Article

new media & society
1–21

DOI: 10.1177/14614448241286881
journals.sagepub.com/home/nms


Perceiving AI intervention does not compromise the persuasive effect of fact-checking

Je Hoon Chae
Department of Communication, University of California, Los Angeles, USA
Department of Statistics and Data Science, University of California, Los Angeles, USA

David Tewksbury
Department of Communication, University of Illinois Urbana-Champaign, USA

Abstract

Efforts to scale up fact-checking through technology, such as artificial intelligence (AI), are increasingly being suggested and tested. This study examines whether previously observed effects of reading fact-checks remain constant when readers are aware of AI's involvement in the fact-checking process. We conducted three online experiments ($N = 3{,}978$), exposing participants to fact-checks identified as either human-generated or AI-assisted, simulating cases where AI fully generates the fact-check or automatically retrieves human fact-checks. Our findings indicate that the persuasive effect of fact-checking, specifically in increasing truth discernment, persists even among participants without a positive prior attitude toward AI. Additionally, in some cases, awareness of AI's role reduced perceived political bias in fact-checks among Republicans. Finally, neither AI-generated nor human fact-checks significantly affected participants' feelings toward or their perceptions of the competence of the targeted politicians.

Keywords

Automated fact-checking, artificial intelligence, large language model, partisan bias, misinformation

Corresponding author:
Je Hoon Chae, Department of Communication, University of California, Los Angeles, 2133 Rolfe Hall, 405 Hilgard Avenue, Los Angeles, CA 90095, USA.
Email: chae@g.ucla.edu

The practice of fact-checking news, serving as a journalistic strategy to counteract the dissemination of misinformation, has seen an upsurge in the United States and across the globe (Graves, 2016). Nevertheless, the production of fact-checked news content is constrained by its dependence on meticulous research conducted by individual fact-checkers. Human research takes time, and this creates a gap between the ready availability of misinformation in the political environment and the supply of fact-checking news. Acknowledging this impediment, there have been noteworthy attempts to amplify the production of fact-checking content by implementing innovative methodologies such as the integration of machine learning algorithms (Thorne et al., 2018), thereby suggesting a shift toward more automated, scalable solutions.

The rapid advancement of machine learning technology, coupled with the increasing relevance of large language models (LLM) across diverse domains, has accelerated the evolution of artificial intelligence (AI)-assisted automated fact-checking strategies. For example, the U.S. fact-checking organization *Snopes* released in mid-2024 a beta version of its LLM-assisted automated fact-checking system, *FactBot*.[1] This system fully generates fact-checks based on *Snopes*'s fact-check archive when a user inputs a claim to be checked. *Meta* uses machine learning algorithms to identify potentially false factual claims within *Facebook* posts, complementing this with third-party human annotations to further enhance the integration of AI into the fact-checking process (Meta AI, 2020). Similarly, *Newtral*, a Spanish fact-checking site, has developed an AI tool called *Claim Hunter*, which automatically detects claims to be checked (Abels, 2022). As these cases illustrate, the incorporation of AI into fact-checking processes is becoming increasingly popular, ranging from simple claim detection to prompted fully automated fact-check generation.

While technological advancements in AI-assisted fact-checking are undeniable, our understanding of its public acceptance remains unclear. This raises critical questions: *Will the empirically observed effects of traditional fact-checking endure when AI intervenes in the fact-checking process? If they do, to what extent? If not, under which conditions might they vary?* This paper addresses these questions with empirical evidence and is structured as follows: First, we briefly review the contexts and purposes for which technology has been developed to automate fact-checking. Next, we derive empirically testable hypotheses from the literature on the effects of fact-checking, focusing particularly on comparisons between AI-generated and human-generated content. Finally, relying on three original survey experiments, we examine the persuasive effects (i.e., effectively increasing truth discernment), political bias perception (i.e., hostile media perception), and impact on attitudinal indicators such as feeling thermometers or perceived competence regarding the original claimant, comparing cases where readers perceive the content as generated by AI versus human.

Intervention of AI in the fact-checking process

AI's role in journalism has become significant, and fact-checking is no exception. Following the taxonomy proposed by Guo et al. (2022), we review how AI has intervened in the production processes of fact-checking, focusing on three sequential primary tasks: (a) claim detection, (b) evidence retrieval, and (c) verdict prediction or justification production.

In the first case, AI is employed to identify factual claims within a corpus that are suitable for assessment by fact-checkers or platforms. The objective of this identification can be guided by one of two slightly different criteria: determining whether a claim is "check-worthy" (Hassan et al., 2015) or whether it is "checkable," meaning it can be verified with available evidence (Konstantinovskiy et al., 2021). The "check-worthy" criterion involves a more subjective interpretation, focusing on the significance of the claim, whereas the "checkable" criterion emphasizes the practical feasibility of verifying the claim. Regardless of the criterion, claim detection is typically approached as a binary classification problem, where trained models predict whether a claim is check-worthy (or checkable). The methodologies employed range from supervised machine learning models (Aker et al., 2017) to advanced deep neural networks (Zhang et al., 2021).

Subsequent to claim detection, the evidence retrieval phase involves gathering relevant data from a variety of sources to support fact-checking efforts. The most straightforward source for fact-checking evidence traditionally comes from existing content on established fact-checking websites (Alhindi et al., 2018). For example, the aforementioned early version of *FactBot* by *Snopes* draws information from its own archived fact-checking articles. However, this method has its limitations, especially when dealing with new claims that have not previously been checked by humans. To overcome this, various alternative sources have been explored, including news headlines (Ferreira & Vlachos, 2016), Wikipedia entries (Thorne et al., 2018), and search engine results (Augenstein et al., 2019). In this phase, the role of AI models is to efficiently and accurately retrieve pertinent evidence from these sources in response to specific queries (Guo et al., 2022).

The final task, verdict prediction or justification production, highlights the potential of contemporary AI. In this phase, sophisticated models are employed to analyze the evidence and assess the veracity of the detected or given claim. Once the necessary evidence is aggregated, algorithms, often built on deep learning foundations, evaluate the claim. Their objective is twofold: to either validate the claim's authenticity (Wang, 2017) or to construct a compelling rationale for the judgment (Atanasova et al., 2020), with several models demonstrating reasonable accuracy in these tasks (Wang, 2017). Most notably, the recent rapid growth of LLMs like *ChatGPT* is accelerating the integration of AI in this area. While it is worth noting that the effectiveness of LLMs in this task is currently debated (DeVerna et al., 2023; Hoes et al., 2023; Yang & Menczer, 2023), their potential in this domain is hard to deny and merits further investigation.

Human-AI interaction and fact-checking

In this section, we review the literature on the effects of fact-checking, including its persuasive impact, hostile media perceptions, and potential downstream effects on other attitudinal indicators. We intentionally focus on fact-checking in the political realm, where cognitive biases are particularly prevalent and can significantly influence how individuals process content. We connect these insights with the literature on human perception of AI. This integration allows us to derive several empirically testable hypotheses and research questions.

