Media and Communication
2025 • Volume 13 • Article 9516
https://doi.org/10.17645/mac.9516

cogitatio

ARTICLE

Open Access Journal

AI Agency in Fact-Checking: Role-Based Machine Heuristics and Publics' Conspiratorial Orientation

Duo Lan [1], Yicheng Zhu [2], Meiyu Liu [2], and Chuge He [2]

[1] School of Digital Media and Design Arts, Beijing University of Posts and Telecommunications, China
[2] School of Journalism and Communication, Beijing Normal University, China

Correspondence: Yicheng Zhu (yicheng@bnu.edu.cn)

Submitted: 29 October 2024 **Accepted:** 6 March 2025 **Published:** 29 May 2025

Issue: This article is part of the issue "AI, Media, and People: The Changing Landscape of User Experiences and Behaviors" edited by Jeong-Nam Kim (University of Oklahoma) and Jaemin Jung (Korea Advanced Institute of Science and Technology), fully open access at https://doi.org/10.17645/mac.i475

Abstract

With a focus on role-based (fact-checker and author) agencies and machine heuristics conceptualized by the modality, agency, interactivity, and navigability model, this study examines the comparative effect of AI (vs. human) agencies in debunking conspiracy theory news. Using a 2 × 2 online experiment with 506 participants, the study explores how conspiratorial orientation influences different role-based AI agencies' relationships with machine heuristics, and therefore news credibility perception and corrective action intentions. Results reveal that AI (vs. human) role-based agencies have separate but also interaction effects on heuristic activation. Moreover, potentially because conspiratorial orientation originates from skepticism towards humans, AI fact-checkers can be associated with higher corrective action intention for individuals with high conspiratorial orientation by activating AI fact-checker's positive machine heuristics.

Keywords

artificial intelligence; conspiratorial orientation; conspiracy theory; fact-checking; machine heuristics

1. Introduction

The role of AI-generated content in misinformation has been widely studied, primarily focusing on AI's capabilities as a content creator (Xu et al., 2023). However, recent research has also explored AI's potential as a fact-checking agent with platforms increasingly adopting AI for moderation and verification tasks (Moon et al., 2023). The MAIN model (modality, agency, interactivity, navigability) offers a useful framework for understanding these dynamics, proposing that users perceive AI and human agencies differently based on their assigned roles in online interfaces (Sundar, 2008). In particular, AI and human agents in fact-checking and authorship roles can influence user responses through distinct cues and heuristics.

Research has explored the differential perceptual and intentional effects of AI versus human agencies when each serves in roles such as fact-checker (Banas et al., 2022) or author (S. Wang & Huang, 2024). Due to differences in thematic contexts, agent roles, and interaction types, studies showed different results. With sources of both roles (fact-checker and author) disclosed, we aim to extend existing work by examining how different role-based AI agencies (vs. human) are associated with news credibility perception and corrective action intention (i.e., the behavior of an individual attempting to address or counteract a perceived negative influence of media messages on others; Talwar et al., 2020).

Individuals' perceptual and intentional outcomes relating to misinformation are often politically motivated (Kahan, 2015), where confirmation bias and motivated reasoning drive user engagement with false narratives (Miller et al., 2016; Zhu, Fitzpatrick, & Bowen, 2024; Zhu, Xu, et al., 2024). Studies show that AI fact-checking can mitigate such biases, potentially reducing the influence of motivated reasoning linked to political identity (Moon et al., 2023; Wischnewski & Krämer, 2022). Conspiratorial thinking is related to political or cultural identity but is distinct in its underlying nature (Federico, 2022; Sutton & Douglas, 2020). Unlike political identity, conspiratorial orientation (CO) reflects a more generalized skepticism toward human intentions (Kim & Lee, 2024). This mindset could affect the relative impact of AI versus human fact-checkers as cues in online news interfaces. Consequently, we aim to explore whether CO conditions the differential effects of AI and human agencies, especially in their respective fact-checking and authorship roles, on corrective action within conspiracy theory contexts.

2. Literature Review

2.1. AI's Agency Cues and Positive/Negative Machine Heuristic

AI's Agency Cues and Positive/Negative Machine Heuristic AI as an agency (instead of a hidden or unseen algorithm) of fact-checker, author, or other types of sources has become increasingly explicit in online news (Chae & Tewksbury, 2024; Tulin et al., 2024). Disclosure of AI agency can have significant perceptual effects among news consumers for online information processing and evaluation. When AI acts and therefore is perceived as a fact-checker or author, it becomes the source of information about a news article. Some recent literature has explored the effect of AI agency on news reception and information processing by building on the MAIN model (Sundar, 2008).

The MAIN model proposes that agency provides important information cues on the human–computer interface, which can further influence information processing and credibility perceptions (Sundar, 2008). More specifically, the AI or machine-related cues may activate users' mental heuristics (mental shortcuts in the form of pre-determined evaluation about the cue) which facilitate information processing (Banas et al., 2022; Garrett et al., 2013; Molina & Sundar, 2024). Among different types of heuristics, machine heuristics refers to cognitive shortcuts that users apply when interpreting AI-driven content, allowing them to quickly assess the reliability or intent of machine-generated information based on pre-existing beliefs about AI's capabilities (Sundar, 2008). These heuristics are a set of prior beliefs about the nature of machines or automated programs such as AI. Based on users' prior engagement and experience with machines and AI, it can either be positive or negative (S. Wang & Huang, 2024).

Positive machine heuristic (PMH), characterized by perceptions of AI as objective, accurate, and unbiased, often leads users to trust AI's assessments more readily. This trust stems from the belief that AI operates without personal biases, fostering a sense of algorithmic impartiality that can influence users' willingness to engage with or accept information (Sundar & Kim, 2019). In contrast, negative machine heuristic (NMH) is driven by skepticism regarding AI's limitations, especially in tasks perceived as requiring human nuance or empathy. Activation of NMH corresponds to the perspective that AI is mechanistic or overly simplistic, leading to lower trust in AI-driven content, particularly on complex topics (Waddell, 2019). Either way, studies have found that AI cues can activate machine heuristics more strongly than human cues (i.e., when certain sources on the news interface are disclosed as human) in experimental settings (Banas et al., 2022; Molina & Sundar, 2024; Pareek et al., 2024).

2.2. Role-Based Agencies: AI as Fact-Checker and Author

AI can serve two distinct roles in a digital news interface: as a fact-checker or as an author. Existing research on AI's agency effects has largely examined these roles separately. Some studies focus on AI as a fact-checker. When PMH is activated, AI fact-checkers are generally perceived as objective and efficient, leading readers to trust the accuracy of flagged content—especially when clear, structured explanations are provided (Pareek et al., 2024; S. Wang, 2021). However, when NMH is activated, the visibility of AI fact-checking can lead to a responsibility shift, where readers feel less inclined to engage in corrective action and instead defer content verification to the AI (Bhandari et al., 2021). This diminished sense of personal responsibility may reduce users' willingness to challenge or verify AI's fact-checkers suggestions.

Similarly, studies on AI as an author (news producer) have reported mixed findings. When PMH is activated, readers may associate AI authorship with objectivity, perceiving AI-generated content as free from ideological bias (Sundar, 2008). However, when NMH is activated, readers may view AI-authored content as lacking depth and empathy, particularly in complex or sensitive topics like conspiracy theories (Graefe et al., 2018; Thurman et al., 2017; Wu et al., 2019). This perceived lack of complexity can reduce reader engagement, including their likelihood of verifying information. Because AI-generated content is often viewed as purely factual, users may default to surface-level trust, reducing their motivation to critically assess or scrutinize AI-authored material, particularly when AI authorship is explicitly disclosed (DeVerna et al., 2024). Therefore, in line with the MAIN model, we propose the following hypothesis:

> H1: Disclosure of AI agency (fact-checker or author) compared to human agency leads to significantly higher activation of machine heuristics.

Existing literature has examined the activation of positive and negative machine heuristics based on AI agency in fact-checking and authorship roles. However, as AI and automation become increasingly prevalent in online news processing, AI can fact-check both human- and AI-generated news, while AI-authored content can be fact-checked by either humans or AI, creating a reciprocal fact-checking dynamic.

Prior research on AI-human collaboration in fact-checking has shown that different combinations of human and AI agreement/disagreement influence user perceptions of both content credibility and news source trustworthiness. Banas et al. (2022) demonstrated that the activation of bandwagon versus machine heuristics depends on whether fact-checking judgments (true vs. false) are aligned between AI and human

sources. In other words, the activation of a particular heuristic is not independent but contingent on contextual cues and the interaction among them.

