The Trust Paradox of Authority and AI-Generated Content: A Drift-Diffusion Model of Cognitive Decision-Making

Authors: Yang Juan, Su Sheng, KANG Chunhua, Kang Chunhua

Date: 2025-11-24T00:00:00+00:00

Abstract

This study investigated the effects of news generation format and source authority on trust decisions and their underlying cognitive mechanisms through experimental design and the construction of a drift-diffusion model. The results revealed that trust rates, response times, and decision-making processes were all subject to interaction effects between generation format and source authority. Specifically, compared to non-authoritative sources, individuals exhibited lower trust rates for AI-generated news from authoritative sources, longer decision response times, and lower drift rates (v) and starting point biases (z) in the decision process. These findings suggest that the blind adoption of AI technology by authoritative media may undermine their credibility, exposing them to a trust crisis during their intelligent transformation.

Full Text

The Paradox of Authority and Trust in AI-Generated Content: Cognitive Decision Mechanisms Based on the Drift Diffusion Model

YANG Juan[1],[2]#, SU Sheng[1],[2]#, KANG Chunhua[1],[2]* (1. School of Psychology, Zhejiang Normal University, Jinhua 321004, China; 2. Zhejiang Philosophy and Social Science Laboratory for the Mental Health and Crisis Intervention of Children and Adolescents, Zhejiang Normal University, Jinhua 321004, China)

Abstract

This study investigates how news generation formats and source authority influence trust decisions and their underlying cognitive mechanisms through ex-

perimental design and drift diffusion modeling. The results demonstrate that trust rates, reaction times, and decision processes are all subject to interactive effects between generation format and source authority. Specifically, compared to non-authoritative sources, AI-generated news from authoritative institutions elicits lower trust rates, longer decision reaction times, and reduced evidence accumulation speed (v) and starting point bias (z). These findings reveal that the blind adoption of AI technology by authoritative media may undermine their credibility and precipitate a trust crisis during their digital transformation. The research underscores the importance of ensuring transparency and comprehensibility when integrating AI technologies, thereby alleviating public concerns about emerging technologies and preserving media platform credibility.

Keywords: source authority, trust decision, AI-generated, news, drift diffusion model

Introduction

Artificial Intelligence Generated Content (AIGC) represents an emerging production paradigm that yields outputs spanning text, images, and videos (Yu, 2023). While this technological advancement has substantially enhanced user experiences, it has also lowered the barrier for generating misinformation, posing severe cognitive challenges for the public in discerning content authenticity (Nightingale, 2022). Consequently, individuals have developed two opposing attitudes toward AIGC: "automation bias," where people overestimate the consistency and accuracy of AI technology and consequently favor its outputs, and "algorithm aversion," where individuals reject AI due to its dehumanized characteristics and lack of comprehension and agency (Lee, 2018). Within the journalism domain, where news media increasingly leverage AI to deliver higher-quality content, examining how AI technology differentially impacts media credibility and individual trust decision mechanisms is crucial for addressing trust risks arising from technological innovation.

Source authority in news information primarily depends on public trust in media institutions and their information reliability (Moran, 2022). "Authoritative news media" is not a natural construct but rather emerges from interactions among journalists, politicians, and experts, with authority serving as a cornerstone of news credibility (Anstead, 2018). Research on trust construction in news authority indicates that in digital environments, the "identity marker" of authoritative platforms has become a critical trust indicator (Tandoc, 2018). Traditional news institutions have established stable trust relationships with the public through professional training and rigorous oversight of news production processes, effectively curbing fake news (Robinson, 2007). Conversely, non-authoritative media, represented by self-media platforms, often prioritize "traffic" and "attention" over news authenticity, leading to rampant misinformation and severely undermining their authority and credibility (Wang, 2022). Therefore, we hypothesize that individuals may activate contrasting cognitive expectations when encountering AI-generated news from sources of varying au-

thority, resulting in an interactive effect between news generation method and source authority on audience trust decisions (H1). Specifically, the introduction of AI technology in authoritative media—where high trust expectations already exist—may be perceived as undermining information certainty and disrupting public cognitive habits, thereby triggering algorithm aversion toward AIGC and resulting in lower trust and longer decision reaction times (H1a). In contrast, for non-authoritative sources like self-media platforms, characterized by weak oversight and unreliable quality, audiences may view AI integration as a means to guarantee content authenticity and objectivity, aligning with core journalistic values (Zheng, 2021). This may foster automation bias toward AIGC, leading to higher trust and shorter reaction times (H1b).