The persuasive effect of political fact-checking

We conceptualize fact-checking on political issues as a form of persuasive messaging, aligning with O'Keefe's (2015, p. 27) definition of persuasion as "a successful intentional effort at influencing another's mental state through communication in a context where the persuadee retains some degree of autonomy." A primary concern in this domain, particularly regarding the autonomy of the persuadee, is motivated reasoning (Kunda, 1990). Specifically, when fact-checking messages are counter-attitudinal (e.g., debunking a claim made by a politician from the recipient's preferred party), there can be a tendency for directional motivation to override accuracy motivation. As a result, individuals may resist accepting the message, leading to a failure in persuasion (Nyhan and Reifler, 2010). However, recent studies consistently show that fact-checking is persuasive (Walter et al., 2020; Wood and Porter, 2019), even when the fact-checking story is counter-attitudinal (Coppock et al., 2023) and when the source of the fact-check is perceived as an out-group (Chae et al., 2024). This finding remains consistent across a spectrum of issues, encompassing both political and nonpolitical domains (e.g., Bode and Vraga, 2018; Chae et al., 2024; Coppock et al., 2023; Walter et al., 2020).

However, the question arises: Will these persuasive effects remain consistent when AI intervenes in the fact-checking process, whether to a greater or lesser extent? Recent studies, while not directly situated in political contexts, reveal the effectiveness of AI-labeled fact-checking in certain scenarios. Specifically, Liu et al. (2023) established that AI-labeled fact-checking is on par with other sources in countering misinformation related to health. Similarly, Bode and Vraga (2018) found that algorithmic interventions are as effective as social interventions in correcting health-related misinformation. Neither of the two studies uncovered evidence suggesting whether algorithm or AI-labeled fact-checking is more (or less) effective than human fact-checking. However, their investigations primarily focused on less-contentious issues, where group identity plays a minimal role in diminishing or amplifying the effects, unlike the case with claims made by polarizing political figures. Our study aims to bridge this gap by evaluating the persuasive impact of AI-labeled fact-checking on statements made by polarizing political figures.

At the same time, it is crucial to acknowledge potential heterogeneity rooted in individuals' unique beliefs about how AI's characteristics differ from humans' (Sundar, 2020). Such beliefs might sway positively, with individuals viewing AI as generally more precise or impartial, or negatively, leading some to perceive AI as inflexible and devoid of the subtleties inherent in human judgment. Sundar and colleagues have labeled readily applied beliefs about AI as machine heuristics, mental devices that guide cognition and attitudes in conditions of low effort or limited processing capacity (Molina and Sundar, 2022a, 2022b). The incorporation of machine heuristics within the context of fact-checking is not a novel concept. Drawing upon the theoretical foundations of the MAIN model (Sundar, 2008), which posits that the heuristic related to machines influences the credibility of machine-assisted or generated technology, Banas et al. (2022) demonstrated that individuals who are more inclined to apply a positive machine heuristic tend to perceive greater trustworthiness in the outcomes of AI-labeled fact-checking. Our study extends this inquiry by examining whether these findings also apply to

persuasive effects. Specifically, we investigate whether individuals with a heightened machine heuristic are more likely to be persuaded by AI-delivered fact-checking. We particularly emphasize the dimension of machine heuristics related to perceptions of AI's unbiasedness. This focus is pertinent in the political realm, where bias perception is crucial (Chae et al., 2024), and in light of recent studies suggesting that AI is perceived as more impartial than humans (Hidalgo et al., 2021). Informed by this theoretical discussion and empirical evidence, we postulate initial hypotheses and research questions:

Hypothesis 1 (H1): People reading human-delivered fact-checking news will demonstrate more accurate beliefs than those not reading fact-checks.

Hypothesis 2 (H2): People reading AI-delivered fact-checking news will demonstrate more accurate beliefs than those not reading fact-checks (**H2a**). However, the persuasive effect of AI-delivered fact-checking news will be stronger as people perceive AI as unbiased compared to humans (**H2b**).

Research Question 1 (RQ1): How does the corrective effect of AI-delivered fact-checking news on beliefs differ from the corrective effect of human-delivered fact-checking news on beliefs?

Hostile media effects and political fact-checking

Whereas individuals typically adjust their factual beliefs consistently, irrespective of personal biases, recent studies have spotlighted a tendency among readers, particularly those with strong partisan leanings, to discern political bias in fact-checking (Chae et al., 2024; Li et al., 2022). This perception of bias intensifies when the fact-checking content challenges audience members' viewpoints or is perceived to emanate from an out-group (Chae et al., 2024). This tendency corresponds with the well-documented phenomenon of hostile media perception (HMP; e.g., Gunther and Schmitt, 2004).

Reassuringly, Chae et al. (2024) demonstrate that the perception of political bias does not impede the persuasive effects of fact-checking, regardless of whether the content is counter-attitudinal or delivered by sources from an opposing party. However, drawing from the extensive research on the HME, we recognize that perceived bias is associated with various consequences, including diminished political participation (Moy et al., 2005), pluralistic ignorance (Gunther and Chia, 2001), and decreased confidence in democracy (Tsfati and Cohen, 2005). While HME of fact-checking may not directly relate to the immediate persuasive effect of fact-checking, the implications of HME in this context are significant, as they could influence real-world political behaviors beyond the experimental setting.

In this study, we address a question hitherto unexplored in the existing literature: Does the perception that AI has intervened in the process of political fact-checking reduce the HME? We hypothesize that, within this dynamic context, AI-labelled fact-checking could potentially serve as a tool to mitigate biased perceptions that are typically fueled by individual motivations. This conjecture is grounded in recent studies investigating human perceptions of material generated by AI.

For example, recent research suggests that AI involvement in various tasks is often linked to perceptions of impartiality. For instance, Hidalgo et al. (2021) investigated differences in how people perceive human versus machine actions across various scenarios, even when the actions were essentially the same. The study found that people tended to be more forgiving of actions performed by machines, particularly in contexts related to the fairness of differing actions (Hidalgo et al., 2021). In a similar vein, Helberger et al. (2020) conducted a study where participants compared the fairness of decisions made by AI and humans. Notably, 54% of participants favored AI as the more equitable decision-maker, a view shared by only 33% of human decision-makers (see, also, Marcinkowski et al., 2020).

Moreover, within the context of the perception of fact-checks, Moon et al. (2023) revealed that individuals with pronounced partisan biases perceive AI-labeled fact-checking as enhancing the credibility of a narrative. Similarly, Banas et al. (2022) showed that individuals may perceive an AI-labeled fact-checking story as more trustworthy than one delivered by humans, even in instances where the results are contradictory. These empirical studies collectively suggest a trend where material generated by AI is generally viewed as impartial, with even the results of fact-checking being perceived as more credible or trustworthy. In our research, we aim to investigate whether this positive perception of AI-driven outputs, as compared to human ones, extends to perceptions of political bias in fact-checking. Specifically, we examine whether AI-labeled fact-checking can diminish the HME among groups that perceive political bias in fact-checking. The political context renders fact-checking persuasive in nature. Thus, people encountering political fact checks will be inherently suspicious of message sources, a tendency that should favor the seemingly less biased machine sources.

Hypothesis 3 (H3): People reading the AI-delivered fact-checking news will perceive less bias regarding the check and its author than those reading the human-delivered fact-checking news.