Building on this idea, we consider two possibilities for heuristic activation in fact-checking and authorship roles. The first is that activation of role-based heuristics may be stronger for the fact-checker role because fact-checkers act as supervisors or evaluators of authored content. Therefore, the fact-checker's agency (AI vs. human) may influence not only fact-checker-based PMH and NMH but also those associated with authorship. For the second possibility, if one AI role (fact-checker or author) provides the context for heuristic activation in the other role, certain agency-role combinations may significantly amplify or suppress PMH and NMH. For example, when a human fact-checker debunks AI-generated news, it may trigger higher skepticism (NMH for AI author), depending on how users perceive the disadvantage of AI fact-checkers for news with complex socio-political backgrounds.

Prior research also illustrates that if an AI fact-checks human-authored conspiracy theory content, users may more readily trust the correction due to the perceived objectivity and distance AI brings as an external reviewer (S. Wang, 2021). Conversely, when AI serves as both fact-checker and author, this dual presence may prompt readers to engage less critically, as they might assume that the information has been pre-vetted by a "neutral" entity. However, with a human author, AI's fact-checking might instead serve as a reinforcing agent, encouraging readers to perceive the content through a lens of human insight balanced by AI's impartial validation (Horne et al., 2020).

Despite the abundance of prior research, there remains a lack of firm evidence of the exact direction of the interaction between different role-based AI vs. human agencies (as fact-checker and author), therefore we propose the following research question:

> RQ1: How do agency (AI vs. human) and role (fact-checker vs. author) interact in activating PMH and NMHs (fact-checker-based and authorship-based)?

2.3. Perceptual and Intentional Effects of AI (vs. Human) Role-Based Agencies

Prior research has studied both perceptual and intentional outcomes of AI vs. human agency in the two roles (fact-checker and author) concerning the current study. For instance, news credibility perception is expected to be modified as are intentional outcomes such as support for restrictions. However, little research has examined the effects of corrective action (i.e., the behavior of an individual attempting to address or counteract a perceived negative influence of media messages on others), which is arguably a desirable outcome of fact-checking.

By leveraging different heuristics, fact-checkers tend to have a significant impact on perceived credibility and quality evaluation of the news content (often fake news or misinformation), regardless of being human/crowdsourced or machine/AI. However, the differential effects brought by AI vs. human fact-checkers are less clear. For instance, Lee and Bissell (2024) found that human and AI interventions do not differ in their effects on readers' belief in misinformation about Covid-19 vaccination. Chae and Tewksbury (2024) reported that knowledge of AI intervention does not hinder the effectiveness of fact-checking labels compared to human fact-checkers. AI's fact-checkers differential effects are more

pronounced for behavioral intentions. For instance, AI fact-checkers or content moderators have an inferior effect than humans on encouraging support for regulation/censorship (Moon et al., 2023) and flagging (Bhandari et al., 2021), as well as reducing the likelihood of information forwarding (i.e., sharing the news; DeVerna et al., 2024).

However, machine authorship's effect seems to be less consistent, as shown in two meta-analyses (Graefe & Bohlken, 2020; S. Wang & Huang, 2024). Earlier studies have illustrated that AI authorship reduces hostile media bias (Cloudy et al., 2023; Craig & Choi, 2024) or slightly enhances the perceived credibility of the message (Kreps et al., 2022); however, more recent studies have found that AI authorship reduced perceived credibility and quality of the message (Jia et al., 2024). In other studies, machine authorship (vs. human) has no significant effect on perceptual outcomes such as credibility, news quality evaluation (Graefe & Bohlken, 2020), or other context-specific perceptions (e.g., how enjoyable, funny, or trustworthy, etc.; Rae, 2024). While Graefe and Bohlken (2020) found conflicting results from experimental designs (human authorship is considered better) and descriptive designs (machine authorship is considered better), Wang and Huang (2024)'s analysis showed a general, but slight, disadvantage in credibility perception when authorship is attributed to automated agents. As for behavioral intentions, AI authorship is found to have only marginally negative effects on information-forwarding behavior (re-sharing the message online; Rae, 2024).

Prior research presents mixed findings regarding AI agency's effects on news credibility and behavioral responses, such as information forwarding and more restrictive actions like support for content regulation. Corrective actions, such as advising others on misinformation, are influenced by how users perceive a message's personal and social impact. AI agency may shape users' evaluations of both fact-checkers and authors, influencing whether a message is seen as socially acceptable or problematic. Specifically, fact-checking warnings and author credibility cues may determine how users assess the reliability of the content and their willingness to take corrective actions. As such, we propose the following research question:

> RQ2: When both roles are shown on the news interface, how are different role-based AI agencies (vs. human) associated with news credibility perception and corrective action intention?

Prior research has also demonstrated that machine heuristics can act as mediators in various behavioral responses. For example, PMH has been shown to mediate trust in automated decision-making, where users may accept machine-generated outcomes without critical scrutiny (Binns et al., 2018). Similarly, NMH can mediate user engagement in contexts requiring high levels of personal involvement or moral judgment, as users tend to question the AI's depth and accuracy in such areas (Graefe et al., 2018). These heuristic responses are particularly relevant in AI's roles as fact-checker and author, where PMH and NMH may influence the extent to which users take corrective action based on the perceived credibility or depth of AI's input.

Molina and Sundar (2024) found that such PMH reinforces a responsibility shift, where users defer to AI's perceived authority, reducing their personal engagement in corrective actions when AI's fact-checking role is visible. Conversely, NMH may be more prevalent when AI is labeled as an author, as users may question the credibility and depth of AI-authored content. This skepticism can lead to reduced corrective engagement, particularly for complex topics like conspiracy theories, where readers may perceive AI as incapable of nuanced

expression (Waddell, 2019).

Therefore, we propose that AI as a fact-checker or author can uniquely shape corrective action, directly or through the mediation of machine heuristics. Therefore, the following hypotheses are proposed:

> H2: Machine heuristic (fact-checker role) mediates the AI fact-checker agency's comparative effect against human on (a) news credibility and (b) corrective action intention.
>
> H3: Machine heuristic (author role) mediates the AI author agency's comparative effect against human on (a) news credibility and (b) corrective action intention.

2.4. Conspiratorial Thinking and CO

Conspiracy theory news is a specific type of misinformation (Kim & Lee, 2024). In this context, past research has explored the potential of AI technologies in identifying and categorizing misinformation and fake news (Jahanbakhsh et al., 2023). In the meantime, researchers have also explored whether disclosure of AI's role as the fact-checker would be perceived as reliable and trustworthy (Molina & Sundar, 2024). On the perceptual level, these studies explored whether AI as a fact-checking source can achieve better or worse effects in comparison to human fact-checking (Banas et al., 2022; Moon et al., 2023). These studies focus on different outcomes of AI vs. human fact-checking, but a common finding is that AI agency has an advantage over humans when the message source or content invites motivated reasoning: a way of reasoning with the purpose of identity protection or with the preference of pre-existing beliefs towards controversial social issues. In comparison to human fact-checker debunking misinformation, AI's fact-checkers agency leads to less perceived hostile media effect (Cloudy et al., 2023), or reduces the extent to which partisans adore in-group misinformation (Moon et al., 2023; Moon & Kahlor, 2025).

In the context of debunking fake news or misinformation, audience responses to misinformation are often shaped by their pre-existing beliefs about a particular story or by alignment with narratives that reflect their personal identity (Hameleers & van der Meer, 2019). Research indicates that individuals who believe in conspiracy theories often share specific cognitive patterns that make them more prone to accepting such theories (Romer & Jamieson, 2022). In this light, Kim and Lee (2024) conceptualized CO, which refers to an individual's tendency to interpret information through a lens of distrust toward mainstream narratives, often involving beliefs in hidden agendas and manipulation by powerful entities. Those with high CO are generally inclined to believe in conspiracy theories. For behavioral intentions, they view corrective interventions with suspicion, particularly when they perceive these as efforts by traditional institutions to control the narrative (Tam & Lee, 2024).

This predisposition to skepticism is rooted in the perceived power imbalance between the subjects of conspiracy theories, the sources of information, and the audience receiving the information. Since conspiratorial thinking functions as a type of quasi-problem-solving, this skepticism is heightened when individuals face a prolonged lack of power, resources, and access to solutions (Kim & Lee, 2024). Compared to humans, AI fact-checkers and AI authors (as communicative agents) may offer cognitive shortcuts that could either promote or hinder this quasi-problem-solving process by acting as a third-party influence on information processing. Additionally, the perceived power distance and resourcefulness of human and AI

communicators may vary between individuals with a high level of conspiratorial thinking and those without such orientation.

