International Journal of Communication 19(2025), 3712–3735 1932–8036/20250005

Echoes of Doubt: Exposure to Information About Generative AI Decreases Believability of News

MARINA TULIN[1]
MYRTO PANTAZI
CHRISTOPHER STARKE
MICHAEL SIVOLAP
TOM DOBBER
University of Amsterdam, The Netherlands

The emergence of generative artificial intelligence (GenAI) has sparked a debate about its potential misuse for creating political disinformation. However, the effects of providing information about generative AI on disinformation perceptions remain unclear. We fill this gap by testing the impact of GenAI literacy interventions on truth discrimination (i.e., the ability to accurately distinguish between genuine and false online news) and deception bias (i.e., the tendency to believe that online news is false) in an online experiment among 897 Canadian adults. Respondents were randomly assigned to a GenAI literacy intervention (explainer videos), showing how ChatGPT and Midjourney can be used to create political disinformation vs. art. The GenAI interventions increased participants' propensity to classify online news as false, yet signal detection analyses showed no improvement in truth discrimination. In addition, we find evidence for a deception bias where participants have a slight tendency to judge online news as false rather than true. We conclude that GenAI literacy interventions need to be carefully crafted to avoid further undermining the believability of genuine news.

Keywords: generative AI, disinformation, misinformation, truth bias, deception bias, media literacy, journalism

Marina Tulin: m.tulin@uva.nl
Myrto Pantazi: pantazi@uva.nl
Christopher Starke: c.d.r.o.starke@uva.nl
Michael Sivolap: m.sivolap@uva.nl
Tom Dobber: t.dobber@uva.nl
Date submitted: 2025-02-20

[1] Acknowledgement: We wish to thank Christofer Talvitie and Marta Maggioni for their support in selecting the stimuli for this study and programming the initial draft of the Qualtrics survey.

The recent proliferation of generative artificial intelligence (GenAI) tools has raised concerns about potential misuse, like creating highly realistic, deceptive content. Such concerns are echoed by the World Economic Forum, which named AI-generated political disinformation the top global risk for 2024. Disinformation has grave consequences for democracy, including declining shared truth and amplifying relativism, fueling polarization, and impeding political decision-making (Bennett & Livingston, 2018; Van Aelst et al., 2017). While the news media widely warn about the disruptive potential of GenAI for democracy, art blogs celebrate the creative possibilities of AI image generators. Similar dissent exists in academic debate. While some researchers warn about the sophisticated deception capabilities of GenAI (Shoaib, Wang, Ahvanooey, & Zhao, 2023), others argue that such fears are overblown because AI is unlikely to increase the quantity, quality, or personalization of disinformation (Simon, Altay, & Mercier, 2023). Recent evidence nuances this by showing that GenAI has increased the scale of certain propaganda efforts (see Wack, Ehrett, Linvill, & Warren, 2025). In this study, we conceptualize disinformation as a perceptual crisis, focusing not on its direct consequences but on the perceptions it raises among citizens (e.g., Hameleers & Brosius, 2022). Online users are exposed to misinformation less often than commonly assumed (Acerbi, Altay, & Mercier, 2022), but public concerns around the exposure to false or misleading information are relatively high (Hameleers et al., 2023; van der Meer & Hameleers, 2024). Such heightened disinformation perceptions may be further exacerbated by GenAI, which may have negative consequences for political participation and trust in democratic institutions (Ognyanova, Lazer, Robertson, & Wilson, 2020).

This article aims to contribute to the ongoing debate on GenAI and its potential use in political disinformation by studying the extent to which GenAI literacy interventions may help citizens accurately distinguish between genuine and false online news—an ability we define as truth discrimination (Batailler, Brannon, Teas, & Gawronski, 2022).[2] In addition, we study citizens' general bias toward believing or disbelieving online news and the potential unintended consequences of GenAI interventions to increase deception bias—the tendency to believe that online information is false. Previous research indicates that providing information on deception techniques can have diverging effects on disinformation perceptions. On the one hand, it could prepare individuals for future deception attempts and increase their ability to spot them (e.g., Traberg, Roozenbeek, & van der Linden, 2022). On the other hand, such warnings can backfire, causing individuals to misclassify true information as false (Guess et al., 2020), even if they correctly identify false information. In fact, some of the most widely used interventions against misinformation can increase skepticism alongside their positive effect in reducing misinformed beliefs (Hoes, Aitken, Zhang, Gackowski, & Wojcieszak, 2024). Yet, it remains unclear how receiving information about the (deceptive) capabilities of GenAI may affect disinformation perceptions. Shedding light on this open research question, this study asks: What are the effects of GenAI literacy interventions on truth discrimination and deception bias? We conducted a pre-registered online experiment with 897 Canadian adults to answer this question.

[2] Truth discrimination is conceptually synonymous to truth discernment (e.g., Epstein, Sirlin, Arechar, Pennycook, & Rand, 2023).

Generative AI in Times of Information Disorder

The current information ecosystem is polluted by harmful information, falsehoods, and half-truths (Freelon & Wells, 2020; Vraga & Bode, 2020; Wardle, 2019). Conceptually, we distinguish between misinformation and disinformation (Wardle & Derakhshan, 2017). Misinformation refers to false information that lacks expertise or supporting evidence, but is not intentionally false (Vraga & Bode, 2020). Disinformation specifically refers to *intentionally* misleading or deceptive information that is fabricated with the intention to harm or to achieve strategic objectives (Chadwick & Stanyer, 2022; Freelon & Wells, 2020). The commonly used term "fake news" can be considered a subcategory of disinformation because it is false information intentionally created to look like news (Egelhofer & Lecheler, 2019). As such, fake news leverages the credibility markers of the news media and benefits from the trust people typically assign to professional journalistic products (Lazer et al., 2018). In this article, we focus specifically on AI-generated fake news, which is political disinformation created with GenAI that imitates the format of a news headline and image.

This study investigates how citizens assess the veracity of AI-generated fake news after receiving a GenAI literacy intervention explaining its functionalities and purposes. To clarify how GenAI literacy interventions shape citizens' perceptions of (fake) news amid information disorder, we draw on truth default theory and theories of truth discrimination. A central question is to what extent GenAI literacy interventions increase citizens' ability to discern genuine from fake news (truth discrimination), or whether it leaves them doubting the accuracy of all information (deception bias).

Generative AI and Truth Discrimination

Truth discrimination, in this context, refers to the ability to accurately distinguish between true and false information. Successful truth discrimination means accurately identifying false information as false and true information as true. High truth discrimination is highly desirable because it means that citizens do not blindly trust or distrust, but they calibrate trust and distrust with information accuracy.

A useful distinction in the realm of misinformation interventions is that of reactive vs. preemptive interventions. Reactive interventions refer to practices that counter misinformation *after* it has already entered the information landscape. Preemptive interventions are preventive practices that are implemented *before* misinformation reaches its audience. Fact-checking and debunking are examples of reactive practices because they react to misinformation already in circulation, including specific false claims, narratives, images, and audio-visual content. Such reactive interventions often seek to inform citizens about the falsehoods contained in this content, explain why it is false, and provide corrective information (Graves, 2017; Graves & Amazeen, 2019). Fact-checks and debunks are generally effective at lowering the credibility of misinformation (e.g., see meta-analysis: Walter, Cohen, Holbert, & Morag, 2020). However, they are unable to fully correct false beliefs, as a 'continued influence effect' remains after receiving corrections (Lewandowsky, Ecker, Seifert, Schwarz, & Cook, 2012; van Huijstee, Vermeulen, Kerkhof, & Droog, 2022; Walter & Tukachinsky, 2020). Preemptive interventions prepare citizens for misinformation before they encounter it (Bruns et al., 2024; Hoes et al., 2024; Walter & Tukachinsky, 2020), such as forewarnings that alert citizens to misinformation they may come across. An alleged advantage is that falsehoods never fully

enter citizens' minds unchecked. This helps prevent the continued influence of misinformation, as it does not fully take root or connect to existing attitudes, memories, and worldviews (Lewandowsky & van der Linden, 2021).