Downstream effects of fact-checking

In addition to the persuasive effects and perceptions of fact-checking, our research explores its impact on attitudinal dimensions, such as impressions or evaluations of politicians and general sentiment toward them. Current studies, including Wintersieck (2017), suggest that fact-checking can indeed influence these aspects, albeit with a modest effect size. For instance, Wintersieck's analysis of the 2013 New Jersey gubernatorial race and subsequent fact-checking of candidates' statements revealed that affirming a claim as true influenced evaluations of the candidate positively. Similarly, Coppock et al. (2023) observed that respondents' emotional reactions toward politicians varied in response to political fact-checking, though the effect size was minimal.

Like the persuasive effects, we do not specifically anticipate a difference in the effect of fact-checking on politician evaluations and feeling thermometers between human and AI-delivered methods for the reasons previously discussed. Consequently, we separately examine the effect of fact-checking on politician evaluations and feelings toward them,

treating the former as a hypothesis-driven inquiry and the latter as a research question, given the marginal or nonexistent effects observed in prior studies for affective/feeling outcomes.

Hypothesis 4 (H4): People reading human-delivered fact-checking news will believe that the person targeted by the fact-checking is less competent (or more competent, depending on predispositions) than will people not reading the fact-check.

Hypothesis 5 (H5): People reading AI-delivered fact-checking news will believe that the person targeted by the fact-checking is less competent (or more competent, depending on predispositions) than will people not reading the fact-check.

Research Question 2 (RQ2): How does the affective reaction toward the figure targeted by fact-checking differ between subjects in the human-delivered fact-checking newsgroup and those in the control group?

Research Question 3 (RQ3): How does the affective reaction toward the figure targeted by fact-checking differ between subjects in the AI-delivered fact-checking news group and those in the control group?

Data and methods

Data and experimental designs

Study 1 and study 2. Data for both studies were collected from participants recruited via the *Prolific Academic* online panel. Study 1, conducted from May 28, 2023, to June 7, 2023, started with an initial sample of $N = 1{,}135$ and, following preregistered quality checks, had a remaining sample size of $N = 1{,}077$. Data collection for Study 2 took place from September 1, 2023, to September 6, 2023. Initially, there were 1,144 responses, but after preregistered quality checks, the sample included $N = 1{,}063$ participants who had not participated in Study 1. For both studies, all treatment materials, survey items, a priori hypotheses and research questions, experimental design, and plans for estimating causal effects and statistical hypothesis testing are preregistered.[2] Studies 1 and 2 follow a common workflow (see Figure A1), though the contexts of the studies differ substantially. In Study 1, participants are given full-text fact-checking articles to read, whereas in Study 2, participants read only the tagged fact-checking headlines on a simulated social media setup. To enhance ecological validity, all treatment materials used in these studies are based on real fact-checking cases.

Participants initially provided pretreatment covariate data, including their partisanship, ideology, media trust, general political knowledge, and political interest. Then, in Study 1, participants are randomly assigned to one of three groups. Those in the control group receive a factual claim made by a politician (either Joe Biden or Donald Trump) and evaluate the claim's factual accuracy and their confidence in this assessment. This process is repeated for four distinct claims, each classified as either true or false and attributed to Biden or Trump, resulting in four scenarios. The order of these claims is randomized to mitigate potential carryover effects. Participants in the human fact-check

group receive a factual claim followed by a fact-checking news item from *FactCheck.org*, which verifies the claim's veracity. Participants then answer questions about their belief in the claim and its certainty, and additionally, they report their perceptions of the bias of the fact-checking news. Participants in the AI fact-check group are presented with a similar setup, again with four order-randomized fact-checking stimuli, but they are informed that the fact-checking was fully generated by AI, akin to *Snopes*'s fully automated fact-checking system, *FactBot*. The remaining procedures are identical to those in the human fact-check group.

Study 2 replicates these procedures, with the key difference being that all factual claims are sourced from social media posts (i.e., Facebook), and the fact-checks are presented in headline format rather than full text. In the human fact-check group, participants are provided only with the headline of the fact-checking report posted by *FactCheck.org*'s official account. The AI fact-check group in Study 2 involves a more limited role for AI compared to Study 1. Here, while the AI system automatically matches the relevant fact-checking results, human experts perform the actual fact-checking.

Study 3. Study 3 was conducted from May 20, 2024, to May 24, 2024, through *Prolific Academic*. Initially, there were 1,901 responses, with 1,838 remaining after quality checks. In this study, we consolidated all experimental conditions from the two previous studies into a single design: 1 (control) plus 2 (strength of intervention: headline only vs. full text) by 2 (deliverer/writer of fact-checking news: AI vs. human) between-subjects factorial design (see Figure A2). The purpose of conducting Study 3 was to replicate the main findings of Studies 1 and 2 while also testing different fact-checking stories from various fact-checking organizations. Additionally, this design allows us to evaluate AI intervention at different levels (i.e., headline vs. full text) within a single research design. Therefore, except for using different treatment materials and combining the experimental groups from previous studies into one design, all other aspects remained the same. In this experiment, we utilized four actual fact-checking reports from *PolitiFact*, each assessing whether claims made by Trump or Biden were false or true. For a detailed examination of the treatment materials, refer to Supplementary Material section B.

Measurement

Our primary outcome variables include a composite measure of *truth discernment with certainty* (Coppock et al., 2023). We first employed a four-point belief score measure for each factual claim by asking, "How accurate is this statement?" and gauged the certainty of that answer via a 4-point scale with endpoints of 1 (i.e., "not at all accurate") and 4 (i.e., "very accurate"). Subsequently, we dichotomized their belief score into correct (value $= 1$) and incorrect (value $= -1$) answers and multiplied the belief certainty by that value. This was then rescaled to range from 0, representing an incorrect belief held with certainty, to 100, indicating a correct belief held with certainty. We opted for this composite measure because we believe it aptly reflects the conceptual definition of misperception, which can be described as a belief in misinformation held with certainty. However, all the findings are replicated using the four-point scale truth discernment

measure, which is coded such that increasing values align more closely with the conclusions of the fact-checking (see section F of the Supplementary Materials).

To gauge hostile media perceptions in fact-checking, we utilized two items from Gunther and Schmitt (2004). These items asked whether (a) the content of the fact-checking news in the article and (b) the writer of the fact-checking news was strictly neutral, or if they were biased in favor of Republicans or Democrats. The scale was adjusted to range between −1 and +1, where +1 indicates a perception of bias unfavorable to the in-group (i.e., hostile media perception), −1 signifies a perception of bias favorable to the in-group (i.e., biased assimilation), and 0 represents a perception of neutrality. Additionally, we evaluated the feeling thermometer toward the target politician, which ranged from 0 to 100. High scores represent positive feelings, 50 indicates neutrality, and low scores indicate negative feelings. We also assessed the *perceived competence* of the politician making the factual claim, which ranged from 1 to 7 using the average score of five traits (i.e., sincere, reliable, trustworthy, expert, accurate). The reliability of the composite measure is higher than .94 in all cases across all studies.

Along with the demographic variables, we measured a group of pretreatment covariates that we theorized as moderators or conceivably related to the dependent variables. The measured covariates include machine heuristic acceptance, political partisanship, political ideology, political interest, trust in news media, and general political knowledge. For details on the measured pre-treatment covariates, refer to section D of the Supplementary Material.

