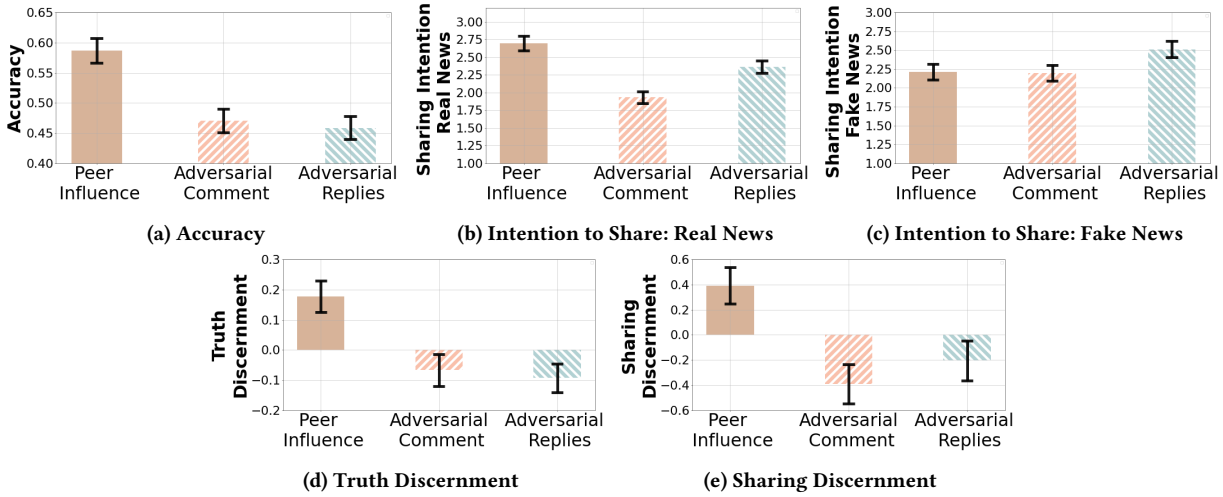
By varying the composition and presentation of social influences, we created three treatments in our experiment.

- **Control** (Peer Social Influence): The social influence consists of three comments on the news story, all generated by human participants.
- **Treatment 1** (Adversarial Comments): The social influence consists of six comments on the news story, including three human-generated comments and three comments generated by LLM-powered bots. The bot-generated comments express opinions opposing the true veracity of the news.
- **Treatment 2** (Adversarial Replies): The social influence consists of three human-generated comments and three bot-generated replies to these comments. The bot-generated replies either refute or agree with the human-generated comments while expressing opinions opposing the true veracity of the news.

The Peer Social Influence treatment was designed to simulate a scenario in which *no* bots are involved in people's news-reviewing procedure; people review the news and, at the same time, are potentially influenced by only other human users. Meanwhile, the other two treatments each pictured one potential way that bots could behave on the social media platform to create misleading social influence: to make news comments to the news story or to reply to existing comments. In both types of behaviors, the bots are prompted to explicitly express opinions contrary to the news stories' veracity. For instance, if the news is fake, the bot will argue that the news is actually real, and vice versa. Similarly, when the news is fake and there exists a comment arguing that the news is fake, the bot will reply to the comment and refute it, thus arguing that the news is actually real. Thus, the LLM-driven adversarial social influences intentionally seek to distort people's perceptions of the news. In practice, we used a pilot study asking participants to freely comment on the news stories to build the peer social influence, and used OpenAI's GPT-4o [30] to build a series of bots to create adversarial social influences.

3.3 Experiment Procedure

We posted our experiment as a crowdsourcing task on Prolific to U.S. workers only. Upon arrival, participants were randomly assigned to one of the three treatments. They were told to complete 10 tasks to review 10 pieces of news stories together with other Prolific workers. Before the formal experiment, participants also needed to



(a) Accuracy (b) Intention to Share: Real News (c) Intention to Share: Fake News

(d) Truth Discernment (e) Sharing Discernment

Figure 2: The impacts of adversarial social influence (i.e., adversarial comments and adversarial replies) on participants' ability to detect misinformation and their intention to spread information in the experiment. Error bars represent the standard errors of the mean.

complete: (1) a consent form, and (2) select an avatar and username to represent themselves in the experiment.