Preemptive interventions may differ in the specific strategies they use to build resilience against future misinformation attempts, such as strengthening media navigation skills, raising awareness of persuasion attempts, or teaching about specific persuasion techniques. Four main types of preemptive interventions exist. Forewarnings alert the recipient about the existence of misinformation, thus increasing their vigilance towards incoming information (Clayton et al., 2020). Similarly, accuracy nudges may help target people's attention towards accuracy concerns, making people less prone to believe the news they read, including fake ones (Pennycook & Rand, 2022). Inoculation techniques combine: 1) a warning against misinformation that supposedly motivates individuals to protect their existing beliefs and 2) specific tips on how to spot and counterargue deception techniques typically employed in mis- and disinformation (Lewandowsky & van der Linden, 2021). In practice, inoculation typically exposes individuals to a small dose of a persuasion attempt to inform them about the underlying techniques used (Roozenbeek, van der Linden, & Nygren, 2020; Traberg et al., 2022). Lastly, media literacy interventions focus on providing individuals with specific tips and decision rules to spot misinformation (Guess et al., 2020). Thus, despite their differences, preemptive interventions generally warn participants about prospective encounters with mis- or disinformation and provide them with guidance on how to spot it.

Finally, an emerging literature focuses on trust-enhancing interventions, which are motivated by the recognition that 1) misinformation is relatively rare (e.g., Acerbi et al., 2022), and 2) misinformation interventions often inadvertently undermine trust in reliable information (Hoes et al., 2024). Indeed, a recent meta-analysis highlights that citizens are better at detecting false news as false than they are at detecting true news as true, suggesting that there is a need to increase trust in reliable news (Pfänder & Altay, 2025). Trust-enhancing interventions therefore aim to enhance confidence in reliable news by addressing either the news production side, like reducing bias and increasing in-depth reporting (Fisher, Flew, Park, Lee, & Dulleck, 2021), or by reassuring news users that most news stories are reliable, especially when headlines are informative and credible (Altay, De Angelis, & Hoes, 2024).

This study speaks to the literature on preemptive misinformation interventions by testing whether viewing a video about the capabilities and dangers of GenAI before exposure to possible AI-generated falsehoods affects citizens' ability to detect them. Specifically, the videos show how different GenAI tools can be used to create different kinds of content, including political disinformation. We focus on this kind of forewarning because it mirrors those commonly presented in news media, which warn about GenAI's potential to disrupt elections (Corera & Wheeler, 2024) and its abilities to create photorealistic depictions of events that never occurred (Verma, 2023). Like the forewarnings in such news media coverage, our intervention does not include concrete tips on how to spot AI-generated content, as inoculation or media literacy interventions might. Our intervention can thus be categorized as a forewarning about the threat of AI-generated false information online, alerting citizens to future persuasion attempts. A 2018 meta-analysis found forewarnings to be effective, though less so than debunking (Walter & Murphy, 2018). More recent work, however, finds forewarnings to be more effective than debunks (Buczel, Siwiak, Szpitalak, & Polczyk,

2024). Because forewarnings have not yet been studied in the context of GenAI, we seek to understand how such a simple and widely used misinformation intervention affects citizens' misinformation perceptions.

An important question when assessing the efficacy of misinformation interventions is what constitutes a desirable outcome (Blair et al., 2024). A meta-analysis comparing various interventions against misinformation found that forewarnings are somewhat effective at lowering misinformed beliefs, albeit not as effective as debunking (Walter & Murphy, 2018). While forewarnings and other preemptive strategies decrease the credibility of fake news, they can also inadvertently reduce the credibility of genuine news (e.g., Clayton et al., 2020; Hoes et al., 2024). Some studies show that people become better at distinguishing genuine news from fake news after receiving a forewarning. However, this improvement happens because the warning reduces the perceived credibility of both fake and genuine news, though the credibility of genuine news decreases less strongly than that of fake news (e.g., Guess et al., 2020). Conceptualizing such an intervention as effective risks further fueling already increased levels of distrust in legacy news media and journalism (Fisher et al., 2021). In this study, therefore, we explicitly conceptualize the efficacy of our GenAI literacy intervention in terms of its effect on truth discrimination. We consider the intervention effective if individuals are more successful at identifying genuine news as genuine and fake news as fake. Unlike past studies that relied on discernment, we use the discrimination index from signal detection theory to properly address this question. While discernment captures how much more truthful people perceive true information to be than fake information, discrimination is a more objective measure of how accurately people can classify true and fake news in their respective categories. As such, it is a more definitive measure of how well people can detect fake news. The discrimination index ranges from 0 (no distinguishing ability) to positive values, with higher values indicating higher ability to distinguish between true and fake news (Batailler et al., 2022). Based on theoretical arguments from the previous literature overview, as well as mixed findings on the efficacy of forewarnings, we ask:

RQ1: What is the effect of GenAI literacy interventions (i.e., forewarnings) on truth discrimination?

Generative AI and Deception Bias

Well-intended misinformation interventions like forewarnings may have undesirable side effects, namely undermining the credibility of all news, fake or genuine (Hoes et al., 2024). We consider the possibility that GenAI literacy interventions may suffer from similar drawbacks. We use the concept of truth default as a starting point for understanding the detrimental effects of forewarnings (Levine, 2014). Generally, people tend to passively accept information as true without immediate skepticism (Pantazi, Kissine, & Klein, 2018), reflecting a broader tendency to take mental shortcuts and preserve cognitive energy (Metzger & Flanagin, 2013). In everyday interactions, people do not spontaneously question the veracity of the information they receive, so unprompted truth assessments rarely occur (less than 10% of the time), and a trigger is needed to prompt people to consider the veracity of information (Clare & Levine, 2019). That said, triggers as simple as asking someone to assess the accuracy of a statement, as forewarnings do, may be sufficient to activate suspicion (Levine, 2022). The inherent inclination to believe rather than doubt information is both adaptive and rational, given that misinformation constitutes only a fraction of the information people encounter online (Acerbi et al., 2022). The truth default, however, may

make people vulnerable to deception attempts in those rare cases where they do encounter false information and uncritically accept it as true (Levine, 2022).

Our study spotlights GenAI as a new informational threat that has the potential to undermine the truth-default in the context of online information. In the mis- and disinformation literature, GenAI has been addressed with mixed reactions, ranging from warnings of its sophisticated deception capabilities (Shoaib et al., 2023) to statements that such fears are overblown (Simon et al., 2023). Within the framework of the truth default theory, people may move away from the truth default when suspicion is triggered (Levine, 2022). We conceptualize warnings about GenAI's capabilities and potential misuse as triggers that can activate veracity considerations, leading people to abandon the truth default.

While the impact of GenAI warnings on the truth default has not yet been studied, it is evident that such warnings are increasingly prevalent in public discourse. Large-scale analyses of circa 70,000 news reports on AI find a high level of fear-mongering and "AI phobia" in news headlines (Samuel, Khanna, & Sundar, 2024). Especially in 2024, when more than two billion voters in 50 countries were asked to cast a ballot, warnings about the disruptive potential of GenAI were particularly prominent (e.g., Corera & Wheeler, 2024). As mentioned above, research on preemptive interventions reveals unintended effects: while people become more accurate at identifying false information, they may also mistakenly judge true information as false (Guess et al., 2020; Hoes et al., 2024). Widespread doubt of genuine news is harmful to democracy if it undermines citizens' ability to form political attitudes or leads to disengagement from politics.

Warnings of the possibilities of GenAI may, thus, contribute to the erosion of truth default in online information. While many people may have heard about AI, only a minority has substantial knowledge about it, and most people are unable to accurately recognize examples of AI in daily life (Faverio & Tyson, 2023). Based on the truth default framework, we expect that receiving a literacy intervention warning about GenAI's capabilities may trigger suspicion and leave citizens doubting the veracity of information. We therefore hypothesize that:

H1: Individuals are more likely to rate online news as false after being exposed to a GenAI literacy intervention.