Results

Persuasive effects of human/AI-delivered fact-checking

To statistically test H1 and H2a, which predict the effectiveness of human-delivered fact-checking and AI-delivered fact-checking, respectively, we employed an OLS regression to estimate the average treatment effect for each. In Figure 1, each point represents the regression coefficient indicating the extent to which the belief levels are aligning closely with the verdict of fact-checking compared to those in the control group, where only the factual claim was presented, and participants were asked about their belief in the claim. Consistently, we observe that in most cases, beliefs about the accuracy of the claim align more closely with the conclusions of the fact-checking, regardless of the party identity of the respondents. Regression results aggregating all partisans in each experimental group indicate that for all treatment groups across all three studies, fact-checking demonstrates statistically significant persuasive effects (see Tables F14 to F16). These results support H1 and replicate previous research when human fact-checkers conduct the fact-checking (Wood & Porter, 2019).

Next, with respect to H2a, which predicts that AI-delivered fact-checking will exhibit a persuasive effect, we observed a pattern consistent with human-delivered fact-checking cases. That is, even when respondents are informed that the fact-checking was conducted by AI or that the fact-checking headline was automatically curated by AI, the general tendency to update their beliefs in the direction suggested by the fact-checking consistently occurs across all three studies (for CATE estimates, see Figure 1). Regarding

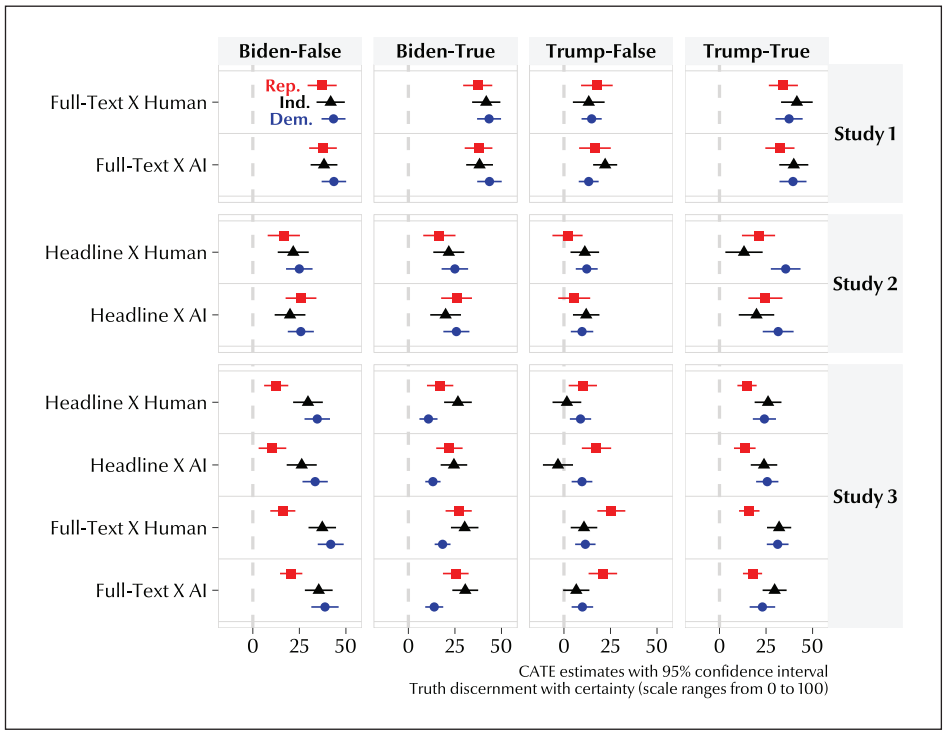

Figure 1. Effect of AI-delivered fact-checking and human-delivered fact-checking on truth discernment with certainty by party ID (PID).

Note. Points are point estimates, and bars are 95% confidence intervals calculated with HC2 standard errors. Red triangles, black squares, and blue circles represent Republican, Independent, and Democrat PID, respectively.

the ATE estimates, regardless of party identity, the smallest effect observed in the studies involving AI-delivered headlines was $b = 10.09$, $SE = 7.72$, $p < .001$, 95% CI [6.03, 14.15]. In cases involving full-text fact-checking by AI, the effect was $b = 13.23$, $SE = 2.04$, $p < .001$, 95% CI [9.23, 17.22] (both were Study 3: Treatment material of Trump-False). Additionally, all findings were consistent when they were estimated with unadjusted differences in means (Tables F17 to F19) and when using the simple four-point scale truth discernment measure (Tables F20 to Table F22). That is, H2a is supported.

Regarding RQ1, concerning any differences in the average treatment effect on truth discernment with certainty between AI-delivered and human-delivered fact-checking, our findings do not consistently support the possibility that one method is universally more or less effective than the other. We compared the differential effect sizes, adjusting the *p*-values using the Benjamini-Hochberg method—which controls the false discovery rate by ranking the *p*-values from multiple tests and then determining a threshold below which *p*-values are considered significant—as preregistered. Only one case was found to be different at the 0.05 level, while the remaining 15 pairs showed no

significant differences (see Table F23). Consequently, our empirical evidence indicates that the persuasive effect of fact-checking is similar, irrespective of whether it is delivered by AI or a human. In Study 3, thanks to a design that unified all full-text and headline treatment materials, we were able to compare the differential effect sizes between full-text and headline treatments for both AI and human-delivered cases. The results indicate that when humans delivered the fact-checks, the persuasive effect of full-text fact-checking was more pronounced compared to the headline-only cases. When AI delivered the fact-checks, this differential effect was observed in the same direction but with a weaker effect size after adjusting the *p*-values (see Table F24).

In addressing H2b, which posits that the effectiveness of AI-delivered fact-checking varies according to an individual's general bias perception toward machines, we employed a binned estimator for each tercile (i.e., low, mid, and high) while also estimating the linear moderation effect as suggested by Hainmueller, Mummolo, and Xu (2019). As illustrated in Figure 2 for Study 1, a clear trend emerges, indicating that the effect of AI-delivered fact-checking intensifies in three out of the four cases as individuals perceive AI to be more unbiased compared to humans. In contrast, for Study 2, where AI had a more subtle intervention during the fact-checking process, a moderation effect was observed for only one stimulus (i.e., asserting Donald Trump's claim as false): $b = 2.58$, $SE = 1.10$, $p = .02$, 95% CI [0.41, 4.74]. In Study 3, where we explored the same concept but with entirely different topics and materials for fact-checking several months after Studies 1 and 2, we did not observe a strong linear relationship between bias perception of AI and its persuasive effect (see also Figure F44 for the four-points scale measure). However, importantly, a notable pattern revealed by our binned estimator approach is that, regardless of a low belief in the unbiasedness of AI, a clear, persuasive effect of fact-checking increasing the truth discernment with certainty was evident. In a nutshell, these results thus provide mixed support for H2b.

Mitigation of hostile media perceptions by AI-delivered fact-checking

Before addressing H3, we first examine whether there is a hostile media effect regarding fact-checking in general. Notably, we confirmed a clear and consistent pattern of the hostile media effect among Republicans, particularly when the fact-checking content contradicts their attitudes—that is, when fact-checking reports suggest that Trump is wrong or Biden is correct. Figure 3 graphically illustrates this pattern, clearly showing the raw distribution scores of bias perception alongside the results of the two-sided one-sample *t*-test. Across all three studies, it is evident that when the fact-checking conclusions are counter-attitudinal, participants perceived the news as being biased in favor of their out-group. Generally, the effect size, quantified using Cohen's *d*, was more pronounced in Studies 1 and 2 (e.g., $d = 0.73$ and $d = 0.78$, respectively, with the fact-checking concluding that Biden's claim is true), compared to Study 3, where $d = 0.58$. For a more detailed quantification of the mean comparison from a neutrality score of 0, see Tables F26 to F28.