Previous investigations of decision-making cognitive mechanisms have typically relied on dual-process theory, which posits two information processing systems in the human brain: a fast, automatic, and unconscious heuristic system that relies on prior knowledge and beliefs to make rapid judgments (Tversky & Kahneman, 1974), and a slow, conscious systematic system that requires greater cognitive resources and follows logical rules and deliberative reasoning (Evans, 1996). In the intelligent era, scholars have begun examining how AI influences individual decision-making mechanisms in human-machine collaboration contexts. Research confirms that to conserve cognitive resources, people tend to employ heuristic processing when judging information accuracy (Hause Lin, 2023), while other work combining decision process data with deep neural networks has revealed the role of systematic processing in modeling decision-making (Nikitin, 2021).

To further quantify cognitive mechanism differences in news trust decisions, this study employs the Drift-Diffusion Model (DDM) to model decision data. The DDM assumes that accumulated evidence fluctuates dynamically over time during decision-making, reaching a preset threshold before a response is made. The model comprises four core parameters: drift rate (v), starting point bias (z), decision boundary (α), and non-decision time (τ) (Zhang et al., 2020). In the context of our decision task, information accumulation begins as individuals read news. Here, v represents the rate at which evidence for a particular choice accumulates, quantifying participants' speed of integrating evidence for trust or distrust; z indicates prior bias before decision-making, quantifying participants' tendency to choose trust/distrust before evidence acquisition; α represents the amount of information accumulated before responding, with the upper boundary indicating "trust" and the lower boundary indicating "distrust" ; and τ reflects other factors affecting reaction time, such as information encoding and response execution (see Figure 1

).

Based on dual-process theory and DDM characteristics, we propose that audience cognitive processes during trust decision tasks integrate features of both heuristic and systematic processing, with news generation method and source authority interactively influencing these processing modes (H2). Specifically,

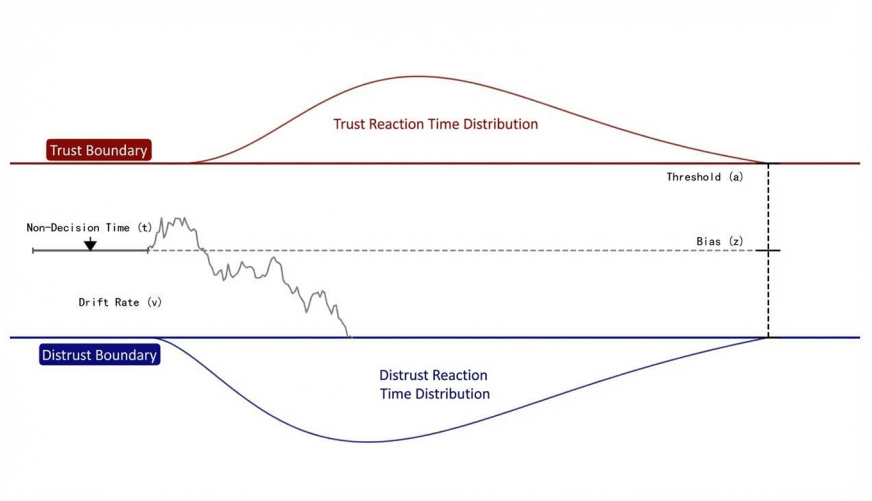

Figure 1: Figure 1

when encountering AI-generated news from authoritative sources, individuals' trust in journalistic professionalism and rigorous media oversight may clash with AI systems' "machine-like" qualities, potentially triggering algorithm aversion toward AIGC. This leads to systematic processing of information, more cautious evaluation of content authenticity, relatively slower information accumulation, and reduced starting point bias and drift rate (H2a). For news from non-authoritative sources, audiences' inherent distrust of social media platforms may lead them to view AI's objectivity as superior to unreliable human creators, fostering automation bias toward AIGC. This prompts heuristic processing that yields quick, coarse-grained, trust-biased decisions, increasing decision bias and drift rate (H2b). Since the experiment strictly controlled for response execution and reading material length, non-decision time and decision boundary were treated as constant parameters and excluded from analysis (Mormann & Russo, 2021).