To gauge the extent to which participants may resort to truth-default or deception bias, we used the bias index from signal detection theory (see Stanislaw & Todorov, 1999). Negative values suggest a bias towards responding "fake" (deception bias), while positive values indicate a bias towards responding "true" (truth bias).

Context-Specific Differences in Using Generative AI

GenAI may be used for different purposes. Our main interest is in studying the consequences of warning about AI-generated disinformation, as these warnings are increasingly highlighted in the news media. We extend our investigation to another area where GenAI is making waves, namely the creative business (Epstein, Hertzmann et al., 2023), for two reasons: First, to enhance external validity, we sought to reflect common real-world applications of GenAI, with art being a prominent example. Numerous art

blogs were promoting the use of GenAI during its initial rise, recommending prompts to generate beautiful visuals, song lyrics, and poetry, enticing users to get creative and generate their own art. Second, art serves as a relatively neutral use case, enabling us to disentangle the effects of simply learning about GenAI's capabilities from the influence of warnings about its risks. By presenting GenAI's artistic applications, we offer a less charged context for evaluating citizens' perceptions, in contrast to the more negative implications associated with its use for creating political disinformation.

Learning about GenAI's capabilities in art creation is less likely to trigger suspicion than learning about its capabilities for creating political disinformation. Thus, awareness of GenAI's role in art might not increase people's skepticism. However, simply learning about its potential to create photo-realistic images and persuasive text may be sufficient to make citizens wonder whether it could be misused (Epstein, Hertzmann et al., 2023). Truth default theory argues that individuals do not question the veracity of information until they experience a trigger event that raises suspicion. The art intervention may serve as such a trigger, bringing to mind questions of content authenticity. Especially in a time when mis- and disinformation concerns are relatively high among citizens (Hameleers et al., 2023; van der Meer & Hameleers, 2024), even a subtle trigger may lead individuals to retrieve memories of fabricated content, previous warnings, and associated concerns. It is, therefore, possible that the phenomenon and mere knowledge of the existence of AI-generated art could have negative consequences on citizens' perceptions of online content.

Since art is generally not assessed for veracity, but AI-generated art still showcases GenAI's capabilities—which may instill doubt—it can serve as a benchmark for understanding both the undermining effect of learning about GenAI and the knock-on effects of warnings about AI-generated disinformation. We therefore hypothesize:

H2: The negative effect of GenAI literacy interventions on truth ratings of online news is stronger for individuals who learn how GenAI tools are used to generate disinformation compared to individuals who learn how GenAI tools are used to create art.

Methods

Sample

To test our research question and hypotheses, we conducted a pre-registered online experiment (see anonymized pre-registration link: https://osf.io/zkaeb/?view_only=e4b786f8046e4eaf8fceeabe79e2509e) among Canadian adults (see Table A1 in the online Appendix for all deviations from the pre-registration via https://osf.io/c632m/files/osfstorage/688b73878748f9a51e130015). We used quotas for age, gender, and education to match the Canadian population. The data were collected between November 16 and November 23, 2023, through an online panel administered by Dynata. The ethics committee of the University of Amsterdam approved the study design under project number FMG-3027. A total of 1,220 respondents gave informed consent and completed the questionnaire. After excluding 42 respondents for speeding and 271 for failing the attention check, our final sample consisted of N = 897 (median completion

time: 8 minutes). Our sample closely resembles the Canadian population in terms of gender, age, and education (see Table 1).

Table 1. Sample Description.

	N	Quota achieved (%)	Quota set (%)
Gender			
Woman	461	.51	.51
Man	431	.48	.49
Non-binary	5	.01	n/a
Age			
18–34 years	246	.28	.29
35–54 years	281	.32	.32
55 years or older	363	.41	.40
Education			
Lower	572	.64	.65
Higher	325	.36	.35

Note. Lower education includes starting or completing elementary or secondary education, high school, technical or community college. Higher education includes starting or completing any university education, that is, Bachelor's, Master's, or PhD.

Materials & Measurements

GenAI Literacy Intervention

Respondents were randomly assigned to one of three conditions: experimental (disinformation or art) vs. control condition (no video). In the experimental conditions, participants watched a circa 2-minute video specifically created for this study. It combined publicly available royalty-free videos with original screen recordings showing prompts being entered into ChatGPT and Midjourney and fragments of their output. No GenAI was used in producing the video interventions, aside from these screen-recordings. A human Canadian narrator delivered the audio in a neutral, calm tone. The short format was chosen for scalability and compatibility with various social media platforms, enhancing ecological validity. The intervention aligns with video platforms like YouTube, the most-used video-based platform in Canada (Mai & Gruzd, 2022), and is also suitable for Instagram (up to 3-minute reels) and TikTok (up to 10-minute videos). YouTube videos are easily sharable across major social media platforms (e.g., Facebook, X, BlueSky) and instant messaging apps (e.g., WhatsApp, Telegram, Signal).

The videos featured the University of Amsterdam logo at the beginning and the end. Because the videos were embedded in Qualtrics from the university's YouTube account, participants also saw the logo in the bottom-left corner of the screen.

In terms of content, the video in the "Disinformation" condition (2 min 11 sec) informed about the possibilities of creating political disinformation by using GenAI, namely ChatGPT for text and Midjourney for images. Specifically, respondents were shown how to use GenAI to create a fake news

article and photo-realistic image of Russian president Vladimir Putin being convicted of war crimes.[3] Respondents in the "Art" condition were exposed to a short video (1 min 54 sec) explaining how GenAI can be used to create art. Specifically, they learned how to use GenAI to create a poem about a car made of potatoes and a matching image.

The intervention resembles brief YouTube explainer videos. To mitigate ethical concerns, the videos are not a step-by-step GenAI tutorial. We deliberately omitted critical steps needed to create the final product, for example, specific prompts and workarounds required to get ChatGPT to generate a fake news article. Similarly, we omitted several steps taken in Midjourney to create the image. Thus, viewers would not be able to reproduce a fake news article themselves. Both videos are available here: https://drive.proton.me/urls/G0CW4EWYJG#t5EgJruBh6bc.

After watching the video, respondents answered the open question: "What do you think of the video you just saw?". Their responses suggest no suspicion that the videos were AI-generated. In the disinformation condition, a minority of participants seemed to be unbothered by the capabilities of GenAI discussed in the video. The majority reported that they felt scared after watching the video, and that they did not know "what to believe anymore." As for the art condition, most participants said that "this did not trigger any thoughts" or that they were not worried. A minority of the participants, however, also used language such as "scary," "very concerning," or "it really made [them] think." We also observed some rare positive responses in both conditions that pointed to fascination with this new technology.

Truth Ratings

Our key dependent variables, namely *Discrimination* and *Bias* (see analytical strategy), are based on respondents' truth ratings of eight news items, each consisting of a headline and an associated image. Four news items were genuine news containing accurate information (see Figure 1 for an example), and four were generated via AI tools and contained false information (see Figure 2 for an example). We conducted a pre-test (n = 53) for 19 news items (10 fake, 9 real). The aim was to select stimuli that would reduce floor or ceiling effects in the main study (e.g., very high agreement that the news item is false prior to manipulation may lead to floor effects). The items were presented in random order, and participants were asked to judge the credibility of the news item via a binary choice: "Did the depicted event ever occur?" (yes, no). We interpret a "yes" as believing that the news depicting the event is genuine and a "no" as believing that the news is fake. Respondents in the main study responded to 8 stimuli based on the same question as in the pre-test. Averaging across all conditions in the main study, participants judged the news to be genuine 46% of the time.

[3] Participants saw a shortened version of this process. In reality, creating the fake news article required several prompts to circumvent the tools' security layers which otherwise prevent (or at least discourage) users from creating disinformation. For ethical reasons we did not display the exact prompts to circumvent these security layers.