Now that we have empirically confirmed that Republicans, in particular, perceive fact-checking as unfavorably biased when it is counter-attitudinal, we will examine whether AI-delivered fact-checking can significantly reduce the extent of this

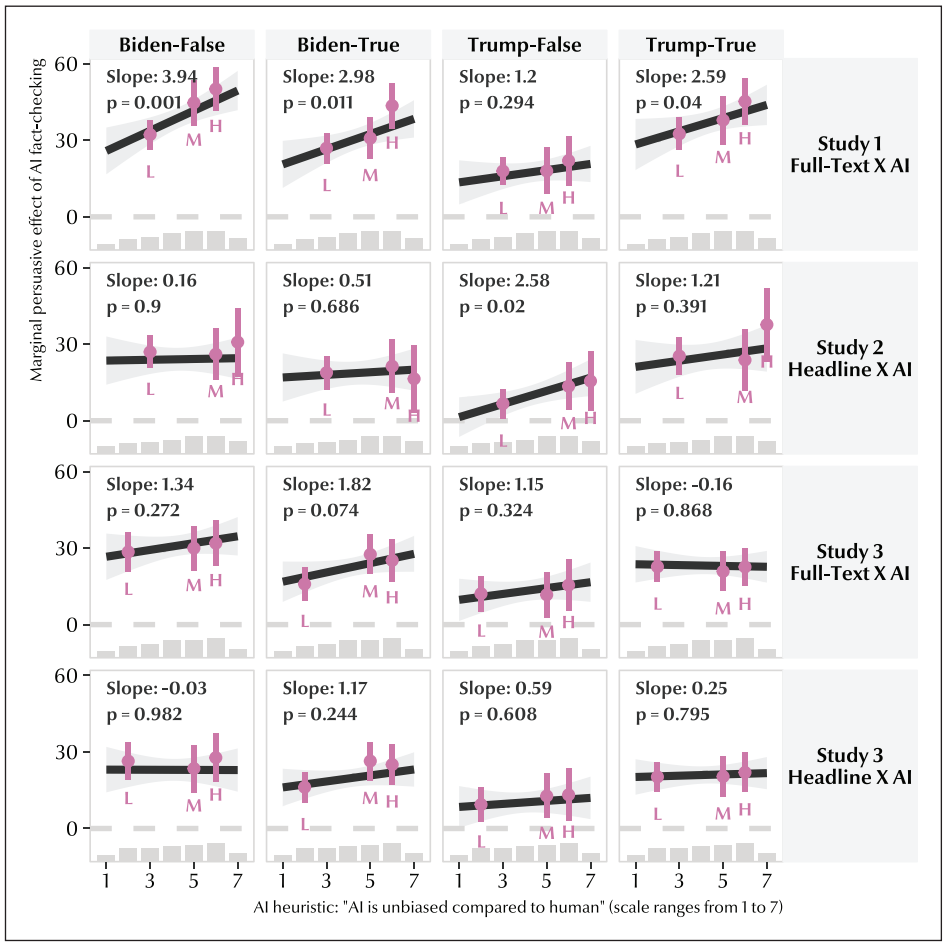

Figure 2. Estimated average treatment effect of AI-delivered fact-checking on truth discernment with certainty marginal on the perception of machine bias.
Note. Points are point estimates, and bars are 95% confidence intervals calculated with HC2 standard errors. The grey bars at the bottom of each panel represent the distribution of the moderator variable measure for each data collection. These bars provide a visual sense of the distribution of the moderator variable and are not related to the scale of the vertical axis (L: low; M: mid; H: high).

perception, as predicted by H3. Figure 4 displays the OLS estimates along with 95% confidence intervals. Each point represents the estimated regression coefficient of hostile media perception when fact-checking is delivered by humans as compared to AI. A negative direction indicates that participants perceive less bias when fact-checking is conducted by AI. As illustrated in Figure 4, across the three studies, in half of the counter-attitudinal cases for Republicans, fact-checking news was perceived as less biased when delivered by AI. However, this pattern was more pronounced in Studies 1

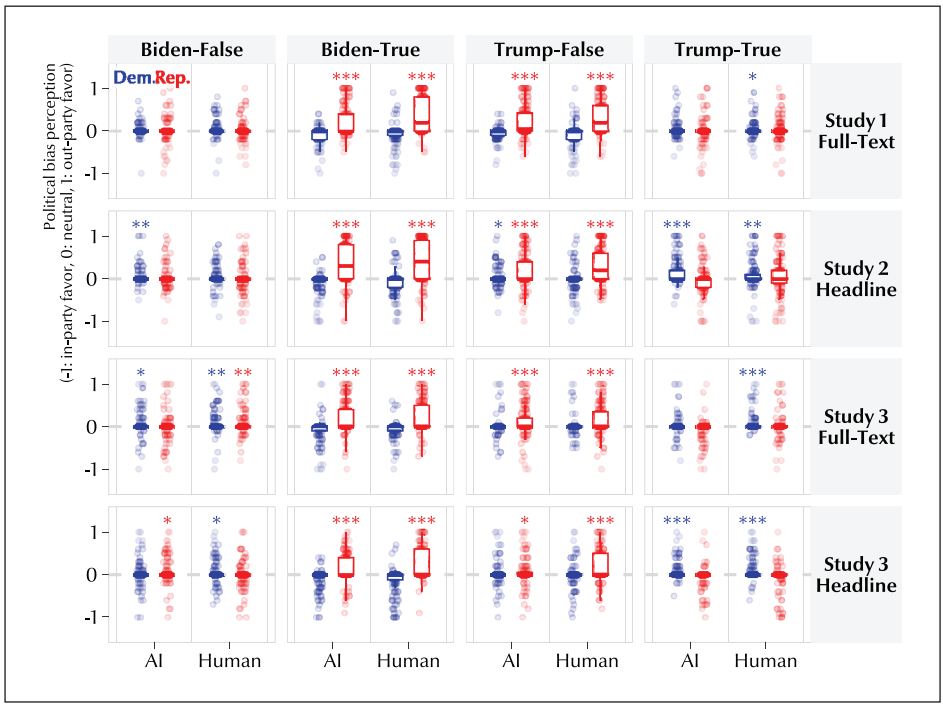

Figure 3. Distribution of hostile media perceptions on AI-delivered fact-checking and human-delivered fact-checking by PID.

Note. The scale of hostile media perception is represented as follows: when it is 1, it indicates that the participant perceives the fact-checking as favorable toward the out-group; 0 means the participant perceives it as neutral; and −1 indicates a perception of the fact-checking being favorable toward the in-group. Asterisks denote the levels of statistical significance from the two-sided *t*-test compared to the neutrality score of zero (*$p < .05$; **$p < .01$; ***$p < .001$).

and 2 compared to Study 3. This discrepancy can be partially explained by the potential floor or ceiling effects, in that the effect sizes of hostility perception from Republicans in Study 3 were generally weaker than those observed in the previous two studies. In sum, we conclude that the results of the three studies provide mixed support for H3.