As participants entered the actual experiment, 12 tasks were selected for them to complete. To minimize the chance that participants were misled by the fake news they reviewed, we debriefed them on the ground truth veracity for each news after all the tasks were completed. To filter potential spammers, we also included an attention check question[2], and only those who passed were considered valid. The base payment for participating in the experiment was $1.2. To encourage participants to analyze the veracity of the news in each task carefully, they could earn a bonus of 5 cents for each correct judgment that they made if the accuracies of their final veracity judgments were over 65%. Thus, participants could receive a bonus payment up to $0.6.

We also had a post-survey to collect participants' demographic information. Specifically, we incorporated a series of measurements widely used in prior work that have been shown to impact people's information processing, especially in the context of partisan news reviewing, including the cognitive reflection test [14], misinformation susceptibility test [25], and the political standing survey (i.e., self-identified political leaning of Republican, Democrat, or Independent) [28]. A compensation of $0.5 was provided.

3.4 Analysis Methods

The main independent variable in our analysis is the experimental treatment that a participant was assigned to. To understand the effects of the presence and presentation format of adversarial social influence on people's perceptions of and willingness to engage with information, we pre-registered the following dependent variables: (1) the accuracy of a participant's judgment on the news veracity, and (2) truth discernment, which was calculated as a participant's frequency of labeling a piece of real news as "Real" minus the participant's frequency of labeling a piece of fake news as "Real" in their judgments. Unlike accuracy, this truth discernment metric captures how much more a participant believes true information relative to false information and, therefore, reflects the participant's sensitivity to distinguish true and false information. Both are well-adopted measurements in misinformation interventions research [16, 32].

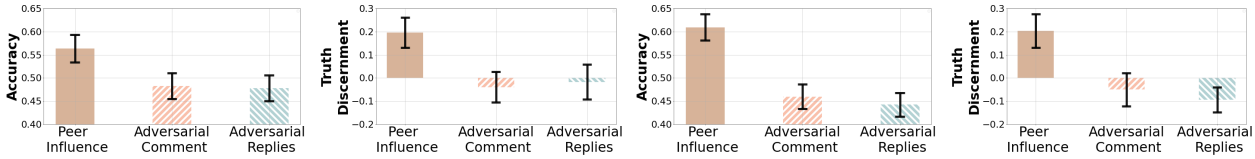
[2]In the attention check question, we asked participants to determine whether the statement "Washington DC is the capital city of the USA" is real or fake.

3.5 Results

In total, we collected data from 176 participants (48.3% self-identified as male, 44.9% self-identified as female, and the most frequent age group reported by subjects was 25-34). In what follows, we analyze the data obtained from the experiment to examine the effects of the LLM-driven adversarial social influences.

Effects of LLM-driven Adversarial Social Influences We begin by examining whether adversarial social influence affects individuals' ability to judge the veracity of news. Figures 2 compare participants' overall accuracy in judging news veracity and their ability to discern true versus false information. The figures clearly show that the presence of adversarial social influence reduces both people's accuracy in identifying misinformation and their sensitivity in differentiating true and false information. The ANOVA results confirm that such a decrease is significant for both people's veracity judgments accuracy and truth discernment (Accuracy: $F(2, 1933) = 12.28, p < 0.001$, Truth Discernment: $F(2, 175) = 8.13, p < 0.001$). Post-hoc Tukey HSD results further reveal that the effect of adversarial comments (Accuracy: $p < 0.001$, Cohen' $d = 0.24$, Truth Discernment: $p = 0.003$, Cohen' $d = 0.61$) and adversarial replies (Accuracy: $p < 0.001$, Cohen' $d = 0.26$, Truth Discernment: $p < 0.001$, Cohen' $d = 0.73$) both lead to significant decreases in people's capabilities in detecting misinformation compared to the control treatment.