Belarus' leader helicopters over Minsk with a rifle as protesters below demand his resignation

Figure 1. Example of a real image.

Thousands of Italians gather to the shore to welcome refugee boats as a demonstration against government's tighter anti-immigration policies

Figure 2. Example of a fake image.

Analytical Strategy

To address RQ1, we followed recent recommendations (Batailler et al., 2022) and used Signal Detection Theory to test participants' ability to discriminate between fake and genuine news, while taking into account response biases. To this end, we calculated the *Discrimination* (*d'*) and *Bias* (*c*) indexes at the individual level. If *d'* equals 0, then participants are unable to distinguish between fake and genuine news; positive values suggest an ability to distinguish. This index allowed us to test whether the intervention improved participants' ability to discriminate between genuine and fake news.

As part of this analytical approach, we also calculated the *c* bias index, which shows whether participants exhibit a truth or deception bias. A *c* of 0 means absence of bias, while negative values indicate a bias towards responding "fake" and positive values indicate a bias towards responding "true" (see Stanislaw & Todorov, 1999). To maximize power and overcome problems of incomputability due to extreme Hit or False Alarm rates, we corrected values of one or zero by .01 (see Kane, Conway, Mirua, & Colflesh, 2007).

Findings of the Signal Detection analyses show whether a truth or deception bias is present, but they do not provide an immediate test that compares the experimental conditions and the control condition, as described in H1 or H2. For a stringent test of H1 and H2, therefore, we analyzed the raw truth-ratings at the item level through a logistic regression model with news type ("True" vs. "Fake"), condition ("Control" vs. "Art" vs. "Disinformation"), and their interaction as predictors. This allowed us to determine whether our interventions increased "fake" ratings (H1) and whether the intervention in the disinformation condition led to more "fake" ratings than in the art condition (H2).

Results

Signal Detection Theory analyses

As Figure 1 suggests, participants varied greatly in their ability to discriminate (*d'*) and in their response bias (*c*). We performed t-tests to determine whether the discrimination and bias indexes across our entire sample significantly differed from the zero value, reflecting indiscriminability and absence of bias, respectively. The mean discrimination index of the overall sample was just above 0, $M = .15$, 95%CI[.07, .24], $t(894) = 3.44$, $p < .001$, suggesting that participants' ability to discriminate overall was slightly but significantly higher than chance. Concerning bias, the entire sample exhibited a very small average bias towards responding "fake," $M = -.16$, 95%CI[−.22, −.10], $t(894) = -5.06$, $p < .001$. That said, there was significant variation, suggesting that whereas several participants exhibited a deception bias (i.e., tended to respond "fake" instead of "true"), others exhibited a truth default one (i.e., tended to respond "true" instead of "fake").

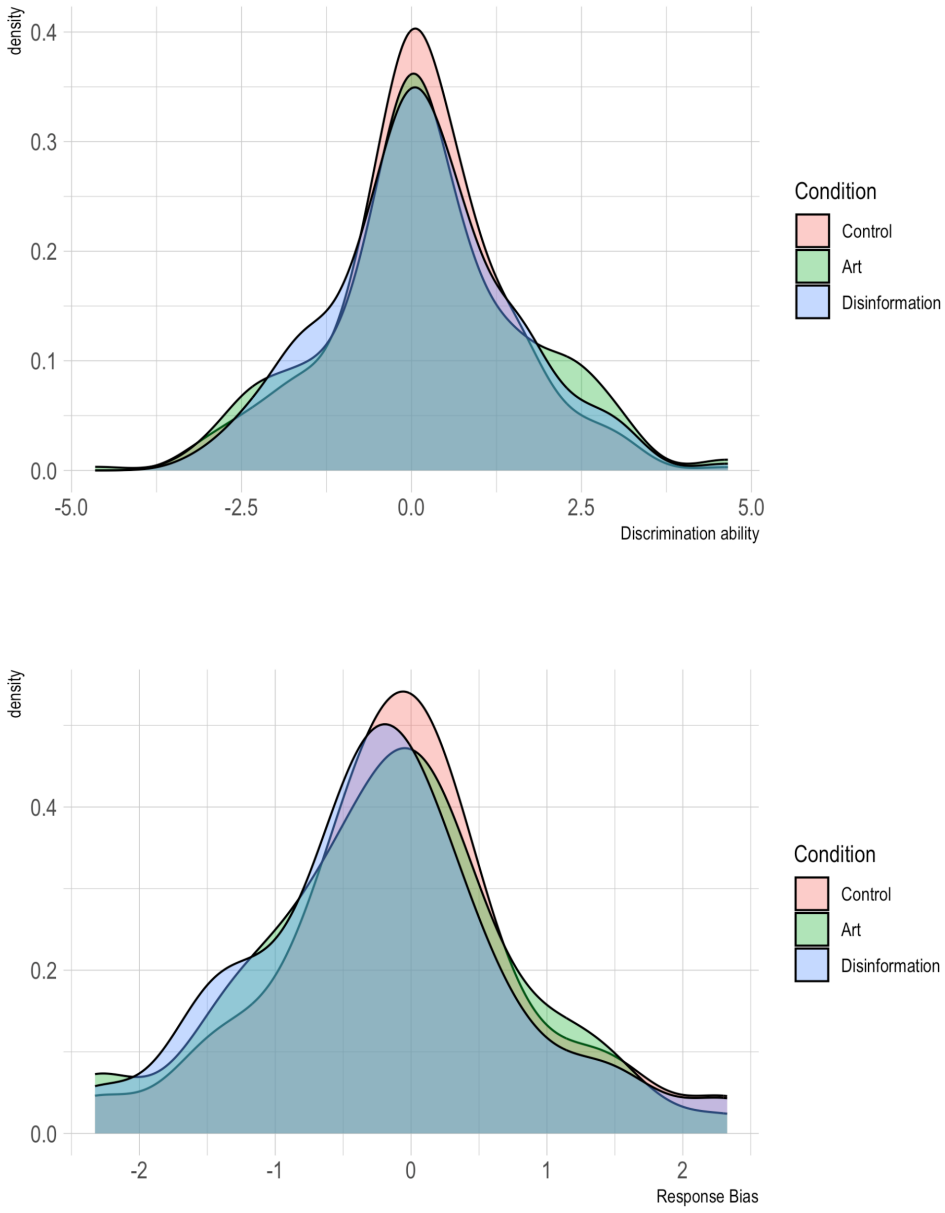

Figure 1. Density plots of the Discrimination (top) and Bias indexes (bottom) per group.

Table 2 presents the mean bias and discrimination indexes per group. Two one-way ANOVAs did not suggest any difference in either the bias or the discrimination indexes across the experimental conditions

(see Table 2). This suggests that the slight increase in the "fake responses" in the experimental conditions probably resulted in small changes in both discrimination and the default response bias across individuals, but did not reflect a unique intervention effect on either of the two signal detection indexes.

Table 2. Mean Sensitivity and Bias Measures and 95% CIs Per Condition.

	Control	Art	Disinformation	F Statistic
Discrimination (d)	.12 [−.02, .25]	.52 [.03, .38]	.14 [−.01, .30]	$F(2, 892) = .319$, $p = .727$
Bias (c)	−.07 [−.18, .03]	−.19 [−.31, −.08]	−.22 [−.33, −.11]	$F(2, 892) = 2.15$, $p = .117$

Truth Ratings

To test H1 and H2, we submitted the truth rating variable to a logistic regression (*glm* command of the "stats" package in R) with news type ("Genuine" vs. "Fake"), condition ("Control" vs. "Art" vs. "Disinformation"), and their interaction as predictor variables. The main effect of Condition was significant, $\chi 2$ (2, 7164) = 7.56, $p = .023$. This effect was because, in line with H1, participants in the "Disinformation" condition rated more news as "fake" than the Control group participants, OR = 1.16, $SE = .05$, $p = .019$ (see Figure 1). Unlike the "Disinformation" group, the "Art" group did not significantly rate more news as fake compared to the control group, OR = 1.11, $SE = .05$, $p = .183$, thus partially supporting H2, although the "Art" group did not differ significantly from the Disinformation group in terms of the amount of news they identified as fake, OR = .95, $SE = .06$, $p = .657$. The main effect of news type was also significant, $\chi 2$ (1, 7164) = 14.95, $p = .001$, suggesting that participants were more likely to rate AI-generated news as fake than genuine news, OR = 1.2, $SE = .06$, $p < .001$. Lastly, the News Type X Group interaction was not significant, $\chi 2$ (2, 7164) = .398, $p = .819$, suggesting that the effect of the intervention was not restricted to fake news but spilled over to genuine news as well. Figure 2 shows the probability of rating news as false or true depending on the news type and the GenAI literacy intervention received.