Individuals with high levels of CO are likely to view debunkers of conspiracy theories in a negative light. They may see human fact-checkers as powerful entities attempting to challenge their beliefs while perceiving AI fact-checkers as comparatively less biased. For high CO individuals, the idea that AI fact-checkers are neutral may be more readily accepted than the idea that human fact-checkers offer context and deeper understanding. This preference for AI may stem from the inherent skepticism toward human intentions associated with conspiracy theories themselves (Frenken & Imhoff, 2023; Imhoff & Bruder, 2014). Therefore, high CO individuals are likely to activate PMH for AI fact-checking, interpreting it as an objective, rule-based system that lacks hidden motives (Sundar & Kim, 2019). This belief in AI's impartiality can lead high CO readers to respond more favorably to AI fact-checking than to human intervention, potentially fostering corrective actions by reducing suspicion. In contrast, these individuals often view human fact-checkers as part of a larger agenda to suppress alternative viewpoints, which could heighten skepticism and diminish the effectiveness of corrective measures when presented by humans.

On the other hand, low CO individuals—those with less inclination to believe in conspiracy theories—are less likely to be skeptical of human fact-checkers. For individuals with lower CO, human fact-checking may reinforce a social norm of collective responsibility in countering misinformation (Gimpel et al., 2021), or activate a perceptual affinity or trust toward human expert fact-checkers. This trust could stem from both authority and machine heuristics, reflecting an inherent confidence in the agent's reliability, regardless of whether the agent is human or AI (Vraga & Bode, 2017; Y. Wang, 2021). Therefore, according to the two-step motivated reasoning model, these could enhance their willingness to engage in corrective actions (Jennings & Stroud, 2023; Liu et al., 2023). The endorsement by a human fact-checker can be perceived as socially responsible and contextually aware, aligning with low CO individuals' trust in mainstream narratives. In this case, PMH is less likely to be activated for AI fact-checkers, as low CO individuals may prefer human intervention, especially for socio-political topics, due to the perceived depth and empathy of human understanding.

When AI or human agencies serve as authors, CO also moderates reader responses, though with different heuristic effects. High CO individuals may activate NMH when AI is the author, perceiving AI-authored content as overly mechanistic and incapable of capturing the complexity of conspiracy theories (Waddell, 2019). This skepticism may limit their engagement with corrective actions, as they question the quality and depth of AI-authored content. Conversely, if a human is the author, high CO individuals may still maintain suspicion, interpreting human-authored content as potentially biased (S. Wang & Huang, 2024). For low CO individuals, human authorship is likely to foster trust, as they value the social accountability associated with human authors. AI-authored content, while perceived as objective, might lack the relational depth that low CO individuals expect, making human-authored interventions more effective for promoting corrective actions.

In sum, CO can potentially moderate the influence of fact-checking and authorship agencies by shaping whether positive or NMH are activated in response to AI and human interventions (therefore activating indirect or direct pathways). In light of this review, we proposed the following hypotheses:

H4: CO moderates the indirect effect (through PMH and/or NMH) and the direct effect of fact-checker-role-based AI agency (vs. human) on (a) perceived credibility and (b) corrective action intention.

H5: CO moderates the indirect effect (through PMH and/or NMH) and the direct effect of author-role-based AI agency (vs. human) on (a) perceived credibility and (b) corrective action intention.

3. Method

3.1. Sample and Sampling Method

To address the research questions and hypotheses, we conducted a 2 (fact-checker source: AI, human) by 2 (author source: AI, human) between-subjects online experiment. We adapted two real-world online news articles containing conspiracy theories as the stimuli of the experiment: one about the cause of an airplane crash that happened in China and the other one about Pfizer's alleged role in mutating the Covid-19 virus for profit.

In regards to the sampling method, the sample for this study was recruited by a major Chinese online panel provider wjx.com (问卷星) using quotas mimicking those of the adult population in Beijing city in terms of age, gender, and education from the sixth national population census (National Bureau of Statistics of China, 2018). Survey invitations were sent to existing randomly selected representative panels of Beijing residents. Thereafter, participants entered the survey experiment procedure. The experiment was performed online between September 9–28, 2024, and from January 13–19, 2025, with no repetitive participants. Detailed demographic information can be found in Table 1.

3.2. Experimental Design

Survey participants from the study panel were randomly assigned by the survey system to one of the four experimental conditions. Participants will first answer a series of questions that can be related or unrelated to the independent variables (e.g., nationalism, prior knowledge, etc.) and are universal across conditions. Then participants in different conditions will be given different message stimulus.

The stimuli are pictures designed to mimic a social media post (see the Supplementary Material for translated articles) showing adapted news articles: one about the cause of the China Eastern Airline 5332 crash and the other one about Pfizer's alleged role in mutating Covid-19 virus for profit. Each news article was shown to half of the participants. The first promotes a conspiracy theory against the airplane manufacturer (Boeing) and its connections with the US government, the second one implies Pfizer has been continuously conducting gain-of-function research to mutate the Covid-19 virus to sell medications. While the title, content, account name, and other features of the article remain the same, it has four different versions with different combinations of fact-checking sources (AI vs. human) and author sources (AI vs. human journalist). The different combinations of the labels can be seen in the Supplementary Material.

Participants are randomly assigned to four experimental conditions. The only differences between these randomized experimental conditions are the different versions of fact-checking and author source labels shown to the participants. For example, in the "AI fact-checker–human author" condition, the author icon

Table 1. Sample demographics ($N = 506$).

	N	%
Age		
18–24	57	11.3
25–34	218	43.1
35–44	164	32.4
45 and above	64	12.6
Gender		
Male	255	50.4
Female	251	49.6
Education		
High school and below	140	27.7
Bachelor's degree	296	58.5
Master's degree	65	12.8
Doctoral degree	5	1
Family Monthly Income (Chinese ¥)		
Less than 5,000	16	3.2
5k–20k	245	48.4
20k–50k	200	39.5
50k–100k	27	5.3
More than 100k	18	3.6

will show a human face with the fictional name of the journalist, and the fact-checking label will show "our AI algorithm suggests that this article may contain unverified information."

3.3. Independent Variables

In factor I—AI vs. human agency (fact-checker role)—a dichotomous variable is created to represent participants' assignment into the AI or human fact-checker groups. It uses indicator coding to represent the groups in this factor (AI fact-checker = 1, and human fact-checker = 0), therefore, the effects and coefficients in Section 4 show the influence brought by an AI fact-checker.

In the same rationale, for factor II—AI vs. human agency (author role)—a dichotomous variable uses indicator coding to represent the groups in this factor (AI author = 1, and human author = 0).

In regards to CO, we follow Kim and Lee's (2024) conceptualization and suggested measurements. The measure includes three dimensions (i.e., conspiratorial realism, susceptibility to popular folklore, workplace conspiratorial realism) and 11 items in total. Each item is measured on a 7-point Likert scale (1 = *absolutely disagree*, 7 = *absolutely agree*). An example item of CO included "those people in power will use shadowy means to gain profit or advantage rather than lose it" ($M = 4.62$, $SD = 1.18$, Cronbach's $\alpha = .92$).

3.4. Dependent Variables

The first variable is PMH and NMH in a fact-checker role. The assessments of machine heuristics are based on the conceptualization and operationalization by Sundar (2020) and Molina and Sundar (2024). It includes four 5-point Likert scale items measuring fact-checker-role-based PMH-C ("C" stands for "checker"): "the *fact-checker* in the news you just read" followed up by "has machine-like precision," "is error free," "has machine-like accuracy," and "has machine-like objectivity" (M = 3.24, SD = .88, α = .83). Those measuring NMH (NMH-C) included "the fact-checker in the news you just read" followed up by items such as "is able to detect human emotion" (reverse coded), "is mechanistic," "is able to understand contextual background" (reverse coded), and "lacks human intuition" (M = 2.80, SD = .93, α = .80).

The second variable is PMH and NMH in an author role (PMH-A and NMH-A; the suffix "A" stands for "author"). Because there are two source agencies in our experiment (fact-checker and author), we also measured machine heuristics for the author role with the same items above. However, the leading sentence was changed to "the *author* of the news you just read." PMH-A (M = 3.12, SD = .94, α = .84) and NMH-A (M = 2.73, SD = .93, α = .80) for author role also has good reliability.

The third variable relates to news credibility perception. We adapted Flanagin and Metzger's (2000) measurement of internet information credibility, with three items asking respondents if they perceive the news to be "credible," "accurate," and "biased" (reversed coded). Items were measured on a 7-point Likert scale (1 = *absolutely disagree*, 7 = *absolutely agree*) and have good reliability (M = 4.69, SD = 1.11, Cronbach's α = .83).