Next, we explore participants' willingness to engage with the news. Our results in Figure 2 show that adversarial social influence



(a) Accuracy: Political Leaning Aligned **(b) Truth Discernment: Political Leaning Aligned** **(c) Accuracy: Political Leaning Conflicted** **(d) Truth Discernment: Political Leaning Conflicted**

Figure 3: The impacts of adversarial social influences (i.e., adversarial comments and adversarial replies) on participants' ability to detect misinformation when their political leanings are aligned or conflicted with the news content. Error bars represent the standard errors of the mean.

primarily reduces participants' intentions to share real news, while its effects on fake news are limited. The one-way ANOVA results confirm the significant differences in people's intention to share real news across the treatments ($F(2, 1072) = 17.09, p < 0.001$); meanwhile, adversarial information did not make people change their intention to share fake news ($p > 0.05$). Furthermore, the post-hoc analysis results show that adversarial replies decrease people's intention to share the news compared with those in the control treatment ($p = 0.03$, Cohen' $d = 0.19$). Furthermore, adversarial comments have a significant impact on people's engagement with the news. Specifically, it makes people less willing to share real news, compared to both the control treatment ($p < 0.001$, Cohen' $d = 0.44$) and the people under adversarial replies ($p = 0.02$, Cohen' $d = 0.26$). Due to such a reduction of willingness to share real news, we also find a significant difference in participants' tendency to share real news over fake news, as measured by the sharing discernment ($F(2, 175) = 6.61, p = 0.002$). The results show that when participants are affected by adversarial influences, they possess a significantly lower sharing discernment than those in the control treatment (Adversarial Comments: $p = 0.002$, Cohen's $d = 0.68$, Adversarial Replies: $p = 0.02$, Cohen's $d = 0.52$).

Overall, adversarial social influence significantly impacts the spread of information, though the effects vary depending on how it is presented. Both adversarial comments and replies reduce participants' ability to detect misinformation and distinguish true news from false news. Importantly, despite both adversarial comments and replies making people less discerning in engagement with real and fake news, adversarial comments have a more substantial negative effect on decreasing people's tendency to share real news.

How Do Adversarial Social Influences Affect People's Capabilities to Detect Misinformation? Our findings indicate that both forms of adversarial social influence negatively impact individuals' abilities to detect misinformation and appropriately share information. As we previously noted, people could possess confirmation bias when judging information veracity. Such biases in this context of reviewing partisan news could refer to their own political leaning or whether their political leaning is aligned with the news content. Although our results did not show that people's own political leaning is strong enough to resist adversarial social influences, we considered the possibility that political leanings might moderate the effects of these influences.

To explore this, we divided the data into two subgroups based on the alignment between the participants' political leanings and the political leanings of the news stories they reviewed. This alignment was measured using a partisan alignment value, calculated as the proportion of survey questions in which participants favored the political party aligned with the news item's leaning, divided by the total number of political stance questions. We then divide the data according to whether the alignment value is greater than the median alignment value of the full data ($Med_{alignment} = 0.5$), thus obtaining two subgroups, Aligned and Conflicted. We then repeat the analysis to examine how adversarial social influences affect people's capability to detect misinformation in the two conditions, respectively.

As shown in Figure 3, despite the trend, when people's political leaning is aligned with the news's leaning (i.e., the Aligned group), their capability to detect misinformation is not significantly affected by the adversarial social influence (Accuracy: $p > 0.05$). Though an overall significant difference is detected in people's truth discernment across three treatments ($p = 0.041$), post-hoc analysis indicates no significant differences between any pair of treatments in this group ($p > 0.05$). However, different patterns emerge in the Conflicted group, where participants' political leaning is different from the news leaning, we found that both types of adversarial social influences significantly harm people's capability of detecting misinformation, both in terms of accuracy ($p < 0.001$) and truth discernment ($p = 0.0048$). The post-hoc analysis further reveals that both adversarial social influences create a significant reduction in people's capability to detect misinformation, both in terms of accuracy (Adversarial Comments: $p < 0.001$, Adversarial Replies: $p < 0.001$) and truth discernment (Adversarial Comments: $p = 0.02$, Adversarial Replies: $p = 0.006$).