Minimal/null downstream effects on feeling thermometer and perceived competence

In relation to RQ2 and RQ3, which inquire about potential downstream effects of both human-delivered and AI-delivered fact-checking on feeling thermometer scores for the target politician, our findings indicate no significant effect across the three studies. This suggests that exposure to fact-checking, irrespective of being human-delivered or AI-delivered, doesn't notably change the warmth or coldness of individuals' sentiments toward politicians who have been the subjects of fact-checking. This observation may be

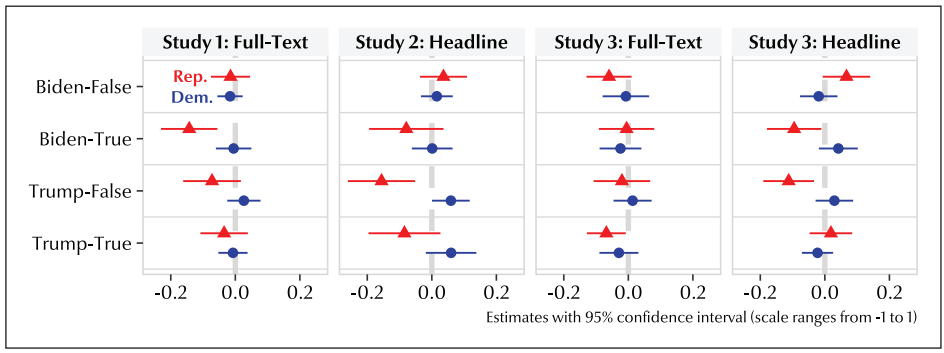

Figure 4. Estimates of the difference of the hostile media perception between human and AI fact-checking (AI compared to human).
Note. Points represent the point estimates, and bars indicate the 95% confidence intervals, which are calculated using HC2 standard errors.

partially attributed to the fact that our chosen politicians for fact-checking are already notably prominent and polarizing. Across the three studies, we measured how participants felt toward Trump and Biden at various points. Regardless of treatment group assignment and measurement timing, Democrats consistently scored Trump between 10 and 15 and Biden around 60. Conversely, Republicans scored Biden similarly low and Trump around 60. For nonparty supporters, the feeling thermometer toward both politicians consistently stayed within the 25 to 35 range, which is generally quite cold. Detailed results are available in Tables F29 to F31 and Figure F45.

H4 and H5 posited that for both human-delivered and AI-delivered fact-checking, respondents' perceptions of the focal politician would be swayed by the verdict of the fact-check. Specifically, if the fact-checking determines that a politician's statement is factually inaccurate, it would lead to a decline in the politician's perceived competence. Conversely, if the fact-checking establishes the claim as true, the perceived competence would increase. However, the influence of fact-checking on the perceived competence of the politician in question is typically negligible or, in instances where it does manifest within a subgroup, it remains minimal. This pattern suggests that fact-checking may not significantly alter public perceptions of a politician's competence, particularly in the case of the well-known political figures featured in our experiments. Similar to findings with the feeling thermometer, this negligible effect may be due in part to the substantial prominence of the politicians being scrutinized. Detailed estimates are provided in Tables F32 to F34 and Figure F46.

Discussion

Against the rapid spread of misinformation, the integration of machine learning or AI with efforts to combat misinformation through fact-checking is coming closer to regular application (Graves, 2018). In a series of studies, we address a not-yet-answered, but crucial, question: What are the implications of AI's intervention in the process of fact-check generation?

Through three experiments, we simulated: (a) a strong intervention (i.e., full-text generation) scenario where AI actively crafts the justification and verdict of fact-checking (Study 1), (b) a weak intervention (i.e., headline tagging) scenario wherein AI merely identifies the factual claim and matches it with the relevant fact-checking headline on social media (Study 2), and (c) a scenario that combines elements of both the strong and weak interventions into a single research design (Study 3).

From our analysis, several key findings emerged. Firstly, the persuasive effect of fact-checking remains consistent whether AI intervenes strongly or weakly. Secondly, although the evidence is mixed, as respondents begin to view AI as less biased compared to humans, the persuasive effect of AI-delivered fact-checking increases, particularly when AI fully generates the fact-checking. Importantly, the persuasive effect of AI fact-checking remains consistent even among those who do not necessarily believe that AI is unbiased. Thirdly, Republicans often perceive fact-checking as being biased in favor of Democrats when the verdict is counter-attitudinal, but this hostile media effect tends to diminish—though very weakly—when fact-checking is delivered by AI. Lastly, irrespective of whether fact-checking is human-delivered or AI-delivered, its effect on a target politician's feeling thermometer or perceived competence is, at best, minimal and most often null.

Implications

Our empirical findings lead to several crucial implications and points worth discussing. First, our results indicate that, despite individuals' awareness of AI's involvement in the fact-checking process, the primary objective of fact-checking—increasing people's truth discernment—remains neither diminished nor amplified by AI's role. This suggests that, despite widespread discussions about public distrust in AI (Dwork and Minow, 2022; Jacovi et al., 2021), at least in the context of fact-checking, people's perceptions of AI's role in information generation do not seem to be a significant concern. This finding is especially important given the rampant spread of misinformation in today's digital age. The consistent, persuasive impact of AI-assisted fact-checking accentuates the immense potential of integrating AI in countering misinformation. Organizations specializing in fact-checking might substantially benefit from incorporating AI technologies into their operational framework. However, this is contingent upon two pivotal factors: the accuracy of the AI-generated information and the transparent communication of the information-generation process to users or audiences.

Second, our findings that Republicans perceive less bias in fact-checks delivered by AI suggest a potential tactic for addressing the high tendency of political bias accusations among partisans, particularly in the United States. The use of AI, which is often seen as more neutral and impartial than humans, could be a potential key to overcoming this barrier. By ensuring transparency in AI's operational algorithms and highlighting its relatively impartial nature, fact-checking institutions might be able to bypass entrenched perceptions of political bias. This approach could increase the acceptance of fact-check outcomes across partisan lines, with broader implications for fostering a more informed and less polarized society, especially in an era where misinformation can significantly influence public opinion and decision-making.

Third, our findings contribute to the ongoing discussions about the role of AI and the future of democracy, both from optimistic and pessimistic perspectives. Drawing from Habermas's concept of the public sphere, the quality of a democracy has traditionally been measured by the quality of its discourse—where open debate, rigorous scrutiny of facts, and respect for diverse viewpoints are essential to its health (Habermas, 1991). However, the proliferation of misinformation and the rise of echo chambers have put these foundations of communicative action under strain (Sunstein, 2018). From an optimistic viewpoint, our study suggests a potential role for neutral AI entities to serve as "mediators" or "moderators" in public discourse. By ensuring the accuracy of claims on a large scale (or in real-time) and being perceived as unbiased, AI could help rejuvenate the public sphere by fostering debates that are more fact-driven than sentiment-driven (Dahlberg, 2001) and at a lower cost. In this light, the future of democratic discourse might embrace a hybrid model, where humans and AI collaboratively ensure a more informed and balanced public conversation, aligning with the democratic ideal of a well-informed citizenry.