2.1 Participants and Design

The study employed a 2 (news generation method: AI-generated vs. human-written) × 2 (source authority: authoritative platform vs. self-media platform) mixed experimental design, with news generation method as a between-subjects factor and source authority as a within-subjects factor. A total of 127 participants were recruited through an online platform. After attention screening and invalid data removal, 116 valid datasets remained (45 male, 38.8%; 71 female, 61.2%; mean age = 22.0 years, SD = 1.81). All participants received 8 RMB compensation upon task completion. Sensitivity analysis using G*Power indicated that, with $\alpha = 0.05$ (two-tailed) and 80% statistical power, the current

sample size was sufficient to detect a small-to-medium effect size of d = 0.25.

2.2 Materials and Procedure

All news materials were collected from social media platforms and categorized into four types: (1) authoritative media news (official outlets such as People' s Daily Online and Xinhua Net); (2) self-media news (Douyin self-media, unverified WeChat public accounts, etc.); (3) AI-generated news (including institutionally AI-generated content and AI-organized text); and (4) human-written news. News content covered diverse topics including technology, culture, education, and health. To optimize experimental control, each news excerpt was limited to 120-150 characters. The experimental program was developed using PsychoPy 2024.2.4.

Participants completed a news trust decision task. Both groups read news from authoritative and self-media platforms, with one group viewing exclusively human-written content and the other viewing exclusively AI-generated content. In each trial, source authority and generation method information were clearly labeled below the content. After reading each news item, participants made a "trust" or "distrust" keypress decision regarding its authenticity as quickly as possible.

2.3 Data Analysis

The study conducted repeated measures ANOVA on trust rates and reaction times using R, and modeled behavioral data with Python' s HDDM (Hierarchical Drift-Diffusion Model) library. Based on hierarchical Bayesian parameter estimation, we calculated posterior distribution differences for parameters across experimental conditions and examined the 95% Highest Density Interval (HDI) of these differences. If the HDI did not contain zero, the two conditions were considered significantly different on that parameter (Johnson et al., 2017).

3.1 Trust Rate Analysis

Repeated measures ANOVA with trust rate as the dependent variable revealed no significant main effect of news creation method ($F(1, 114) = 1.34$, $p = .248$, $^2G = 0.006$, $BF10 = 0.28$). The main effect of source authority was significant ($F(1, 114) = 114.71$, $p < 0.001$, $^2G = 0.348$, $BF10 = 5.35$), with trust rates significantly higher for authoritative sources ($M = 0.82$, $SD = 0.18$) than non-authoritative sources ($M = 0.53$, $SD = 0.21$). The interaction was significant ($F(1, 114) = 4.35$, $p = 0.039$, $^2G = 0.020$, $BF10 = 2.04$). Simple effects analysis showed no significant difference between AI-generated ($M = 0.54$, $SD = 0.21$) and human-written ($M = 0.52$, $SD = 0.21$) news in the non-authoritative condition ($t = 0.726$, $p = 0.470$, $BF10 = 0.24$). However, in the authoritative condition, trust rates for AI-generated news ($M = 0.77$, $SD = 0.17$) were significantly lower than for human-written news ($M = 0.85$, $SD = 0.18$; $t = -2.32$, $p = 0.022$, $BF10 = 3.22$) (see Figure 2