These findings suggest that the impact of adversarial social influences in reducing individuals' ability to detect misinformation is more prominent when there is a conflict between people's political leanings and the political leaning of the news content in the context of reviewing partisan news. A plausible explanation is that, in such cases, adversarial social influences actually align with the people's political stance, making them appear more credible and convincing. Therefore, it indicates that such adversarial influences may exploit the cognitive dissonance or uncertainty arising from conflicting leanings, leading to a greater susceptibility to misinformation.

Moderating Effects of Individual Factors As we previously revealed, certain individual differences, exemplified by political alignment, could probe into how adversarial social influence affects people's ability to detect misinformation. To formally investigate

Idp Variable/Dep Variable		y = Final Accuracy	
Intercept (β_0)		0.32	
Ad Comment (β_1)	-0.76^{**}	Ad Comment: Political Alignment (β_5)	0.59
Ad Replies (β_2)	-0.87^{**}	Ad Comment: CRT (β_6)	-0.01
Political Alignment (β_3)	-0.53	Ad Replies: Political Alignment (β_7)	0.79^{*}
CRT (β_4)	0.18^{*}	Ad Replies: CRT (β_8)	-0.02

Table 1: The regression for understanding how the political alignment between participants and news content, and participants' cognitive reflection test results moderates the impacts of adversarial social influences on people's veracity judgment accuracy. *,**,*** represent significance levels of 0.05, 0.01 and 0.001.

the potential moderating effects of such contextual factors in information processing, we conducted a moderation analysis. This analysis involves a logistic regression, using participants' misinformation detection accuracy (i.e., correct or not) as the dependent variable and the experimental treatments as independent variables, moderated by the individual factors. In addition to the political alignment discussed in the previous section, we incorporated participants' scores on CRT.

As shown in Table 1, we can first confirm the negative impacts of adversarial social influences on people's misinformation detection accuracy ($\beta_1 < 0, \beta_2 < 0$). Political alignment moderates these effects, making adversarial influences less effective when participants' political leanings align with the news and more effective when leanings differ ($\beta_5 > 0, \beta_7 > 0$). Such moderation is statistically significant when adversarial social influences are introduced by bots replying to existing comments ($p = 0.041$). Additionally, the CRT results reveal a main effect, showing that individuals with a stronger tendency for analytical thinking are better at distinguishing real from fake news. Notably, CRT scores do not interact with adversarial social influences ($\beta_6, \beta_8, p > 0.05$), indicating that maintaining analytical thinking is consistently effective in helping to detect misinformation even under adversarial social influences.

4 Directions for Future Work

Finding that social influence by bots affect people's views of news stories, we propose possible future directions.

Bot Indicators Drawing from a body of work on accuracy prompts that shift people's attention towards being accurate in assessing online content [33] and literacy interventions that aim to improve people's knowledge of online media [17], we propose placing a page that describes what LLMs are capable of, such as fluent natural language and extensive world knowledge. Another possible intervention borrows the concept of credibility indicators [24, 42], where instead of automatically labeling the veracity of content, the focus is on whether or not the content was detected to be LLM-written. These interventions aim to raise the users' caution towards the presence of content generated by bots.

Adaptive Intervention based on Contextual Information Our moderation analysis shows that adversarial social influences show prominent effects when people have different political leanings than the news. Conversely, when people have the same political leaning as the news, they possess better capabilities to detect misinformation. Malicious actors may exploit such polarized scenarios to deploy adversarial social influences. In response, it could be effective to implement adaptive credibility indicators that dynamically activate in contexts of conflicting political leanings, aiming to reduce the influence of adversarial social influences in polarized environments.

Improve Misinformation Resistance When added as a covariate to the regression analyses, we found that the incorporation of MIST measures [25] (i.e., capabilities to judge news veracity without social influence) also impacts the ability to detect fake news. People who were less susceptible to misinformation were also more resistant towards social influence in the comments. The literature points towards various factors that make people more discriminate of news. This includes having greater numeracy skills and stronger analytical reasoning, among others [39, 41]. Corroborating these works, our study suggests that educational programs to develop these skills would be beneficial in scenarios of judging the veracity of content under social influence.