The last variable concerns corrective action intention. We adapted Talwar et al.'s (2020) measurement of active fake news corrective action, which included three items asking for agreement: "If my friends share this kind of news, I will try to correct their views," "if I see this kind of news on social media, I will comment to oppose its content," and "I will search for authoritative information, in order to rectify misunderstandings about the news among other people." The items of the scale (M = 3.08, SD = .95, Cronbach's α = .82) were measured on a 5-point Likert scale (1 = *absolutely disagree*, 5 = *absolutely agree*).

3.5. Analytical Strategy

Data analysis was performed in SPSS® (Version 26.0). To investigate hypothesized main effects and interaction effects on machine heuristics, credibility, and corrective action intention (as probed by H1, RQ1, and RQ2), we use univariate general linear models for analysis. Given that the targets of attribution of responsibility of the conspiracy theory in our stimuli are foreign entities, we controlled for nationalism hoping to mitigate this limitation to some extent. We also controlled pre-existing knowledge and demographic variables (age, gender, education, and income level; Jia & Luo, 2023). Given the strong relationships between the conspiratorial thinking mechanism and the news article's perceptual and behavioral effect (Kim & Lee, 2024), we also controlled CO and its two-way interaction terms with the two factors, which will be further explored by H4 and H5. Given this setup, a priori estimation of the required sample size using G-Power suggests 500 for a small effect size (.01) with a statistical power of .90.

To investigate hypothesized simple mediation effects on dependent variables (H2 and H3), the bootstrap method with an SPSS application (PROCESS, model 4) provided by Hayes (2015) was used. To test the moderated mediation hypotheses (H4 and H5), Model 8 in PROCESS was used. Inference regarding moderated mediation is assessed using Hayes (2015) index of moderated mediation; 5,000 bootstrap samples were specified to generate bias-corrected CIs. Given Model 8's set-up, a priori estimation of the required sample size using G-Power suggests 132 for medium effect size (.10) with a statistical power of .95.

4. Results

4.1. Manipulation Check

After the participants have seen the stimuli, they will be asked a single question: "In the news article you just read, who assisted in verifying the content of the article?." The participants will be provided with a multiple-choice question with answers corresponding to the two types of fact-checkers: "AI" and "human." Another question asks, "In the news article you just read, who is the author of the article?" and provides two choices: "AI journalist" and "Haibo Wang" (the fictional name of the human journalist). For each of the four conditions, more than 90% of the respondents correctly specified the condition they were assigned to. Data of those who failed were obtained for further analyses.

4.2. H1, RQ1, and RQ2: Main Effects and Interaction Effects

To test H1 and answer RQ1 and RQ2, a 2 (Fact-checker Source: AI vs. human) by 2 (Author Source: AI vs. human) MANCOVA was conducted with role-based machine heuristics (PMH-C, NMH-C, PMH-A, NMH-A), perceived news credibility and corrective action intention as dependent variables, followed by separate t-tests.

The MANCOVA revealed a significant main effect of fact-checker agency on NMH-C (F (1, 488) = 23.84, $p < .001$, $\eta_p^2 = .05$, $M\_AI = 3.30$, $M\_Human = 2.30$). Fact-checker agency also significantly influenced NMH-A (F (1, 488) = 4.28, $p < .001$, $\eta_p^2 = .01$), though pairwise comparisons did not indicate a significant mean difference. Author agency also had a significant main effect on NMH-A (F (1, 488) = 9.36, $p = .002$, $\eta_p^2 = .02$, $M\_AI = 3.25$, $M\_Human = 2.23$), suggesting that AI authors were associated with higher skepticism. These findings support H1.

For RQ1, a significant interaction effect emerged between fact-checker and author agency on PMH-A (F (1, 488) = 58.03, $p = .04$, $\eta_p^2 = .01$). Post-hoc comparisons revealed that the AI fact-checker and AI author combination led to the highest PMH-A ($M = 3.50$), while human author presence, regardless of fact-checker type, resulted in lower PMH-A ($M\_HumanAuthor/HumanChecker = 2.90$, $M\_HumanAuthor/AIChecker = 2.75$; see Figure 1). No significant interaction effects were found for NMH-C, NMH-A, or PMH-C.

For RQ2, fact-checker agency had a significant main effect on corrective action intention (F (1, 488) = 4.47, $p = .03$, $\eta_p^2 = .01$, $M\_AI = 3.09$, $M\_Human = 3.06$), but there is no significant effect on news credibility. AI (vs. human) agency in the author role did not significantly influence news credibility and corrective action intention, no significant interaction effects were found for these outcomes.

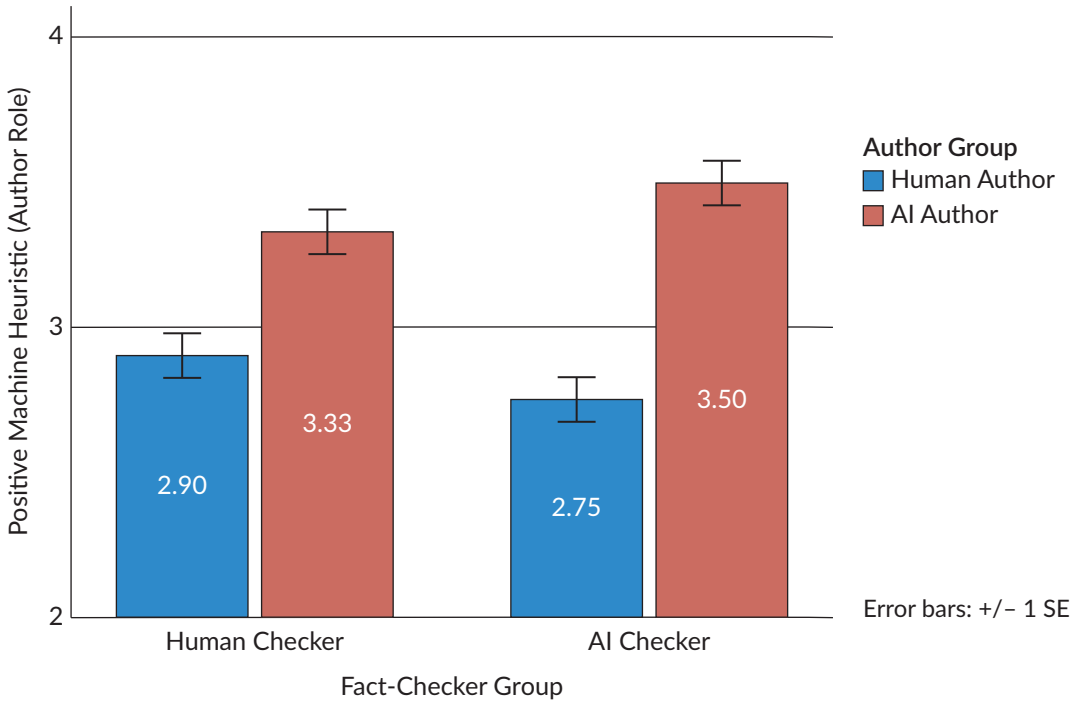

Figure 1. Fact-checker–and author-role-based AI agency (vs. human) on PMH-A.

The results of the MANCOVA also show that interaction between CO and fact-checker agency influences PMH-C (F (1, 488) = 4.50, p = .04, η_p^2 = .01), which indicates that the effect of fact-checker agency on PMH-C is dependent on CO levels (this will be explored in H4). Pair-wise comparisons indicate a significant mean difference of PMH-C between fact-checker groups (M_AI = 3.49, M_Human = 2.98).

4.3. Mediation Analyses

4.3.1. H2: Simple Mediation of Fact-Checker's Source Effect

As shown in Figure 2, results for H2a indicated no significant direct or indirect effect on the news credibility perception. However, results of H2b showed a significant indirect effect through PMH-C (effect = .15, SE = .04, 95% CI [.08, .22]), indicating that AI fact-checkers (vs. human) increased corrective action intention via PMH-C. However, the indirect effect through NMH-C was not significant (effect = −.03, SE = .06, 95% CI [−.14, .09]).

4.3.2. H3: Simple Mediation of Author's Source Effect

Results of H3a (illustrated in Figure 3) showed a significant indirect effect through PMH-A (effect = .18, SE = .05, 95% CI [.09, .27]), indicating that AI authors (vs. human) increased news credibility when PMH-A is activated. Similarly, NMH-A significantly mediated the effect in the opposite direction (effect = −.21, SE = .07, 95% CI [−.35, −.06]), suggesting that AI authors can also trigger negative heuristics that lowered credibility. Results of H3b showed a significant indirect effect through NMH-A (effect = .14, SE = .06, 95% CI [.01, .26]), indicating that AI authors (vs. human) increased corrective action intention via NMH-A. However, the indirect effect through PMH-A was not significant.