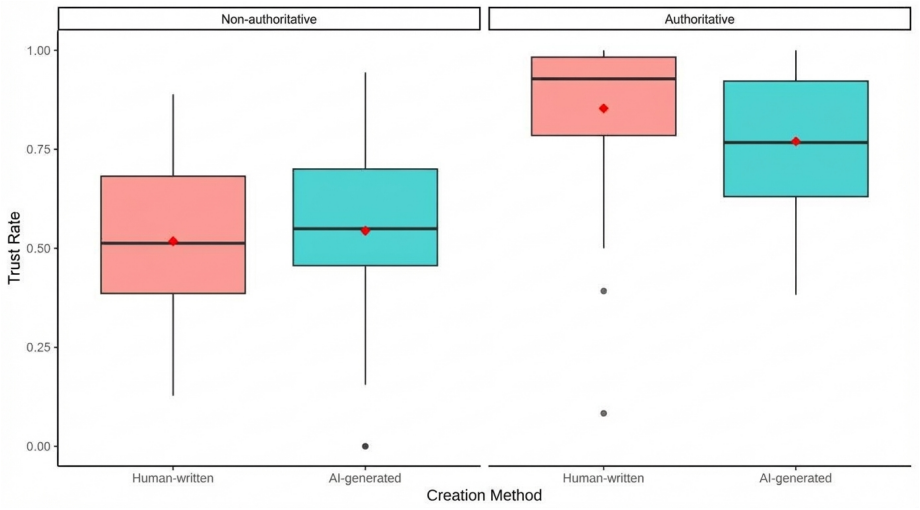

Figure 2: Figure 2

).

3.2 Reaction Time Analysis

Repeated measures ANOVA with reaction time as the dependent variable showed no significant main effect of news creation method ($F(1, 114) = 1.85$, $p = .176$, $^2G = 0.015$, $BF10 = 0.79$). The main effect of source authority was significant ($F(1, 114) = 33.98$, $p < .001$, $^2G = 0.014$, $BF10 > 100$), with reaction times longer for authoritative sources ($M = 7.20$, $SD = 3.36$) than non-authoritative sources ($M = 8.06$, $SD = 3.47$). The interaction was significant ($F(1, 114) = 19.95$, $p < .001$, $^2G = 0.008$, $BF10 > 100$). In the authoritative condition, reaction times were significantly longer for AI-generated news ($M = 7.98$, $SD = 3.83$) than human-written news ($M = 6.52$, $SD = 2.74$; $t = 2.315$, $p = 0.022$, $BF10 = 2.45$). In the non-authoritative condition, no significant difference emerged between AI-generated ($M = 8.17$, $SD = 3.87$) and human-written ($M = 7.96$, $SD = 3.11$) news ($t = 0.339$, $p = 0.735$, $BF10 = 0.21$) (see Figure 3

).

3.3 Drift Diffusion Model Results

To investigate the cognitive mechanisms underlying the effects of news generation method and source authority, we conducted hierarchical Bayesian regression analyses on two core DDM parameters: drift rate (v) and starting point bias (z).

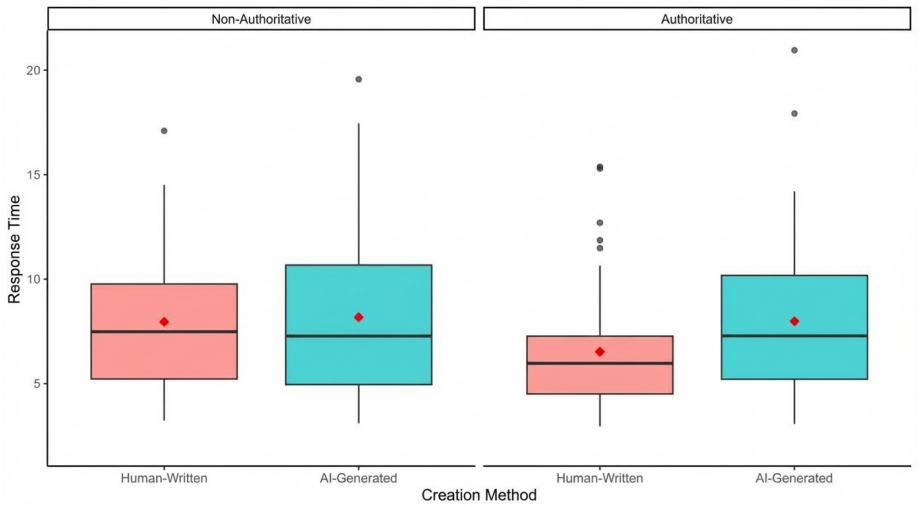

Figure 3: Figure 3

Drift rate analysis revealed that authoritative sources significantly accelerated evidence accumulation rate (b = 0.488, 95% HDI: [0.396, 0.578]), while news generation method showed no significant main effect (b = 0.025, 95% HDI: [-0.058, 0.112]). The interaction was significant (b = -0.195, 95% HDI: [-0.312, -0.068]), indicating that source authority moderated the effect of creation method. In the non-authoritative condition, AI-generated news had minimal impact on drift rate (increasing from 0.041 to 0.066). In the authoritative condition, however, AI-generated news significantly reduced evidence accumulation, with drift rate decreasing from 0.529 to 0.359. In this model, "label" indicates news creation method (0 = human-written, 1 = AI-generated) and "authority" indicates source authority (0 = non-authoritative, 1 = authoritative).