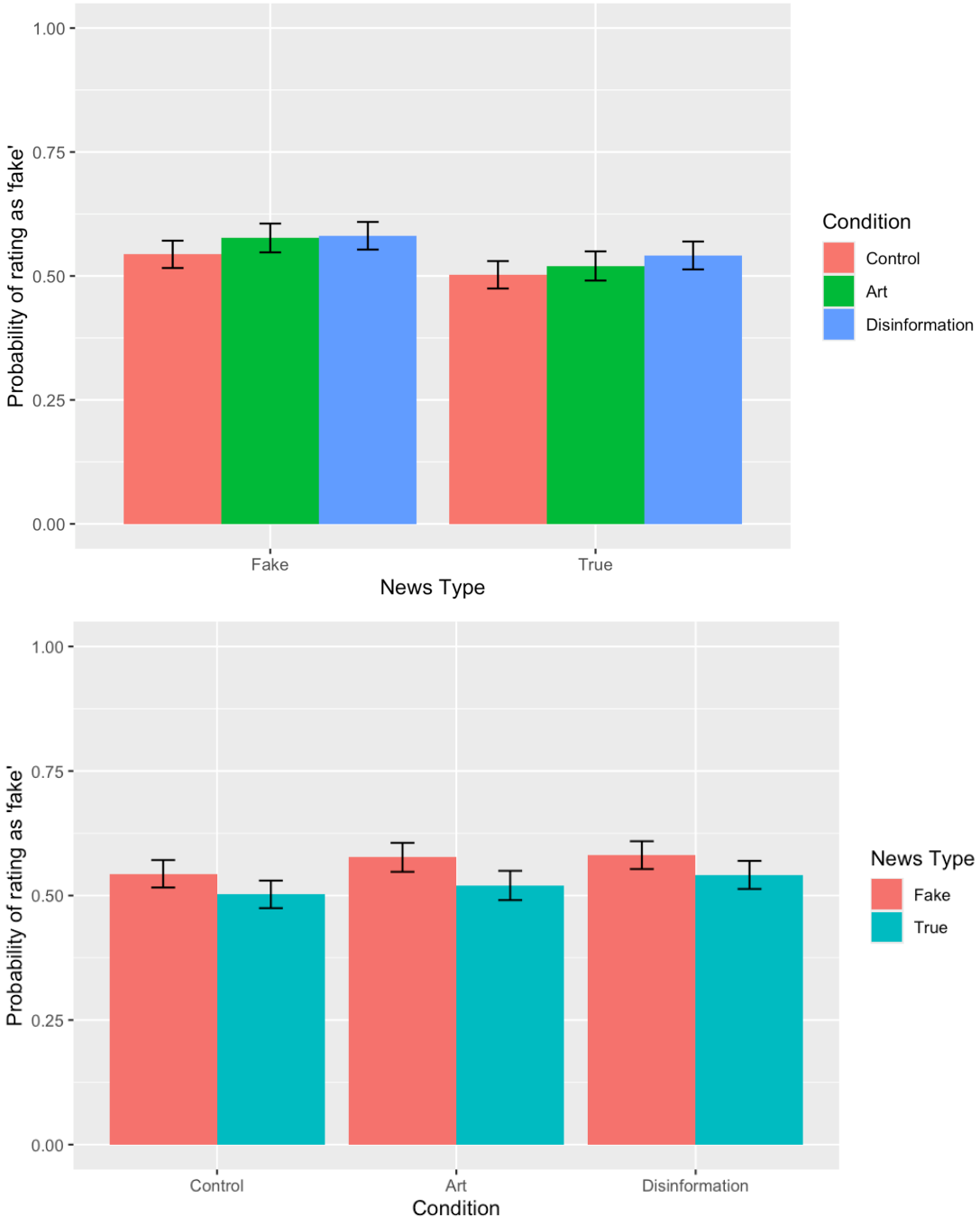

Figure 2. Probability of rating the news as "fake" as a function of Condition (top panel) and News type (bottom panel).

As a robustness check, we also ran post-hoc analyses to test whether the effects of the manipulation might depend on participants' previous familiarity with the GenAI technology, which we had also measured in the survey. We thus added to the above logistic regression self-reported familiarity with GenAI (Yes vs. No) and its interaction with all other terms, namely news type ("Genuine" vs. "Fake") and condition ("Control" vs. "Art" vs. "Disinformation"). Aside from confirming the previously reported effects of condition $\chi 2$ (2, 6960) = 10.31, p = .006, and news type, $\chi 2$ (2, 6960) = 15.53, p < .001, this new analysis only revealed a main effect of Familiarity, $\chi 2$ (2, 6960) = 35.34, p < .001, reflecting that people not previously familiar with GenAI had a significantly higher probability of rating news as fake, OR = 1.33, *SE* = .06, p < .001, compared to people who were previously familiar with GenAI. Notably, none of the interactions involving familiarity were significant.[4]

Discussion

This study aimed to provide insights into the effects of a GenAI literacy intervention that served as a forewarning informing citizens about the capabilities and potential misuse of GenAI for creating political disinformation. The intervention served as a forewarning, similar to the warnings that citizens typically receive via the news media (e.g., White, 2024). We tested predictions based on the literature on preemptive misinformation interventions and truth default theory. Specifically, we investigated whether the GenAI literacy intervention increased citizens' ability to accurately distinguish between genuine and false online news (truth discrimination) and whether it undermined their general tendency to believe that information is true (deception bias).

Our findings partially align with our predictions. Regarding RQ1, the GenAI literacy intervention did not increase truth discrimination, although participants exhibited a small, above-chance discrimination ability. In addition, we observe a small deception bias, where citizens are more likely to state that news is fake rather than true. In addition, citizens were more likely to judge online news as fake following exposure to the GenAI literacy intervention, thus supporting H1. We find this for both genuine and fake news, which suggests that while the GenAI literacy intervention can decrease the credibility of fake news, it can also inadvertently decrease the credibility of genuine news. We find partial support for H2 that participants shown how GenAI can be used to create political disinformation would show a stronger tendency to classify news as fake than those shown how GenAI can be used to create art. While participants in the disinformation group were more likely to rate news as false compared to the control group, those in the art condition did not differ significantly from the control condition.

Our study has implications for three research strands: truth default theory (Levine, 2022), mis- and disinformation, and literacy interventions. First, with changes in technology and rising concerns about false information and deception attempts (Van Aelst et al., 2017), the general tendency to trust information may be challenged in online spaces. Our findings spotlight GenAI as a new technology with the potential to undermine the truth default. Citizens are highly concerned about the informational threats of mis- and

[4] NewsType x Condition: χ^2 (2, 6960) = .397, p = .820; NewsType x Familiarity: χ^2 (1, 6960) = 1.39, p = .237; Condition x Familiarity: χ^2 (2, 6960) = 2.50, p = .285; NewsType x Condition x Familiarity: χ^2 (2, 6960) = .729, p = .694.

disinformation (van der Meer & Hameleers, 2024) and are beginning to display a deception bias when judging news shared on social media (Luo, Hancock, & Markowitz, 2022). Rather than assuming that online information is true, awareness of GenAI leaves citizens with echoes of doubt about the veracity of information. Reminders and warnings about the possibilities of GenAI may contribute to an erosion of trust in online information. Our study contributes to the literature on the truth default theory and foreshadows a possible shift from general trust and passive acceptance of information to skepticism due to the potential for AI-generated content. A concrete theoretical implication is that we may need to reconsider whether the truth bias—long assumed to be the default state—should still be considered the starting point for how humans perceive information, particularly in an online landscape riddled with warnings about mis- and disinformation. A shift towards deception bias, where citizens default to suspicion, is harmful if citizens doubt information that is actually true because the functioning of democracy depends on a well-informed citizenry.