Acknowledgments

We are grateful to the anonymous reviewers who provided many helpful comments. We thank the support of the National Science Foundation under grant IIS-2229876 and IIS-2340209 on this work. Any opinions, findings, conclusions, or recommendations expressed here are those of the authors alone.

References

[1] Hunt Allcott and Matthew Gentzkow. 2017. Social media and fake news in the 2016 election. *Journal of economic perspectives* 31, 2 (2017), 211–36.

[2] Dennis Assenmacher, Lena Clever, Lena Frischlich, Thorsten Quandt, Heike Trautmann, and Christian Grimme. 2020. Demystifying Social Bots: On the Intelligence of Automated Social Media Actors. *Social Media + Society* 6, 3 (01 Jul 2020), 2056305120939264. doi:10.1177/2056305120939264

[3] Dipto Barman, Ziyi Guo, and Owen Conlan. 2024. The Dark Side of Language Models: Exploring the Potential of LLMs in Multimedia Disinformation Generation and Dissemination. *Machine Learning with Applications* 16 (2024), 100545. doi:10.1016/j.mlwa.2024.100545

[4] Canyu Chen and Kai Shu. 2023. Can llm-generated misinformation be detected? *arXiv preprint arXiv:2309.13788* (2023).

[5] Canyu Chen and Kai Shu. 2024. Combating misinformation in the age of LLMs: Opportunities and challenges. *AI Magazine* 45, 3 (2024), 354–368. doi:10.1002/aaai.12188 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/aaai.12188

[6] Chun-Wei Chiang, Zhuoran Lu, Zhuoyan Li, and Ming Yin. 2023. Are two heads better than one in ai-assisted decision making? comparing the behavior and performance of groups and individuals in human-ai collaborative recidivism risk assessment. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–18.

[7] Kon Shing Kenneth Chung, Liaquat Hossain, and Joseph Davis. 2007. Individual performance in knowledge intensive work through social networks. In *Proceedings of the 2007 ACM SIGMIS CPR conference on Computer personnel research: The global information technology workforce*. 159–167.

[8] Jonas Colliander. 2019. "This is fake news": Investigating the role of conformity to other users' views when commenting on and spreading disinformation in social media. *Computers in Human Behavior* 97 (2019), 202–215.

[9] Nicholas Dufour, Arkanath Pathak, Pouya Samangouei, Nikki Hariri, Shashi Deshetti, Andrew Dudfield, Christopher Guess, Pablo Hernández Escayola, Bobby Tran, Mevan Babakar, and Christoph Bregler. 2024. AMMeBa: A Large-Scale Survey and Dataset of Media-Based Misinformation In-The-Wild. arXiv:2405.11697

[10] Melissa Dunne. 2023. Data Dive: Fake news in the age of AI. Retrieved June 19, 2024 from https://www.ipsos.com/en-us/data-dive-fake-news-age-ai

[11] Ullrich K. H. Ecker, Stephan Lewandowsky, John Cook, Philipp Schmid, Lisa K. Fazio, Nadia Brashier, Panayiota Kendeou, Emily K. Vraga, and Michelle A. Amazeen. 2022. The psychological drivers of misinformation belief and its resistance to correction. *Nature Reviews Psychology* 1, 1 (01 Jan 2022), 13–29. doi:10.1038/s44159-021-00006-y

[12] Emilio Ferrara, Onur Varol, Clayton Davis, Filippo Menczer, and Alessandro Flammini. 2016. The rise of social bots. *Commun. ACM* 59, 7 (jun 2016), 96–104. doi:10.1145/2818717

[13] Foreign, Commonwealth & Development Office, The Rt Hon Elizabeth Truss, and The Rt Hon Nadine Dorries. 2022. UK exposes sick Russian troll factory plaguing social media with Kremlin propaganda. Retrieved December 1, 2024 from https://www.gov.uk/government/news/uk-exposes-sick-russian-troll-factory-plaguing-social-media-with-kremlin-propaganda