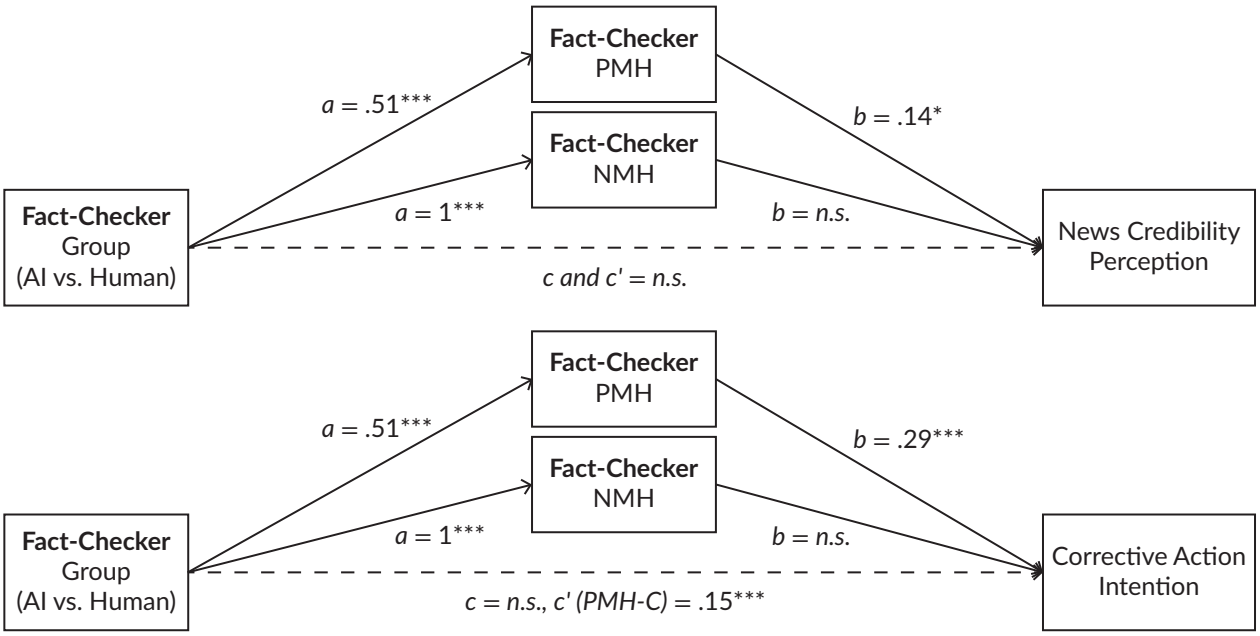

Figure 2. Simple mediation of AI (vs. human) fact-checker agency's effect. Notes: Mediating effects of PMH, NMH; news credibility perception as a dependent variable (top); corrective action intention as dependent variable (bottom); *** $p < .001$, ** $p < .01$, * $p < .05$.; c' = direct effects of agency type on dependent variables; c = total effect of agency type; n.s. = not significant.

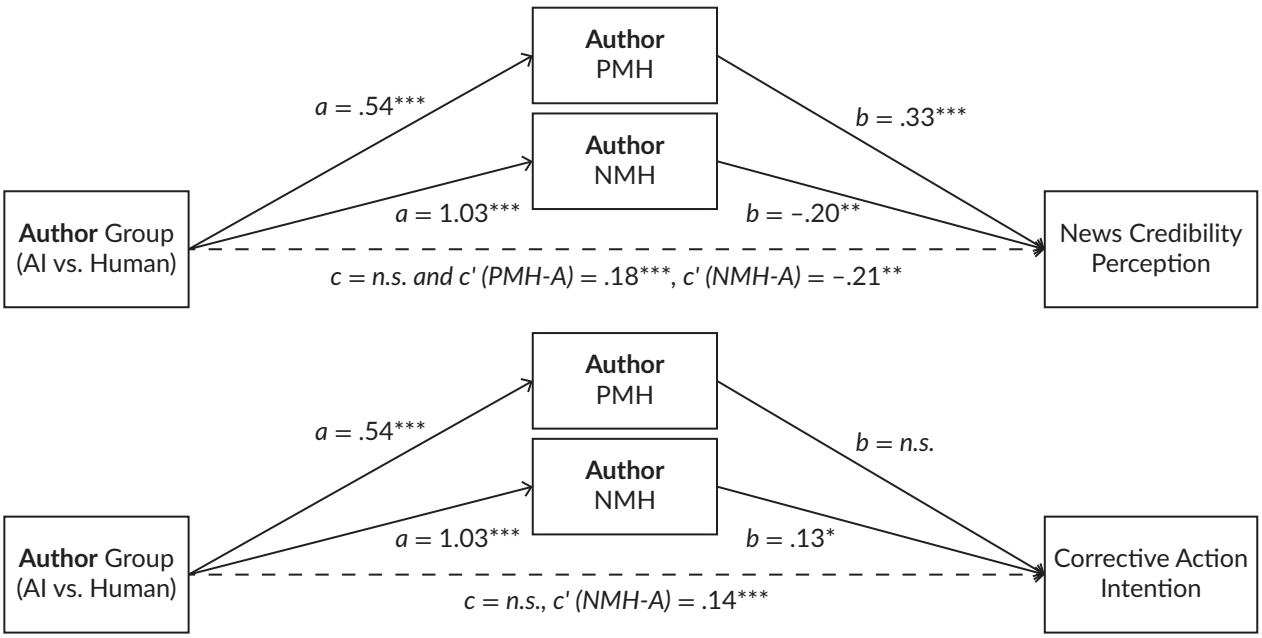

Figure 3. Simple mediation of AI (vs. human) author agency's effect. Notes: Mediating effects of PMH, NMH; news credibility perception as a dependent variable (top); corrective action intention as a dependent variable (bottom); *** $p < .001$, ** $p < .01$, * $p < .05$; c' = direct effects of agency type on dependent variables; c = total effect of agency type; n.s. = not significant.

4.4. Mediation Analyses Moderated by CO

4.4.1. H4: Moderation on the Effect of Fact-Checker Role-Based AI Agency

For H4a, results suggest that CO does not moderate the indirect or direct effect of AI (vs. human) fact-checker agency on perceived news credibility. For H4b, results showed a significant moderated mediation effect through PMH-C (effect = .12, *SE* = .05, 95% CI [.06, .19]). For 90.1% of the participants who have CO > 2.72, higher CO strengthens the mediation effect on corrective action intention (Figure 4a). However, the indirect effect through NMH-C was not significant. For the conditional direct effects on corrective action intention, AI had a significant disadvantage to human fact-checker for inducing corrective action intention at lower levels of CO, however, this effect became non-significant at higher levels of CO (for 68.9% of the participants who have CO > 4.32; Figure 4b).

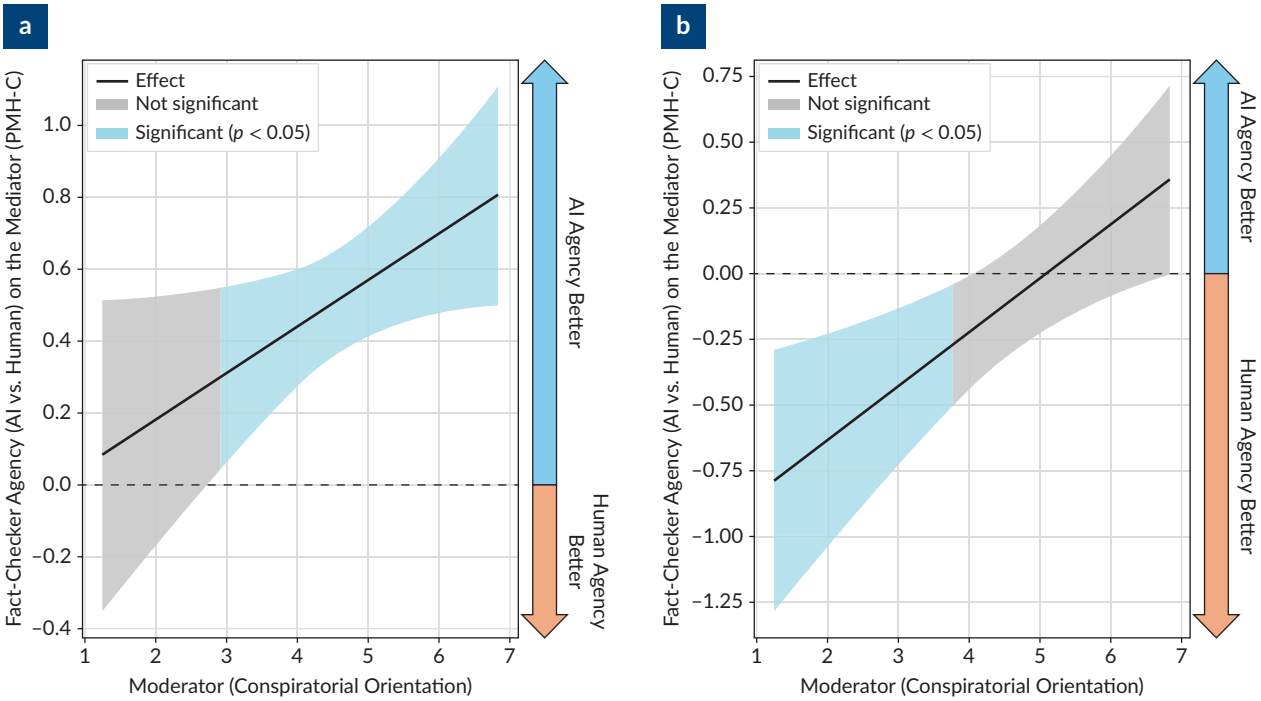

Figure 4. Effects of AI vs. human fact-checker on PMH-C (a) and corrective action intention (b).