A deception bias may lead to more wrong judgments than a truth bias because—despite warnings and fears—studies estimate the vast majority of information in the average news diet to be accurate (Acerbi et al., 2022). That said, the amount of false information likely varies by platform and topic, raising the question of when skepticism is appropriate. Higher skepticism may be desirable in contexts of uncertainty and high levels of mis- and disinformation, such as during crises (Talvitie, Hameleers, Tulin, & de Vreese, 2023). Recent research also highlights the importance of motivations underlying skepticism, with accuracy-driven motivations linked to healthy skepticism (Li, 2025), which encourages verification behaviors like reading fact-checks (Tulin et al., 2024). In general, behavioral extensions of skepticism offer a promising avenue for future research. If (healthy) skepticism promotes verification behaviors, it may offset concerns about deviating from the truth default.

Another nuance is that while our study focused on a binary true/false distinction, grey areas of truthfulness matter, especially in the context of GenAI, which can be used to illustrate real or plausible events. We accounted for this by asking, "Did the depicted event ever occur?" rather than asking whether the image or text is AI-generated. As such, our measure allows for the possibility that the (potentially AI-generated image) depicts a true event. Future work could explore these grey areas of partial truthfulness, which are likely to occur in the information ecology.

Our findings show that even individuals who were not exposed to the GenAI literacy intervention displayed a small deception bias, which may be the result of repeated exposure to warnings about GenAI and mis- or disinformation online. Since we conducted a single-shot experiment, we were unable to examine the potential effects of repeated exposure, which is a limitation.

In addition, the static environment of a survey experiment may limit our ability to generalize our findings to the dynamic context of social media platforms. Innovative data collection methods, like data donations or researcher access to online platform data under the Digital Services Act, may allow researchers to go back in time and reconstruct how often citizens encountered warnings about GenAI and mis- or disinformation. The role of sources, contexts, or stated purpose of interventions could be studied to better understand the underlying mechanisms and efficacy of GenAI interventions in the wild. A

longitudinal perspective would allow us to model how media diets may, over time, turn a truth bias into a deception bias.

Second, our study also speaks to the literature on mis- and disinformation as a perceptual crisis. While online users are exposed to misinformation less often than commonly assumed (Acerbi et al., 2022), public concerns about exposure to false or misleading information are relatively high (Hameleers et al., 2023; van der Meer & Hameleers, 2024). Our study shows that disinformation perceptions are further fueled by knowledge of GenAI's capabilities, even though fears that AI will increase the quantity, quality, or personalization of disinformation are possibly overblown (Simon et al., 2023). Irrespective of whether GenAI will actually be used to create more, higher-quality, or more personalized disinformation, it has secondary effects via increased skepticism towards news due to GenAI's mere possibilities.

Third, our study also engages with the literature on misinformation interventions (for meta-analyses, see Walter & Murphy, 2018; Walter & Tukachinsky, 2020). Even though preemptive measures like forewarnings are well-documented strategies against the harms of mis- and disinformation (Guess et al., 2020; Lu et al., 2023; Roozenbeek et al., 2020; Traberg et al., 2022), some studies find no effects on discernment (Modirrousta-Galian & Higham, 2023) or negative side effects (Hoes et al., 2024). Our findings contribute additional evidence to this literature by showing that generic forewarnings about GenAI, similar to those employed in news media reports, are not only ineffective in actually improving discrimination between genuine and fake news but can also be counterproductive by eroding the credibility of all information. Rather than specifically preparing citizens against disinformation, they raise general suspicion about online news. Furthermore, our intervention showed the inputs and outputs of GenAI tools, which may have raised some problematic curiosity among some participants, even though none of them explicitly perceived the intervention as instructional. Instead, educating citizens about GenAI's capabilities made them more skeptical. This correlates with the broader misinformation literature documenting increased skepticism following misinformation interventions (e.g., Hoes et al., 2024). Related research suggests that media literacy interventions need to be targeted depending on levels of distrust (Hameleers & van der Meer, 2023). Our findings likewise highlight the need for targeted interventions that strike a balance between reaping the benefits of AI literacy and mitigating the possible risks of news cynicism.

The study used a negative and potentially extreme example of disinformation, which might raise questions about the representativeness of this type of content within real-world literacy interventions. However, we argue that the stimulus depicting a fabricated trial of Vladimir Putin is ecologically valid since similar AI-generated images of Vladimir Putin and other prominent politicians circulated online in 2023 (Lajka & Marcelo, 2023). Compared to other more sensational stimuli in previous work, such as deepfakes of politicians accusing immigrants of violent crimes like rape (Hameleers, 2024), Christian politicians joking about Christ's suffering (Dobber et al., 2021), Nancy Pelosi supporting the violent acts of capitol hill rioters (Hameleers et al., 2024), or politicians requiring doctors to treat cancer patients with essential oils (Ternovski, Kalla, & Aronow, 2022), our stimulus is relatively moderate and grounded in actual disinformation trends. While some degree of extremity may be necessary to convey risk, we maintain that our example remains appropriate and realistic. Still, findings based on a single topic that may be perceived as politically charged (i.e., Putin's arrest) remain limited in their generalizability.

The findings of this study can also have real-world implications. For instance, journalists are facing great pressure to produce high-quality news content at an incredible pace (Caswell & Dörr, 2019). News outlets have started using GenAI to assist with news production (Opdahl et al., 2023). As our study shows, news users are already suspicious of the veracity of news, and learning that GenAI can be used for news and disinformation may further lower their trust in reliable news content. In addition, news users seem to fail to distinguish between true and false news content, which puts an additional responsibility on news producers to inform them when GenAI is being used, as will soon be required by legal frameworks like the EU AI Act or China's Labeling Measures for Content Generated by Artificial Intelligence. Journalists should critically weigh the benefits and potential harms of using GenAI in news reporting, considering professional and ethical implications—an approach that recent research suggests they are already beginning to adopt (Cools & de Vreese, 2025).

Our findings also have serious consequences for democracy. A general deception bias toward all news undermines citizens' ability to form informed opinions about political issues, a prerequisite for a functioning democracy. Interventions such as forewarnings that primarily instill fear about misinformation, similar to the media response during the emergence of GenAI, are insufficient on their own. To effectively enhance discernment, interventions should offer concrete guidance on identifying AI-generated content (where feasible) and direct users to sources of reliable information (Hameleers, 2024). Emphasizing the accessibility and abundance of trustworthy content in professional news outlets, as trust-enhancing interventions do (Altay et al., 2024), may be instrumental in restoring public trust in political information.