Now, consider the less optimistic perspective. Although AI has the potential to serve as a significant independent agent in democratic societies, its sophistication compared to simpler machine learning algorithms does not guarantee freedom from the social biases inherent in human behavior (Ananny & Crawford, 2018). AI is intrinsically linked to the data it is trained on, which is sourced from humans—who are prone to various biases. These biases can become ingrained in the algorithms in ways that are often opaque (Shin et al., 2022). In the case of automated fact-checking, if the training data contain biases (which are difficult to quantify), the rapid large-scale generation of fact-checking content could merely amplify those biases. Thus, it would be overly naive or optimistic to assume that AI can function as a truly unbiased and independent agent in mediating or moderating roles within democratic societies. Although current research does not definitively point to whether we should be optimistic or pessimistic, our main empirical finding—that the persuasive effect of fact-checking remains consistent regardless of whether it is delivered by AI or humans and irrespective of whether AI is perceived as more unbiased—will be crucial in shaping future discourse.

Lastly, as a caveat rather than an implication, we emphasize that the potential advantages of incorporating AI into fact-checking and broader strategies to counter misinformation critically depend on proper execution. The positive implications of this study's findings can only fully materialize if certain key elements are meticulously addressed. First, quality control is paramount. The automated fact-checking process must maintain high precision, ensuring that claims are analyzed accurately and minimizing the potential for false positives or negatives. Second, transparency is crucial. It is essential to clearly delineate the extent of AI's involvement and intervention in the fact-checking process. Stakeholders and the public must be informed about key details such as the training data used, the algorithms' decision-making processes, and any inherent biases that may influence outcomes. Without this level of transparency, trust in AI-assisted fact-checking could erode, leading to skepticism and potentially undermining genuine fact-checking efforts. In the fight against misinformation, it is not enough to have advanced tools at our disposal; these tools must be both reliable and transparent, paving the way for informed public trust.

Limitations

Whereas this study offers valuable insights, we must acknowledge several limitations. First, to prioritize internal validity and isolate the effect of AI intervention, we stylized the context of fact-checking exposure. Although there are cases like *Snopes*'s *FactBot*, where AI fully generates fact-checks, this is not yet common practice. At the time of writing, we remain uncertain whether fully automated fact-checking will become widespread, given the many technical barriers that must be overcome—such as hallucination, where AI models generate incorrect or nonsensical information presented as factual. Additionally, it is challenging to predict what connotations AI will carry if its intervention becomes prevalent, which could be a potential confounder when applying our empirical findings to the future should such a scenario occur. In this context, the conditions of providing only a headline versus the full text may serve as conceptual lower and upper bounds of this real-world phenomenon.

Second, our focus on political fact-checking omits nonpolitical subjects like health or international affairs, limiting the generalizability of our findings. Additionally, our study's applicability is further restricted by its exclusive focus on the United States, one of the most exceptionally polarized countries globally. Consequently, the insights concerning political bias perception may only be relevant to the U.S. context or to other countries where ideological or affective polarization levels are comparably high.

Third, fact-checking is a complex process, and issues are often more nuanced than simply categorizing them as true or false. Consequently, many fact-checking articles conclude with relatively inconclusive results. This study operates under the assumption that fact-checking is conducted correctly and effectively, which may not always be the case in real-world scenarios. Readers should keep in mind that this is a limitation of the experimental design used in this study and likely in most experimental studies assessing the persuasive effect of fact-checking.

Lastly, our focus on polarizing figures like Donald Trump and Joe Biden may not fully represent fact-checking's impact on perceived competence or feeling thermometer, as their prominence could limit the effects. Readers should consider these potential measurement issues.

Concluding remarks

Despite the growth of the fact-checking industry over the past decade and the efforts of numerous fact-checkers, very few people are exposed to fact-checking content (Guess et al., 2020), which represents a significant societal inefficiency. Enhancing the speed and volume of fact-checking production with AI could undeniably transform the landscape of misinformation and fact-checking, provided the quality of the content is maintained optimally. Consequently, cautiously envisioning a future where real-time fact-checking occurs would not merely be an overly optimistic projection. In fact, the automation of the fact-checking process—whether in part or in whole—is advancing rapidly. This study provides substantial empirical evidence to support preparations for this emerging reality.

Author Notes

The authors have agreed to the submission, and this article is not currently being considered for publication by any other print or electronic journal.

Data Availability Statement

The data underlying this article will be shared on reasonable request by the corresponding author.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

Je Hoon Chae https://orcid.org/0000-0001-7620-2184

Supplemental Material

Supplemental material for this article is available online.

Notes

1. https://www.snopes.com/factbot/. Readers should note that the *Snopes* beta version was not yet released at the time the data were collected for this study.
2. Refer to https://doi.org/10.17605/OSF.IO/T7UES (Study 1), https://doi.org/10.17605/OSF.IO/ZK6VN (Study 2), https://doi.org/10.17605/OSF.IO/XBRNK (Study 3).

References

Abels G (2022). What is the future of automated fact-checking? Fact-checkers discuss. *Poynter*. Available at: https://www.poynter.org/fact-checking/2022/how-will-automated-fact-checking-work/ (accessed 11 August 2024).

Adair B (2023) Fact-checking needs a reboot. *Nieman Lab*. Available at: https://www.niemanlab.org/2023/12/fact-checking-needs-a-reboot/ (accessed 22 December 2023)

Aker A, Derczynski L and Bontcheva K (2017) Simple open stance classification for rumour analysis. *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*: 31–39

Alhindi T, Petridis S and Muresan S (2018) Where is your evidence: Improving fact-checking by justification modeling. *Proceedings of the First Workshop on Fact Extraction and Verification (FEVER)*: 85–90.

Ananny M and Crawford K (2018) Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society* 20(3): 973–989.

Augenstein I, Lioma C, Wang D, et al. (2019) MultiFC: A real-world multi-domain dataset for evidence-based fact checking of claims. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*: 4685–4697.

Banas JA, Palomares NA, Richards AS, et al. (2022) When machine and bandwagon heuristics compete: Understanding users' response to conflicting AI and crowdsourced fact-checking. *Human Communication Research* 48(3): 430–461.

Bode L and Vraga EK (2018) See something, say something: Correction of global health misinformation on social media. *Health Communication* 33(9): 1131–1140.

Chae JH, Lee SY and Song H (2024) Perceiving as biased but nevertheless persuaded? Effects of fact-checking news delivered by partisan media. *Political Psychology* 45(1):69–89.

Coppock A, Gross K, Porter E, et al. (2023) Conceptual replication of four key findings about factual corrections and misinformation during the 2020 US election: Evidence from panel-survey experiments. *British Journal of Political Science* 53(4): 1328–1341.

Dahlberg L (2001) Computer-mediated communication and the public sphere: A critical analysis. *Journal of Computer-Mediated Communication* 7(1): JCMC714.

DeVerna MR, Yan HY, Yang K, et al. (2023). Artificial intelligence is ineffective and potentially harmful for fact checking. *arXiv Preprint arXiv:2308.10800*. DOI: 10.48550/arXiv.2308.10800

Dwork C and Minow M (2022) Distrust of artificial intelligence: Sources & responses from computer science & law. *Daedalus* 151(2): 309–321.