4.4.2. H5: Moderation on the Effect of Author Role-Based AI Agency

For H5a and H5b, we used Model 8 again to test if CO moderates the indirect and direct effects examined by H3a and H3b. Results show that while there were some marginal trends, no significant mediation or moderation effects were found for H5b and H6b, suggesting that the proposed mechanisms did not hold for AI authorship's indirect and direct effects.

5. Discussion

5.1. Role-Specific and Cross-Role Effects on Machine Heuristics and Dependent Variables

AI agency and machine heuristics' relationship is associated with the specific role that AI is playing. In line with prior studies, we find that AI author influences author-role-based machine heuristics (Cloudy et al., 2023; Craig & Choi, 2024; Molina & Sundar, 2024; S. Wang & Huang, 2024), and AI fact-checker has an effect on the fact-checker-role based machine heuristics (Banas et al., 2022; Moon & Kahlor, 2025; Tulin et al., 2024; S. Wang, 2021). Our results from RQ1, however, extend existing results by illustrating the possibility that different role-based agencies have an interaction effect in activating PMH of the AI author agency.

The result of this significant interaction can be interpreted together with AI fact-checker agency's activation of NMH-A (NMH for AI's author role). Potentially, when AI appears both as a debunking fact-checker and author, participants' prior negative belief about AI as the fact-checker (NMH-C) overshadows their negative views on AI as the author (NMH-A), making their prior positive beliefs on AI as the author (PMH-A) more salient. However, we acknowledge that the AI fact-checker–AI author combination is currently uncommon for online news interfaces. Nonetheless, with AI's dual roles becoming increasingly more prominent in both fact-checking and news production, the possibility of encountering such circumstances exists in the future.

In line with prior research, we identified significant effects of AI fact-checker-role-based agency on behavioral intentions against fake news (Bhandari et al., 2021; Moon et al., 2022), which in our case is corrective action intention. However, our results showed that AI agency (either fact-checker or author role) is not associated with different levels of news credibility perception compared to when these roles are played by humans. This result does not surprise us given marginal and situational results as summarized by Graefe and Bohlken (2020) and Wang and Huang's (2024) meta-analyses.

5.2. Mediation Effects of Distinct Role-Based Machine Heuristics

Our mediation analysis shows that the AI agency's advantage or disadvantage compared to human agency in its relationship with news credibility perception and corrective action intention is contingent upon two things. The first is that they are dependent on the activation of a corresponding machine heuristics: by comparing results from the MANCOVA and the mediation analyses, we noticed that activation of machine heuristics is critical in determining whether AI agency created any difference from human agency. Although the MANCOVA result does not support AI agency's direct relationship with the perceptual dependent variable (news credibility perception), there are significant indirect, mediated relationships when author-role-based machine heuristics are activated.

Secondly, our results support the idea that author-role-based machine heuristics have more pervasive mediating effects on news credibility and corrective action, while fact-checking-role-based machine heuristics target intentional outcomes more specifically. This is also in accordance with prior studies along separate lines, but our results provide a comparison when agencies of both roles are disclosed on the news interface: Mediation effect exists for news credibility perception when either one of the author-based machine heuristics are activated (Figure 3, top), not when fact-checker-role-based machine heuristics are (Figure 2, top). For corrective action intention, AI agency is associated with higher intentions

than humans when the positive fact-checker-role-based machine heuristics is activated and when negative author-role-based machine heuristics is activated.

5.3. Debunking Conspiracy Theory News: Does AI Agents Matter?

In the case of debunking conspiracy theory news, is AI (vs. human) fact-checker agency associated with more corrective action intention? Our findings suggest that the answer depends on CO levels. The results from the MANCOVA illustrate that news-specific prior beliefs, such as CO in the context of conspiracy theory news, moderate the activation of individuals' prior beliefs about AI as a more precise, error-free, accurate, and objective news fact-checker (in our case, PMH-C difference between AI and human agents). As a continuation of this finding, moderated mediation analyses show that CO emerged as a significant moderator, influencing both the effect of AI vs. human fact-checking agency on corrective action intention and the mediation of such effect through PMH-C.

From the perspective of motivated reasoning, this effect is not surprising, as prior studies have shown that existing beliefs, political inclinations, or ideological orientation (Walter et al., 2020) moderates the effect of misinformation fact-checking. They also interact in the specific domain of AI vs. human fact-checking agency moderating their comparative relationship with misinformation debunking outcomes, such as hostile media effects across partisan lines (Cloudy et al., 2023), or preferences on in-group over out-group fake news (Moon et al., 2023; Moon & Kahlor, 2025). Fact-checkers labeled as "AI" are found to be perceived as "apolitical" compared to human expert fact-checkers, and therefore induce less mistrust against the fact-checking message caused by partisan or ideological preferences (Chung et al., 2024).

However, different from partisanship or hostile media effects, the moderating effect of CO in this study may not have stemmed from an identity-protection motivation (Kahan, 2015; Moon & Kahlor, 2025). If it was, then the analytical focus would be CO's negative relationship with the activation of positive beliefs (PMH) about both AI (non-significant negative correlation) and human (Pearson correlation $r = -.13$, $p < .05$) debunkers. In the current study, individuals with high CO levels are not conceptualized to share a "conspiracy-theory-lover" identity, but rather a common distrust in powerful entities and disgust of power imbalance (Kim & Lee, 2024). A distinctive feature of CO to partisanship is that it is not context dependent. Rather, it represents a long-standing inclination toward skepticism about human motives and intentions. Therefore, in the current study, such skepticism, rather than an identity-protection motivation against any debunkers of conspiracy theory news, was conceptualized and examined as the motivator of differential evaluation of the AI (vs. human) fact-checker agency.

Current results support this idea. We witness a stronger activation of PMH-C by AI (vs. human) fact-checker agency among individuals with high levels of CO (Figure 4a). Because the higher the CO, the less PMH-C (good qualities of a fact-checker) was attributed to human fact-checker: one who potentially holds certain governmentally or organizationally imposed fact-checking agenda. Conversely, individuals with lower levels of CO do not view AI fact-checker agency (vs. human) as more qualified. Moreover, similar Moon and Kahlor's (2025) findings, when the author-based agency is controlled (in the mediation models), our results indicate that AI fact-checking agency (vs. human) is associated with a poor fact-checking result (less corrective action intention in Figure 4b), but only for individuals with lower levels of CO. As CO increases, AI (vs. human) fact-checker agency's lower direct association with correct action intention ceased to be

significant at CO = 4.32. Taking these results together, it is plausible that higher CO activates an AI-fact-checker-centered PMH, therefore activating a positive mediation for AI fact-checker's and a comparatively stronger association with corrective action intentions than human fact-checkers.

While prior research has largely focused on the activation of machine heuristics to explain responses to AI versus human fact-checking, our study extends this framework by exploring how fundamental psychological traits like CO shape the activation of machine heuristics. Specifically, our findings suggest that CO can influence whether individuals apply machine heuristics to AI or human agents. This indicates that the application of machine heuristics—whether positive or negative—is not exclusively linked to AI. Instead, individuals may attribute machine-like characteristics (e.g., objectivity, neutrality) to human agents if they view humans as more competent in certain roles.

5.4. Practical Implications

Our findings emphasize the importance of considering CO when designing fact-checking interventions. For individuals with high CO, AI fact-checkers are perceived as more objective and neutral, making them a more effective tool for promoting corrective action intentions. In contrast, human fact-checkers may be more trusted by those with lower CO who value relational cues and nuanced judgment. Platforms should consider segmenting audiences by CO levels to tailor interventions, using AI fact-checkers for those with high CO and human fact-checkers for others.