References

Acerbi, A., Altay, S., & Mercier, H. (2022). Research note: Fighting misinformation or fighting for information? *Harvard Kennedy School Misinformation Review*. doi:10.37016/mr-2020-87

Altay, S., De Angelis, A., & Hoes, E. (2024). Media literacy tips promoting reliable news improve discernment and enhance trust in traditional media. *Communications Psychology, 2*(1), 1–9. doi:10.1038/s44271-024-00121-5

Batailler, C., Brannon, S. M., Teas, P. E., & Gawronski, B. (2022). A signal detection approach to understanding the identification of fake news. *Perspectives on Psychological Science, 17*(1), 78–98. doi:10.1177/1745691620986135

Bennett, W. L., & Livingston, S. (2018). The disinformation order: Disruptive communication and the decline of democratic institutions. *European Journal of Communication,* 2, 122–139. doi:10.1177/0267323118760317

Blair, R. A., Gottlieb, J., Nyhan, B., Paler, L., Argote, P., & Stainfield, C. J. (2024). Interventions to counter misinformation: Lessons from the Global North and applications to the Global South. *Current Opinion in Psychology, 55*, 101732. doi:10.1016/j.copsyc.2023.101732

Bruns, H., Dessart, F. J., Krawczyk, M., Lewandowsky, S., Pantazi, M., Pennycook, G., … Smillie, L. (2024). Investigating the role of source and source trust in prebunks and debunks of misinformation in online experiments across four EU countries. *Scientific Reports, 14*(1), Article 1. doi:10.1038/s41598-024-71599-6

Buczel, K. A., Siwiak, A., Szpitalak, M., & Polczyk, R. (2024). How do forewarnings and post-warnings affect misinformation reliance? The impact of warnings on the continued influence effect and belief regression. *Memory & Cognition, 52*(5), 1048–1064. doi:10.3758/s13421-024-01520-z

Caswell, D., & Dörr, K. (2019). Automating complex news stories by capturing news events as data. *Journalism Practice, 13*(8), 951–955. doi:10.1080/17512786.2019.1643251

Chadwick, A., & Stanyer, J. (2022). Deception as a bridging concept in the study of disinformation, misinformation, and misperceptions: Toward a holistic framework. *Communication Theory, 32*(1), 1–24. doi:10.1093/ct/qtab019

Clare, D. D., & Levine, T. R. (2019). Documenting the truth-default: The low frequency of spontaneous unprompted veracity assessments in deception detection. *Human Communication Research, 45*(3), 286–308. doi:10.1093/hcr/hqz001

Clayton, K., Blair, S., Busam, J. A., Forstner, S., Glance, J., Green, G., … Nyhan, B. (2020). Real solutions for fake news? Measuring the effectiveness of general warnings and fact-check tags in reducing belief in false stories on social media. *Political Behavior, 42*(2), 1073–1095. doi:10.1007/s11109-019-09533-0

Cools, H., & de Vreese, C. H. (2025). From automation to transformation with AI-tools: Exploring the professional norms and the perceptions of responsible AI in a news organization. *Digital Journalism*. doi:10.1080/21670811.2025.2505982

Corera, G., & Wheeler, B. (2024). AI could "supercharge" election disinformation, US tells the BBC. *BBC*. Retrieved from https://www.bbc.com/news/world-68295845

Egelhofer, J. L., & Lecheler, S. (2019). Fake news as a two-dimensional phenomenon: A framework and research agenda. *Annals of the International Communication Association, 43*(2), 97–116. doi:10.1080/23808985.2019.1602782

Epstein, Z., Hertzmann, A., the Investigators of Human Creativity, Akten, M., Farid, H., Fjeld, J., … Smith, A. (2023). Art and the science of generative AI. *Science, 380*(6650), 1110–1111. doi:10.1126/science.adh4451

Epstein, Z., Sirlin, N., Arechar, A., Pennycook, G., & Rand, D. (2023). The social media context interferes with truth discernment. *Science Advances, 9*(9), eabo6169. doi:10.1126/sciadv.abo6169

Faverio, M., & Tyson, A. (2023). *What the data says about Americans' views of artificial intelligence*. Pew Research Center. Retrieved from https://www.pewresearch.org/short-reads/2023/11/21/what-the-data-says-about-americans-views-of-artificial-intelligence/

Fisher, C., Flew, T., Park, S., Lee, J. Y., & Dulleck, U. (2021). Improving trust in news: Audience solutions. *Journalism Practice, 15*(10), 1497–1515. doi:10.1080/17512786.2020.1787859

Freelon, D., & Wells, C. (2020). Disinformation as political communication. *Political Communication, 37*(2), 145–156. doi:10.1080/10584609.2020.1723755

Graves, L. (2017). Anatomy of a fact check: Objective practice and the contested epistemology of fact checking. *Communication, Culture & Critique, 10*(3), 518–537. doi:10.1111/cccr.12163

Graves, L., & Amazeen, M. A. (2019). *Fact-checking as idea and practice in journalism*. Oxford Research Encyclopedia of Communication. Retrieved from https://oxfordre.com/communication/abstract/10.1093/acrefore/9780190228613.001.0001/acrefore-9780190228613-e-808

Guess, A. M., Lerner, M., Lyons, B., Montgomery, J. M., Nyhan, B., Reifler, J., & Sircar, N. (2020). A digital media literacy intervention increases discernment between mainstream and false news in the United States and India. *Proceedings of the National Academy of Sciences of the United States of America, 117*(27), 15536–15545. doi:10.1073/pnas.1920498117

Hameleers, M. (2024). Cheap versus deep manipulation: The effects of cheapfakes versus deepfakes in a political setting. *International Journal of Public Opinion Research*, *36*(1), 1–9. doi:10.1093/ijpor/edae004

Hameleers, M., & Brosius, A. (2022). You are wrong because I am right! The perceived causes and ideological biases of misinformation beliefs. *International Journal of Public Opinion Research, 34*(1), 1–12. doi:10.1093/ijpor/edab028

Hameleers, M., Tulin, M., de Vreese, C. H., Aalberg, T., Van Aelst, P., Cardenal, A. S., … Zoizner, A. (2023). Mistakenly misinformed or intentionally deceived? Mis- and disinformation perceptions on the Russian War in Ukraine among citizens in 19 countries. *European Journal of Political Research, 63*(4), 1642–1654. doi:10.1111/1475-6765.12646

Hameleers, M., & van der Meer, T. G. L. A. (2023). Striking the balance between fake and real: Under what conditions can media literacy messages that warn about misinformation maintain trust in accurate information? *Behaviour & Information Technology*, 1–13. Advance online publication. doi:10.1080/0144929X.2023.2267700

Hameleers, M., van der Meer, T. G. L. A., & Dobber, T. (2024). Distorting the truth versus blatant lies: The effects of different degrees of deception in domestic and foreign political deepfakes. *Computers in Human Behavior*, 152, 1–13. doi:10.1016/j.chb.2023.108096

Hoes, E., Aitken, B., Zhang, J., Gackowski, T., & Wojcieszak, M. (2024). Prominent misinformation interventions reduce misperceptions but increase scepticism. *Nature Human Behaviour, 8*(8), 1545–1553. doi:10.1038/s41562-024-01884-x

Kane, M. J., Conway, A. R., Miura, T. K., & Colflesh, G. J. (2007). Working memory, attention control, and the N-back task: A question of construct validity. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33*(3), 615–622. doi:10.1037/0278-7393.33.3.615

Lajka, A., & Marcelo, P. (2023). Fake AI images of Putin, Trump being arrested spread online. *PBS News*. Retrieved from https://www.pbs.org/newshour/politics/fake-ai-images-of-putin-trump-being-arrested-spread-online.