Ferreira W and Vlachos A (2016) Emergent: A novel dataset for stance classification. *Proceedings of NAACL-HLT* 2016: 1163–1168.

Graves L (2016) *Deciding what's true: The rise of political fact-checking in American journalism*. New York, NY: Columbia University Press.

Graves L (2018) Understanding the promise and limits of automated fact-checking. *Reuters Institute for the Study of Journalism*. Available at: https://ora.ox.ac.uk/objects/uuid:f321ff43-05f0-4430-b978-f5f517b73b9b (accessed 28 September 2024).

Guess AM, Nyhan B and Reifler J (2020) Exposure to untrustworthy websites in the 2016 US election. *Nature Human Behaviour* 4(5): 472–480.

Gunther AC and Schmitt K (2004) Mapping boundaries of the hostile media effect. *Journal of Communication* 54(1): 55–70.

Gunther AC and Chia SCY (2001) Predicting pluralistic ignorance: The hostile media perception and its consequences. *Journalism & Mass Communication Quarterly* 78(4): 688–701.

Guo Z, Schlichtkrull M and Vlachos A (2022) A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics* 10:178–206.

Habermas J (1991) *The structural transformation of the public sphere: An inquiry into a category of bourgeois society*. Cambridge, MA: MIT Press.

Hainmueller J, Mummolo J and Xu Y (2019) How much should we trust estimates from multiplicative interaction models? Simple tools to improve empirical practice. *Political Analysis* 27(2): 163–192.

Hassan N, Li C and Tremayne M (2015) Detecting check-worthy factual claims in presidential debates. *Proceedings of the 24th ACM International Conference on Information and Knowledge Management CIKM* 2015: 1835–1838.

Hassan N, Zhang G, Arslan F, et al. (2017) Claimbuster: The first-ever end-to-end fact-checking system. *Proceedings of the VLDB Endowment* 10(12): 1945–1948.

Helberger N, Araujo T and de Vreese CH (2020) Who is the fairest of them all? Public attitudes and expectations regarding automated decision-making. *Computer Law & Security Review* 39: 105456.

Hidalgo CA, Orghian D, Canals JA, et al. (2021) *How humans judge machines*. Cambridge, MA: MIT Press.

Hoes E, Altay S and Bermeo J (2023) Leveraging ChatGPT for efficient fact-checking. *PsyArXiv Preprints*. DOI: 10.31234/osf.io/qnjkf

Jacovi A, Marasović A, Miller T, et al. (2021) Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in AI. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*: 624–635.

Konstantinovskiy L, Price O, Babakar M, et al. (2021) Toward automated factchecking: Developing an annotation schema and benchmark for consistent automated claim detection. *Digital Threats: Research and Practice* 2(2): 1–16.

Kunda Z (1990) The case for motivated reasoning. *Psychological Bulletin* 108(3): 480–498.

Li J, Foley JM, Dumdum O, et al. (2022). The power of a genre: Political news presented as fact-checking increases accurate belief updating and hostile media perceptions. *Mass Communication and Society* 25(2): 282–307.

Liu X, Qi L, Wang L, et al. (2023). Checking the fact-checkers: The role of source type, perceived credibility, and individual differences in fact-checking effectiveness. *Communication Research*. Epub ahead of print 27 October 2023. DOI: 10.1177/00936502231206419.

Marcinkowski F, Kieslich K, Starke C, et al. (2020) Implications of AI (un-) fairness in higher education admissions: The effects of perceived AI (un-) fairness on exit, voice and organizational reputation. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*: 122–130.

Meta AI (2020) Here's how we're using AI to help detect misinformation. Available at: https://ai.meta.com/blog/heres-how-were-using-ai-to-help-detect-misinformation/ (accessed 18 January 2024).

Molina MD and Sundar SS (2022a) Does distrust in humans predict greater trust in AI? Role of individual differences in user responses to content moderation. *New Media & Society*. Epub ahead of print 23 June 2022. DOI: 10.1177/14614448221103534.

Molina MD and Sundar SS (2022b) When AI moderates online content: Effects of human collaboration and interactive transparency on user trust. *Journal of Computer-Mediated Communication* 27(4): zmac010.

Moon WK, Chung M and Jones-Jang SM (2023) How can we fight partisan biases in the COVID-19 pandemic? AI source labels on fact-checking messages reduce motivated reasoning. *Mass Communication and Society* 26(4): 646–670.

Moy P, Torres M, Tanaka K, et al. (2005). Knowledge or trust? Investigating linkages between media reliance and participation. *Communication Research* 32(1): 59–86.

Nyhan B and Reifler J (2010) When corrections fail: The persistence of political misperceptions. *Political Behavior* 32(2): 303–330.

O'Keefe DJ (2015) *Persuasion: Theory and Research*. Thousand Oaks, CA: SAGE Publications.

Shin D, Hameleers M, Park YJ, et al. (2022) Countering algorithmic bias and disinformation and effectively harnessing the power of AI in media. *Journalism & Mass Communication Quarterly* 99(4): 887–907.

Sundar SS (2008) The MAIN model: A heuristic approach to understanding technology effects on credibility. In: Metzger M and Flanagin A (eds) *Digital media, youth, and credibility*. Cambridge, MA: MIT Press, pp. 73–100.

Sundar SS (2020) Rise of machine agency: A framework for studying the psychology of human–AI interaction (HAII). *Journal of Computer-Mediated Communication* 25(1): 74–88.

Sunstein C (2018) # *Republic: Divided democracy in the age of social media*. Princeton, NJ: Princeton University Press.

Thorne J, Vlachos A, Christodoulopoulos C, et al. (2018). FEVER: A large-scale dataset for Fact Extraction and VERification. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies Volume 1 (Long Papers)*: 809–819.

Tsfati Y and Cohen J (2005) Democratic consequences of hostile media perceptions: The case of Gaza settlers. *Harvard International Journal of Press/Politics* 10(4): 28–51.

Walter N, Cohen J, Holbert RL, et al. (2020) Fact-checking: A meta-analysis of what works and for whom. *Political Communication* 37(3): 350–375.

Wang WY (2017) "Liar, liar pants on fire": A new benchmark dataset for fake news detection. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Short Papers)*: 422–426.

Wintersieck AL (2017) Debating the truth: The impact of fact-checking during electoral debates. *American Politics Research* 45(2): 304–331.

Wood T and Porter E (2019) The elusive backfire effect: Mass attitudes' steadfast factual adherence. *Political Behavior* 41: 135–163.

Yang K and Menczer F (2023) Large language models can rate news outlet credibility. *arXiv Preprint arXiv:2304.00228*. DOI: 10.48550/arXiv.2304.00228.

Zhang X, Cao J, Li X, et al. (2021) Mining dual emotion for fake news detection. *Proceedings of the Web Conference* 2021: 3465–3476.

Zittrain J (2014) Engineering an election. *Harvard Law Review Forum* 127(8): 335–341.

Author biographies

Je Hoon Chae (MS, Yonsei University) is a PhD student in the Department of Communication at the University of California, Los Angeles, and MS student in Department of Statistics and Data Science at University of California, Los Angeles.

David Tewksbury (PhD, University of Michigan) is Emeritus Professor in the Department of Communication at the University of Illinois Urbana-Champaign.