Moreover, platforms should adopt public segmentation strategies to address high CO individuals who may be more susceptible to conspiracy theories but would trust AI agency more than human. Insights from this study suggest that interventions based on AI's neutrality could be more effective for these users, especially in environments where conspiracy theories are rampant. Delivering fact-checking content through AI might reduce the resistance these users have toward corrective messages and mitigate the spread of misinformation, ultimately fostering a more informed and engaged user base.

Our findings also show that PMH-C is activated for users with high CO, especially when AI is used as a fact-checker. News and social media platforms can leverage this by incorporating AI-driven fact-checking interventions that resonate with users' preferences for neutrality and objectivity. At the same time, human-based fact-checking can be better suited for addressing users with lower CO who are more likely to engage with human-authored content. This adaptive approach to messaging can help increase engagement with fact-checked content and promote corrective actions, ultimately enhancing the credibility of news sources and reducing the spread of misinformation.

5.5. Limitations

This study has several limitations. First, the sample was skewed toward a younger population, with limited representation of older participants. This may have influenced responses to AI and human fact-checking agencies, as younger individuals may engage differently with technology. Future studies should aim for a more balanced demographic representation to assess how age influences these perceptions.

Second, while we explored two-way interactions between agency type (AI vs. human) and CO, more complex interactions, such as three-way interactions involving fact-checker agency, author agency, and CO, were not investigated. Exploring these interactions could offer deeper insights, though interpreting such models would present significant challenges.

Lastly, factors like the third-person effect or social desirability bias, which can influence corrective actions, were not examined in this study. Incorporating these factors in future research could provide a more comprehensive understanding of the drivers behind corrective behavior, particularly in the context of AI and human fact-checking agencies.

Acknowledgments

We thank the three anonymous reviewers for their insightful feedback and constructive suggestions, which greatly improved the quality of this manuscript. We are also grateful to the thematic issue editors for their guidance throughout the review process.

Funding

This work was supported in part by the Beijing Major Science and Technology Project under Contract No. Z231100007423015e.

Conflict of Interests

The authors declare no conflict of interests.

Data Availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Supplementary Material

Supplementary material for this article is available online in the format provided by the author (unedited).

References

Banas, J. A., Palomares, N. A., Richards, A. S., Keating, D. M., Joyce, N., & Rains, S. A. (2022). When machine and bandwagon heuristics compete: Understanding users' response to conflicting AI and crowdsourced fact-checking. *Human Communication Research*, *48*(3), 430–461.

Bhandari, A., Ozanne, M., Bazarova, N. N., & DiFranzo, D. (2021). Do you care who flagged this post? Effects of moderator visibility on bystander behavior. *Journal of Computer-Mediated Communication*, *26*(5), 284–300.

Binns, R., Van Kleek, M., Veale, M., Lyngs, U., Zhao, J., & Shadbolt, N. (2018). It's reducing a human being to a percentage: Perceptions of justice in algorithmic decisions. In R. Mandryk & M. Hancock (Eds.), *CHI '18: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Paper 337). ACM. https://doi.org/10.1145/3173574.3173951

Chae, J. H., & Tewksbury, D. (2024). Perceiving AI intervention does not compromise the persuasive effect of fact-checking. *New Media & Society*. Advance online publication. https://doi.org/10.1177/14614448241286881

Chung, M., Moon, W.-K., & Jones-Jang, S. M. (2024). AI as an apolitical referee: Using alternative sources to decrease partisan biases in the processing of fact-checking messages. *Digital Journalism*, *12*(10), 1548–1569. https://doi.org/10.1080/21670811.2023.2254820

Cloudy, J., Banks, J., & Bowman, N. D. (2023). The Str(AI)ght Scoop: Artificial intelligence cues reduce perceptions of hostile media bias. *Digital Journalism*, *11*(9), 1577–1596. https://doi.org/10.1080/21670811.2021.1969974

Craig, M. J., & Choi, M. (2024). The role of affective and cognitive involvement in the mitigating effects of AI source cues on hostile media bias. *Telematics and Informatics*, *88*, Article 102097. https://doi.org/10.1016/j.tele.2024.102097

DeVerna, M. R., Yan, H. Y., Yang, K.-C., & Menczer, F. (2024). *Fact-checking information from large language models can decrease headline discernment*. arXiv. http://arxiv.org/abs/2308.10800

Federico, C. M. (2022). The complex relationship between conspiracy belief and the politics of social change. *Current Opinion in Psychology*, *47*, Article 101354. https://doi.org/10.1016/j.copsyc.2022.101354

Flanagin, A. J., & Metzger, M. J. (2000). Perceptions of internet information credibility. *Journalism & Mass Communication Quarterly*, *77*(3), 515–540. https://doi.org/10.1177/107769900007700304

Frenken, M., & Imhoff, R. (2023). Don't trust anybody: Conspiracy mentality and the detection of facial trustworthiness cues. *Applied Cognitive Psychology*, *37*(2), 256–265. https://doi.org/10.1002/acp.3955

Garrett, R. K., Nisbet, E. C., & Lynch, E. K. (2013). Undermining the corrective effects of media-based political fact checking? The role of contextual cues and naïve theory. *Journal of Communication*, *63*(4), 617–637. https://doi.org/10.1111/jcom.12038

Gimpel, H., Heger, S., Olenberger, C., & Utz, L. (2021). The effectiveness of social norms in fighting fake news on social media. *Journal of Management Information Systems*, *38*(1), 196–221. https://doi.org/10.1080/07421222.2021.1870389

Graefe, A., & Bohlken, N. (2020). Automated journalism: A meta-analysis of readers' perceptions of human-written in comparison to automated news. *Media and Communication*, *8*(3), 50–59. https://doi.org/10.17645/mac.v8i3.3019

Graefe, A., Haim, M., Haarmann, B., & Brosius, H.-B. (2018). Readers' perception of computer-generated news: Credibility, expertise, and readability. *Journalism*, *19*(5), 595–610. https://doi.org/10.1177/1464884916641269

Hameleers, M., & van der Meer, T. G. (2019). Misinformation and polarization in a high-choice media environment: How effective are political fact-checkers? *Communication Research*, *47*(2), 227–250. https://doi.org/10.1177/0093650218819671

Hayes, A. F. (2015). An index and test of linear moderated mediation. *Multivariate Behavioral Research*, *50*(1), 1–22. https://doi.org/10.1080/00273171.2014.962683

Horne, B. D., Nevo, D., Adali, S., Manikonda, L., & Arrington, C. (2020). Tailoring heuristics and timing AI interventions for supporting news veracity assessments. *Computers in Human Behavior Reports*, *2*, Article 100043. https://doi.org/10.1016/j.chbr.2020.100043

Imhoff, R., & Bruder, M. (2014). Speaking (un–)truth to power: Conspiracy mentality as a generalised political attitude. *European Journal of Personality*, *28*(1), 25–43. https://doi.org/10.1002/per.1930

Jahanbakhsh, F., Katsis, Y., Wang, D., Popa, L., & Muller, M. (2023). Exploring the use of personalized AI for identifying misinformation on social media. In A. Schmidt, K. Väänänen, T. Goyal, P. O. Kristensson, A. Peters, S. Mueller, J. R. Williamson, & M. L. Wilson (Eds.), *CHI '23: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Article 105). ACM. https://doi.org/10.1145/3544548.3581219

Jennings, J., & Stroud, N. J. (2023). Asymmetric adjustment: Partisanship and correcting misinformation on Facebook. *New Media & Society*, *25*(7), 1501–1521. https://doi.org/10.1177/14614448211021720

Jia, H., Appelman, A., Wu, M., & Bien-Aimé, S. (2024). News bylines and perceived AI authorship: Effects on source and message credibility. *Computers in Human Behavior: Artificial Humans*, *2*(2), Article 100093.

Jia, H., & Luo, X. (2023). I wear a mask for my country: Conspiracy theories, nationalism, and intention to adopt Covid-19 prevention behaviors at the later stage of pandemic control in China. *Health Communication*, *38*(3), 543–551. https://doi.org/10.1080/10410236.2021.1958982

Kahan, D. M. (2015). The politically motivated reasoning paradigm, part 1: What politically motivated reasoning is and how to measure it. In R. Scott, M. Buchmann., & S. Kosslyn. (Eds.), *Emerging trends in the social and behavioral sciences* (pp. 1–16). Wiley. https://doi.org/10.1002/9781118900772

Kim, J.-N., & Lee, S. (2024). Conceptualizing conspiratorial thinking: Explicating public conspiracism for effective debiasing strategy. *American Behavioral Scientist*, *68*(10), 1366–1394. https://doi.org/10.1177/00027642231175637

Kreps, S., McCain, R. M., & Brundage, M. (2022). All the news that's fit to fabricate: AI-generated text as a tool of media misinformation. *Journal of Experimental Political Science*, *9*(1), 104–117.