Lazer, D. M. J., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., … Zittrain, J. L. (2018). The science of fake news. *Science, 359*(6380), 1094–1096. doi:10.1126/science.aao2998

Levine, T. R. (2014). Truth-default theory (TDT): A theory of human deception and deception detection. *Journal of Language and Social Psychology, 33*(4), 378–392. doi:10.1177/0261927x14535916

Levine, T. R. (2022). Truth-default theory and the psychology of lying and deception detection. *Current Opinion in Psychology, 47*, 101380. doi:10.1016/j.copsyc.2022.101380

Lewandowsky, S., Ecker, U. K. H., Seifert, C. M., Schwarz, N., & Cook, J. (2012). Misinformation and its correction: Continued influence and successful debiasing. *Psychological Science in the Public Interest, 13*(3), 106–131. doi:10.1177/1529100612451018

Lewandowsky, S., & van der Linden, S. (2021). Countering misinformation and fake news through inoculation and prebunking. *European Review of Social Psychology, 32*(2), 348–384. doi:10.1080/10463283.2021.1876983

Li, J. (2025). Not all skepticism is "healthy" skepticism: Theorizing accuracy- and identity-motivated skepticism toward social media misinformation. *New Media & Society, 27*(1), 522–544. doi:10.1177/14614448231179941

Lu, C., Hu, B., Li, Q., Bi, C., & Ju, X.-D. (2023). Psychological inoculation for credibility assessment, sharing intention, and discernment of misinformation: Systematic review and meta-analysis. *Journal of Medical Internet Research, 25*, e49255. doi:10.2196/49255

Luo, M., Hancock, J. T., & Markowitz, D. M. (2022). Credibility perceptions and detection accuracy of fake news headlines on social media: Effects of truth-bias and endorsement cues. *Communication Research, 49*(2), 171–195. doi:10.1177/0093650220921321

Mai, P., & Gruzd, A. (2022). *The state of social media in Canada 2022.* Social Media Lab Toronto Metropolitan University. Retrieved from https://socialmedialab.ca/2022/09/14/survey-finds-canadians-are-spending-less-time-on-social-media-but-tiktok-is-the-exception/

Metzger, M. J., & Flanagin, A. J. (2013). Credibility and trust of information in online environments: The use of cognitive heuristics. *Journal of Pragmatics, 59*, 210–220. doi:10.1016/j.pragma.2013.07.012

Modirrousta-Galian, A., & Higham, P. A. (2023). Gamified inoculation interventions do not improve discrimination between true and fake news: Reanalyzing existing research with receiver operating characteristic analysis. *Journal of Experimental Psychology: General, 152*(9), 2411–2437. doi:10.1037/xge0001395

Ognyanova, K., Lazer, D., Robertson, R. E., & Wilson, C. (2020). Misinformation in action: Fake news exposure is linked to lower trust in media, higher trust in government when your side is in power. *Harvard Kennedy School Misinformation Review*. doi:10.37016/mr-2020-024

Opdahl, A. L., Tessem, B., Dang-Nguyen, D.-T., Motta, E., Setty, V., Throndsen, E., … Trattner, C. (2023). Trustworthy journalism through AI. *Data & Knowledge Engineering, 146*, 102182. doi:10.1016/j.datak.2023.102182

Pantazi, M., Kissine, M., & Klein, O. (2018). The power of the truth bias: False information affects memory and judgment even in the absence of distraction. *Social Cognition, 36*(2), 167–198. doi:10.1521/soco.2018.36.2.167

Pennycook, G., & Rand, D. G. (2022). Accuracy prompts are a replicable and generalizable approach for reducing the spread of misinformation. *Nature Communications, 13*(1), 1–12. doi:10.1038/s41467-022-30073-5

Roozenbeek, J., van der Linden, S., & Nygren, T. (2020). Prebunking interventions based on the psychological theory of "inoculation" can reduce susceptibility to misinformation across cultures. *Harvard Kennedy School Misinformation Review*. doi:10.37016//mr-2020-008

Samuel, J., Khanna, T., & Sundar, S. (2024). Fear of artificial intelligence? NLP, ML and LLMs based discovery of AI-phobia and fear sentiment propagation by AI news. *Computer Science and Mathematics*. Advance online publication. doi:10.20944/preprints202403.0704.v1

Shoaib, M. R., Wang, Z., Ahvanooey, M. T., & Zhao, J. (2023). Deepfakes, misinformation, and disinformation in the era of frontier AI, generative AI, and large AI Models. *2023 International Conference on Computer and Applications (ICCA)*, 1–7. doi:10.1109/ICCA59364.2023.10401723

Simon, F. M., Altay, S., & Mercier, H. (2023). Misinformation reloaded? Fears about the impact of generative AI on misinformation are overblown. *Harvard Kennedy School Misinformation Review*. doi:10.37016/mr-2020-127

Stanislaw, H., & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior Research Methods, Instruments, & Computers*, *31*(1), 137–149. doi:10.3758/bf03207704

Talvitie, C., Hameleers, M., Tulin, M., & de Vreese, C. (2023). *Finding truth amidst turmoil: Evidence-based recommendations for fact-checking in times of crisis*. Retrieved from https://benedmo.eu/wp-content/uploads/2023/12/Whitepaper_factchecking-crisis-times-PREV1_spread.pdf

Ternovski, J., Kalla, J., & Aronow, P. (2022). The negative consequences of informing voters about deepfakes: Evidence from two survey experiments. *Journal of Online Trust and Safety, 1*(2), Article 2. doi:10.54501/jots.v1i2.28

Traberg, C. S., Roozenbeek, J., & van der Linden, S. (2022). Psychological inoculation against misinformation: Current evidence and future directions. *The ANNALS of the American Academy of Political and Social Science, 700*(1), 136–151. doi:10.1177/00027162221087936

Tulin, M., Hameleers, M., de Vreese, C., Aalberg, T., Corbu, N., Van Erkel, P., … Theocharis, Y. (2024). Why do citizens choose to read fact-checks in the context of the Russian War in Ukraine? The role of directional and accuracy motivations in nineteen democracies. *The International Journal of Press/Politics, 30*(3), 679–704. doi:10.1177/19401612241233533

Van Aelst, P., Strömbäck, J., Aalberg, T., Esser, F., de Vreese, C., Matthes, J., … Stanyer, J. (2017). Political communication in a high-choice media environment: A challenge for democracy? *Annals of the International Communication Association, 41*(1), 3–27. doi:10.1080/23808985.2017.1288551

van der Meer, T. G. L. A., & Hameleers, M. (2024). Misinformation perceived as a bigger informational threat than negativity: A cross-country survey on challenges of the news environment. *Harvard Kennedy School Misinformation Review*. doi:10.37016/mr-2020-142

van Huijstee, D., Vermeulen, I., Kerkhof, P., & Droog, E. (2022). Continued influence of misinformation in times of COVID-19. *International Journal of Psychology, 57*(1), 136–145. doi:10.1002/ijop.12805

Verma, P. (2023). As AI-generated hands get more realistic, it'll be hard to spot the fakes. Washington Post. Retrieved from https://www.washingtonpost.com/technology/2023/03/26/ai-generated-hands-midjourney/

Vraga, E. K., & Bode, L. (2020). Defining misinformation and understanding its bounded nature: Using expertise and evidence for describing misinformation. *Political Communication, 37*(1), 136–144. doi:10.1080/10584609.2020.1716500

Wack, M., Ehrett, C., Linvill, D., & Warren, P. (2025). Generative propaganda: Evidence of AI's impact from a state-backed disinformation campaign. *PNAS Nexus, 4*(4), pgaf083. doi:10.1093/pnasnexus/pgaf083

Walter, N., Cohen, J., Holbert, R. L., & Morag, Y. (2020). Fact-checking: A meta-analysis of what works and for whom. *Political Communication, 37*(3), 350–375. doi:10.1080/10584609.2019.1668894

Walter, N., & Murphy, S. T. (2018). How to unring the bell: A meta-analytic approach to correction of misinformation. *Communication Monographs, 85*(3), 423–441. doi:10.1080/03637751.2018.1467564

Walter, N., & Tukachinsky, R. (2020). A meta-analytic examination of the continued influence of misinformation in the face of correction: How powerful is it, why does it happen, and how to stop it? *Communication Research, 47*(2), 155–177. doi:10.1177/0093650219854600

Wardle, C. (2019). A new world disorder. *Scientific American, 321*(3), 88–95. Retrieved from https://www.jstor.org/stable/27265334

Wardle, C., & Derakhshan, H. (2017). *Information disorder: Toward an interdisciplinary framework for research and policymaking*. Council of Europe. Retrieved from https://tverezo.info/wp-content/uploads/2017/11/PREMS-162317-GBR-2018-Report-desinformation-A4-BAT.pdf

White, J. (2024, May 20). See how easily A.I. chatbots can be taught to spew disinformation. *The New York Times*. Retrieved from https://www.nytimes.com/interactive/2024/05/19/technology/biased-ai-chatbots.html