[14] Shane Frederick. 2005. Cognitive Reflection and Decision Making. *Journal of Economic Perspectives* 19, 4 (December 2005), 25–42. doi:10.1257/089533005775196732

[15] Josh A. Goldstein, Renee DiResta, Girish Sastry, Micah Musser, Matthew Gentzel, and Katerina Sedova. 2023. Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations. Retrieved June 19, 2024 from https://fsi.stanford.edu/publication/generative-language-models-and-automated-influence-operations-emerging-threats-and

[16] Brian Guay, Adam J. Berinsky, Gordon Pennycook, and David Rand. 2023. How to think about whether misinformation interventions work. *Nature Human Behaviour* 7, 8 (01 Aug 2023), 1231–1233. doi:10.1038/s41562-023-01667-w

[17] Guanxiong Huang, Wufan Jia, and Wenting Yu. 2024. Media Literacy Interventions Improve Resilience to Misinformation: A Meta-Analytic Investigation of Overall Effect and Moderating Factors. *Communication Research* (04 Oct 2024), 00936502241288103. doi:10.1177/00936502241288103

[18] Silvia Knobloch-Westerwick, Cornelia Mothes, and Nick Polavin. 2020. Confirmation Bias, Ingroup Bias, and Negativity Bias in Selective Exposure to Political Information. *Communication Research* 47, 1 (01 Feb 2020), 104–124. doi:10.1177/0093650217719596

[19] Nicole M Krause, Isabelle Freiling, Becca Beets, and Dominique Brossard. 2020. Fact-checking as risk communication: the multi-layered risk of misinformation in times of COVID-19. *Journal of Risk Research* 23, 7-8 (2020), 1052–1059.

[20] 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.

[21] Lindsey C. Levitan and Brad Verhulst. 2016. Conformity in Groups: The Effects of Others' Views on Expressed Attitudes and Attitude Change. *Political Behavior* 38, 2 (01 Jun 2016), 277–315. doi:10.1007/s11109-015-9312-x

[22] Zhuoyan Li, Hangxiao Zhu, Zhuoran Lu, and Ming Yin. 2023. Synthetic data generation with large language models for text classification: Potential and limitations. *arXiv preprint arXiv:2310.07849* (2023).

[23] Lin Long, Rui Wang, Ruixuan Xiao, Junbo Zhao, Xiao Ding, Gang Chen, and Haobo Wang. 2024. On llms-driven synthetic data generation, curation, and evaluation: A survey. *arXiv preprint arXiv:2406.15126* (2024).

[24] Zhuoran Lu, Patrick Li, Weilong Wang, and Ming Yin. 2022. The Effects of AI-based Credibility Indicators on the Detection and Spread of Misinformation under Social Influence. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2, Article 461 (Nov. 2022), 27 pages. doi:10.1145/3555562

[25] Rakoen Maertens, Friedrich M. Götz, Hudson F. Golino, Jon Roozenbeek, Claudia R. Schneider, Yara Kyrychenko, John R. Kerr, Stefan Stieger, William P. McClanahan, Karly Drabot, James He, and Sander van der Linden. 2024. The Misinformation Susceptibility Test (MIST): A psychometrically validated measure of news veracity discernment. *Behavior Research Methods* 56, 3 (01 Mar 2024), 1863–1899. doi:10.3758/s13428-023-02124-2

[26] Miller McPherson, Lynn Smith-Lovin, and James M Cook. 2001. Birds of a Feather: Homophily in Social Networks. *Annual Review of Sociology* 27, 1 (2001), 415–444. doi:10.1146/annurev.soc.27.1.415 arXiv:https://doi.org/10.1146/annurev.soc.27.1.415

[27] Raphael Meier. 2024. *LLM-Aided Social Media Influence Operations.* Springer Nature Switzerland, Cham, 105–112. doi:10.1007/978-3-031-54827-7_11

[28] Amy Mitchell, Jeffrey Gottfried, Galen Stocking, Mason Walker, and Sophia Fedeli. 2019. Many Americans say made-up news is a critical problem that needs to be fixed. *Pew Research Center* 5 (2019), 2019.