Lee, J., & Bissell, K. (2024). User agency–based versus machine agency–based misinformation interventions: The effects of commenting and AI fact-checking labeling on attitudes toward the Covid-19 vaccination. *New Media & Society*, *26*(12), 6817–6837. https://doi.org/10.1177/14614448231163228

Liu, X., Qi, L., Wang, L., & Metzger, M. J. (2023). Checking the fact-checkers: The role of source type, perceived credibility, and individual differences in fact-checking effectiveness. *Communication Research*. Advance online publiaction. https://doi.org/10.1177/00936502231206419

Miller, J. M., Saunders, K. L., & Farhart, C. E. (2016). Conspiracy endorsement as motivated reasoning: The moderating roles of political knowledge and trust. *American Journal of Political Science*, *60*(4), 824–844. https://doi.org/10.1111/ajps.12234

Molina, M. D., & Sundar, S. S. (2024). Does distrust in humans predict greater trust in AI? Role of individual differences in user responses to content moderation. *New Media & Society*, *26*(6), 3638–3656. https://doi.org/10.1177/14614448221103534

Moon, W.-K., Atkinson, L., Kahlor, L. A., Yun, C., & Son, H. (2022). US political partisanship and Covid-19: Risk information seeking and prevention behaviors. *Health Communication*, *37*(13), 1671–1681.

Moon, W.-K., Chung, M., & Jones-Jang, S. M. (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. https://doi.org/10.1080/15205436.2022.2097926

Moon, W.-K., & Kahlor, L. A. (2025). Fact-checking in the age of AI: Reducing biases with non-human information sources. *Technology in Society*, *80*, Article 102760. https://doi.org/10.1016/j.techsoc.2024.102760

National Bureau of Statistics of China. (2018). *China statistical year book 2018*. China Statistics Press. https://www.stats.gov.cn/sj/ndsj/2018/indexeh.htm

Pareek, S., van Berkel, N., Velloso, E., & Goncalves, J. (2024). Effect of explanation conceptualisations on trust in AI-assisted credibility assessment. *Proceedings of the ACM on Human-Computer Interaction*, *8*(CSCW2), Article 383. https://dl.acm.org/doi/abs/10.1145/3686922

Rae, I. (2024). The effects of perceived AI use on content perceptions. In F. Floyd Mueller, P. Kyburz, J. R. Williamson, C. Sas, M. L. Wilson, P. Toups Dugas, & I. Shklovski (Eds.), *CHI '24: Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Article 978). ACM. https://doi.org/10.1145/3613904.3642076

Romer, D., & Jamieson, K. H. (2022). Conspiratorial thinking as a precursor to opposition to Covid-19 vaccination in the US: A multi-year study from 2018 to 2021. *Scientific Reports*, *12*(1), Article 18632.

Sundar, S. S. (2008). *The MAIN model: A heuristic approach to understanding technology effects on credibility*. MacArthur Foundation Digital Media and Learning Initiative.

Sundar, S. S. (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. https://doi.org/10.1093/jcmc/zmz026

Sundar, S. S., & Kim, J. (2019). Machine heuristic: When we trust computers more than humans with our personal information. In S. Brewster & G. Fitzpatrick (Eds.), *CHI '19: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Article 538). ACM. https://doi.org/10.1145/3290605.3300768

Sutton, R. M., & Douglas, K. M. (2020). Conspiracy theories and the conspiracy mindset: Implications for political ideology. *Current Opinion in Behavioral Sciences*, *34*, 118–122. https://doi.org/10.1016/j.cobeha.2020.02.015

Talwar, S., Dhir, A., Singh, D., Virk, G. S., & Salo, J. (2020). Sharing of fake news on social media: Application of the honeycomb framework and the third-person effect hypothesis. *Journal of Retailing and Consumer Services*, *57*, Article 102197. https://doi.org/10.1016/j.jretconser.2020.102197

Tam, L., & Lee, H. (2024). From conspiracy orientation to conspiracy attribution: The effects of institutional trust and demographic differences. *American Behavioral Scientist*, *68*(10), 1395–1411. https://doi.org/10.1177/00027642231174330

Thurman, N., Dörr, K., & Kunert, J. (2017). When reporters get hands-on with robo-writing: Professionals consider automated journalism's capabilities and consequences. *Digital Journalism*, *5*(10), 1240–1259. https://doi.org/10.1080/21670811.2017.1289819

Tulin, M., Hameleers, M., de Vreese, C., Opgenhaffen, M., & Wouters, F. (2024). Beyond belief correction: Effects of the truth sandwich on perceptions of fact-checkers and verification intentions. *Journalism Practice*. Advance online publication. https://doi.org/10.1080/17512786.2024.2311311

Vraga, E. K., & Bode, L. (2017). Using expert sources to correct health misinformation in social media. *Science Communication*, *39*(5), 621–645. https://doi.org/10.1177/1075547017731776

Waddell, T. F. (2019). Can an algorithm reduce the perceived bias of news? Testing the effect of machine attribution on news readers' evaluations of bias, anthropomorphism, and credibility. *Journalism & Mass Communication Quarterly*, *96*(1), 82–100. https://doi.org/10.1177/1077699018815891

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. https://doi.org/10.1080/10584609.2019.1668894

Wang, S. (2021). Moderating uncivil user comments by humans or machines? The effects of moderation agent on perceptions of bias and credibility in news content. *Digital Journalism*, *9*(1), 64–83. https://doi.org/10.1080/21670811.2020.1851279

Wang, S., & Huang, G. (2024). The impact of machine authorship on news audience perceptions: A meta-analysis of experimental studies. *Communication Research*, *51*(7), 815–842. https://doi.org/10.1177/00936502241229794

Wang, Y. (2021). Debunking misinformation about genetically modified food safety on social media: Can heuristic cues mitigate biased assimilation? *Science Communication*, *43*(4), 460–485. https://doi.org/10.1177/10755470211022024

Wischnewski, M., & Krämer, N. (2022). Can AI reduce motivated reasoning in news consumption? Investigating the role of attitudes towards AI and prior-opinion in shaping trust perceptions of news. In S. Schlobach, M. Pérez-Ortiz, & M. Tielman (Eds.), *HHAI2022: Augmenting human intellect* (pp. 184–198). IOS Press. https://doi.org/10.3233/FAIA220198

Wu, S., Tandoc, E. C., Jr., & Salmon, C. T. (2019). Journalism reconfigured: Assessing human–machine relations and the autonomous power of automation in news production. *Journalism Studies*, *20*(10), 1440–1457. https://doi.org/10.1080/1461670X.2018.1521299

Xu, D., Fan, S., & Kankanhalli, M. (2023). Combating misinformation in the era of generative AI models. In A. El Saddik, T. Mei, & R. Cucchiara (Eds.), *MM '23: Proceedings of the 31st ACM International Conference on Multimedia* (pp. 9291–9298). ACM. https://doi.org/10.1145/3581783.3612704

Zhu, Y., Fitzpatrick, M. A., & Bowen, S. A. (2024). Factors related to compliance with CDC Covid-19 guidelines: Media use, partisan identity, science knowledge, and risk assessment. *Western Journal of Communication*, *88*(3), 567–594. https://doi.org/10.1080/10570314.2023.2219239

Zhu, Y., Xu, J., Zhang, R., Lan, D., & Jiang, Y. (2024). Prior attitude, individualism and perceived scientists' expertise: Exploring motivated reasoning of scientific information about HIV risks of homosexuals in China. *Journal of Media Psychology: Theories, Methods, and Applications*. Advance online publication. https://doi.org/10.1027/1864-1105/a000437

About the Authors

Duo Lan (PhD, Beijing Normal University) is an assistant professor at Beijing University of Posts and Telecommunications, PR China. Her research interests include media technology effects, transcultural communication, and film studies.

Yicheng Zhu (PhD, University of South Carolina, USA) is an associate professor at Beijing Normal University, PR China. His research focuses on international strategic communication, social identities, and media technologies.

Meiyu Liu is a graduate student at Beijing Normal University. She is committed to studying the behavioral logic of interaction between humans and intelligent robots, keen to explore the changes in media credibility in the era of intelligence, and seeking crisis communication solutions.

Chuge He is a graduate student at Beijing Normal University. Her research interests include intelligent communication, new media, and risk communication.