Analysis of pre-decision prior preference showed a significant positive main effect of source authority (b = 0.047, 95% HDI: [0.023, 0.073]), but no significant main effect of news generation method (b = 0.021, 95% HDI: [-0.003, 0.044]). The interaction was significant (b = -0.041, 95% HDI: [-0.072, -0.010]), again indicating moderation by source authority. In the non-authoritative condition, AI-generated news shifted the starting point slightly toward "trust" (from 0.475 to 0.496). In the authoritative condition, however, AI-generated news significantly weakened the prior trust preference associated with authority, shifting the starting point from 0.522 to 0.502. Variable coding for "label"and "authority" remains consistent with the previous analysis.

Discussion

This study examined how news generation method and source authority influence audience trust judgments and revealed the underlying cognitive mecha-

nisms through drift diffusion modeling. The findings demonstrate interactive effects of news generation method and source authority on both trust rates and reaction times in trust decisions, partially supporting hypothesis H1. Specifically, for non-authoritative sources, AIGC exerted modest but non-significant positive effects on trust rates and reaction times. This may reflect how countless non-authoritative outlets, particularly self-media platforms, have compromised news norms and sensationalized content in pursuit of traffic and attention (Jia, 2019), leading audiences to remain vigilant and skeptical about content authenticity and accuracy (Newman, 2022). Consequently, even the introduction of more objective and precise AI technology fails to dispel public doubts about these platforms' professionalism. This reveals that in the AI era, credibility building for non-authoritative institutions must prioritize developing professional expertise and rigorous content oversight rather than mere technological innovation (Xu & Fan, 2025). In contrast, for authoritative sources, AI-generated news elicited significantly lower trust rates and longer reaction times compared to human-written news. This suggests that AI technology undermines the trust advantage of authoritative institutions, likely due to algorithm aversion stemming from concerns about AI' s "black box" technology and its dehumanized characteristics. This aversion compels individuals to invest greater cognitive resources, adopt more cautious attitudes, and gather more information to make decisions, thereby prolonging decision times (Luo et al., 2023). This finding confirms that even amid growing technological worship, maintaining news credibility requires platforms to uphold high standards of responsibility and transparency in AI application (Zhang, 2020).

DDM modeling revealed significant interactive effects of news generation method and source authority on both evidence accumulation rate (v) and prior preference (z), strongly supporting hypothesis H2. For non-authoritative sources, AIGC positively influenced information processing speed and initial bias in trust decisions, triggering heuristic processing strategies for rapid, coarse-grained information evaluation. This phenomenon likely relates to automation bias stemming from trust in AI systems' powerful information processing capabilities and rigorous computational programming (Jones-Jang, 2023). Conversely, for authoritative news institutions, AI intervention significantly reduced both v and z parameters. When authoritative sources adopt AIGC, it triggers algorithm aversion, disrupts audiences' default trust, and prompts systematic processing characterized by careful evaluation and deep reflection. Individuals collect more information and evidence to assess content authenticity and accuracy, addressing potential risks from algorithmic uncertainty (Hong & Chen, 2022). These results demonstrate that in journalism, AI intervention is not universally beneficial; against the backdrop of authoritative sources, AI-generated news activates a potential cognitive shift, prompting audiences to switch from default heuristic processing to cautious systematic processing. This reduces decision efficiency and initial trust, posing a potential threat to authoritative media' s credibility (Evans, 2011).

In summary, this study reveals from both behavioral and cognitive perspectives

the "AI paradox" facing authoritative media during intelligent transformation: the introduction of AI technology to enhance efficiency may instead damage the core asset of journalism—public trust. These findings sound an alarm for authoritative news institutions, indicating that they must exercise caution when embracing technology and employ AI in more transparent and responsible ways to preserve hard-won credibility.