[29] Mehdi Moussaïd, Juliane E Kämmer, Pantelis P Analytis, and Hansjörg Neth. 2013. Social influence and the collective dynamics of opinion formation. *PloS one* 8, 11 (2013), e78433.

[30] OpenAI. 2023. ChatGPT. https://openai.com/chatgpt. Version 4.0.

[31] Mariam Orabi, Djedjiga Mouheb, Zaher Al Aghbari, and Ibrahim Kamel. 2020. Detection of Bots in Social Media: A Systematic Review. *Information Processing & Management* 57, 4 (2020), 102250. doi:10.1016/j.ipm.2020.102250

[32] Gordon Pennycook, Jabin Binnendyk, Christie Newton, and David G. Rand. 2021. A Practical Guide to Doing Behavioral Research on Fake News and Misinformation. *Collabra: Psychology* 7, 1 (07 2021), 25293. doi:10.1525/collabra.25293 arXiv:https://online.ucpress.edu/collabra/article-pdf/7/1/25293/470915/collabra_2021_7_1_25293.pdf

[33] Gordon Pennycook and David G. Rand. 2022. Accuracy prompts are a replicable and generalizable approach for reducing the spread of misinformation. *Nature Communications* 13, 1 (28 Apr 2022), 2333. doi:10.1038/s41467-022-30073-5

[34] Tom Postmes, S Alexander Haslam, and Roderick I Swaab. 2005. Social influence in small groups: An interactive model of social identity formation. *European review of social psychology* 16, 1 (2005), 1–42.

[35] David N Rapp and Nikita A Salovich. 2018. Can't we just disregard fake news? The consequences of exposure to inaccurate information. *Policy Insights from the Behavioral and Brain Sciences* 5, 2 (2018), 232–239.

[36] Tate Ryan-Mosley. 2023. How generative AI is boosting the spread of disinformation and propaganda. Retrieved June 19, 2024 from https://www.technologyreview.com/2023/10/04/1080801/generative-ai-boosting-disinformation-and-propaganda-freedom-house/

[37] Chengcheng Shao, Giovanni Luca Ciampaglia, Onur Varol, Kai-Cheng Yang, Alessandro Flammini, and Filippo Menczer. 2018. The spread of low-credibility content by social bots. *Nature Communications* 9, 1 (20 Nov 2018), 4787. doi:10.1038/s41467-018-06930-7

[38] Massimo Stella, Emilio Ferrara, and Manlio De Domenico. 2018. Bots increase exposure to negative and inflammatory content in online social systems. *Proceedings of the National Academy of Sciences* 115, 49 (2018), 12435–12440. doi:10.1073/pnas.1803470115 arXiv:https://www.pnas.org/doi/pdf/10.1073/pnas.1803470115

[39] Mubashir Sultan, Alan N. Tump, Nina Ehmann, Philipp Lorenz-Spreen, Ralph Hertwig, Anton Gollwitzer, and Ralf H. J. M. Kurvers. 2024. Susceptibility to online misinformation: A systematic meta-analysis of demographic and psychological factors. *Proceedings of the National Academy of Sciences* 121, 47 (2024), e2409329121. doi:10.1073/pnas.2409329121 arXiv:https://www.pnas.org/doi/pdf/10.1073/pnas.2409329121

[40] Joshua Aaron Tucker, Andrew Guess, Pablo Barbera, Cristian Vaccari, Alexandra Siegel, Alexandra Siegel, Sergey Sanovich, Denis Stukal, and Brendan Nyhan. 2018. Social Media, Political Polarization, and Political Disinformation: A Review of the Scientific Literature. Retrieved October 10, 2022 from https://ssrn.com/abstract=3144139

[41] Sander van der Linden. 2022. Misinformation: susceptibility, spread, and interventions to immunize the public. *Nature Medicine* 28, 3 (01 Mar 2022), 460–467. doi:10.1038/s41591-022-01713-6

[42] Waheeb Yaqub, Otari Kakhidze, Morgan L. Brockman, Nasir Memon, and Sameer Patil. 2020. Effects of Credibility Indicators on Social Media News Sharing Intent. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–14. doi:10.1145/3313831.3376213