Limitations and Future Directions

This study has several limitations. First, stimulus materials were limited to pure text, whereas real-world news often involves multimodal content including images and videos. Future research should extend investigations to multimodal contexts to assess how AIGC influences audience trust across more complex presentation formats. Second, this study focused on revealing potential risks of AI intervention; future research should shift toward constructive solutions, such as exploring how to enhance transparency in AI application processes, developing AI-assisted tools designed to augment rather than undermine content credibility, and implementing effective communication strategies.

References

Hong, J., & Chen, R. (2022). Consciousness stimulation and rule imagination: The tactical orientation and practical paths of user resistance to algorithms. *Journalism & Communication*, 29(8), 38-56+126-127.

Jia, W. (2019). Breaking out of the pan-entertainmentism cycle. *People' s Tribune*, (2), 18-20.

Luo, Y., Zhu, G., Qian, W., et al. (2023). Algorithm aversion in the era of artificial intelligence: Research framework and future prospects. *Management World*, 39(10), 205-233.

Wang, D., & Yang, J. (2022). Analysis of citizen journalism production and dissemination in the new media era. *Media*, 23(1), 94-96.

Xu, X., & Fan, L. (2025). The practical path and value pursuit of data news from the perspective of journalistic authority. *Publishing Wide Angle*, (2), 3-8.

Yu, G. (2023). An important measure for social collaborative governance in the rising environment of generative content production: On the importance and necessity of full-process AIGC labeling. *Youth Journalist*, 11(2), 74-76.

Zhang, K. (2020). Research on news credibility issues in the context of algorithmic recommendation. *Modern Communication*.

Zhang, Y., Li, H., & Wu, Y. (2020). Application of computational models in moral cognition research. *Advances in Psychological Science*, 28(7).

Zheng, Z. (2021). Ethical crisis and legal regulation of artificial intelligence algorithms. *Science of Law (Journal of Northwest University of Political Science*

and Law), 39(1), 14-26.

Anstead, N. & Chadwick A. (2018). A primary definer online: The construction and propagation of a think tank' s authority on social media. *Media, Culture & Society* 40(2): 246-266.

Evans, J. S. B. T. (2011). Dual-Process Theories of Reasoning: Contemporary Issues and Developmental Applications. *Developmental Review*, 31(2), 86-102.

Hause Lin, Pennycook Gordon & Rand David G. (2022). Thinking more or thinking differently? Using drift-diffusion modeling to illuminate why accuracy prompts decrease misinformation sharing. *Cognition*, 105312-105312.

Johnson, D. J., Hopwood, C. J., Cesario, J., & Pleskac, T. J. (2017). Advancing research on cognitive processes in social and personality psychology: A hierarchical drift diffusion model primer. *Social Psychological and Personality Science*, 8(4), 413-423.

Jones-Jang, S. M., & Park, Y. J. (2023). How do people react to AI failure? Automation bias, algorithmic aversion, and perceived controllability. *Journal of Computer-Mediated Communication*, 28(1), 29-30.

Lee, M. K. (2018). Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data & Society*, 5(1), 14-15.

Moran, R. E. & Nechushtai, E. Before Reception: Trust in the News as Infrastructure.(2022). *Journalism*, 24(3): 456-474.

Mormann, M., & Russo, J. E. (2021). Does Attention Increase the Value of Choice Alternatives? *Trends in Cognitive Sciences*, 25(4), 305-315.

Newman, N., Fletcher, R., Robertson, C. T., Eddy, K. & Nielsen, R. K. Reuters Institute Digital News Report 2022. Oxford: Reuters Institute for the Study of Journalism, 2022.

Nightingale, S. J., & Farid, H. (2022). AI-synthesized faces are indistinguishable from real faces and more trustworthy. *Proceedings of the National Academy of Sciences of the United States of America*.

Nikitin, A., & Kaski, S. (2021). Decision rule elicitation for domain adaptation. *Proceedings of the 26th International Conference on Intelligent User Interfaces*, 244-248.

Tandoc, E. C., Hellmueller, L., & Vos, T. P. (2018). The roles of the gatekeepers in digital news credibility. *Journalism*, 20(5), 658-675.

Tversky, A., & Kahneman, D. (1974). Judgment Under Uncertainty: Heuristics and Biases: Biases in Judgments Reveal Some Heuristics of Thinking Under Uncertainty. *Science*, 185(4157), 1124-1131.