World Journal of Advanced Research and Reviews
eISSN: 2581-9615 CODEN (USA): WJARAI
Cross Ref DOI: 10.30574/wjarr
Journal homepage: https://wjarr.com/

(RESEARCH ARTICLE)

Check for updates

From "Information Cocoons" to "Cognitive Hallucination": The Dissemination Mechanisms and Governance Dilemmas of Health Information Disorder in Algorithmic Recommendation Environments

Zihong WU [1] and Yuexing WU [2,*]

[1] *International Education College, Hunan City University, Yiyang Hunan 413000, China.*
[2] *School of Management, Hunan City University, Yiyang, Hunan 413000, China.*

World Journal of Advanced Research and Reviews, 2026, 30(01), 1042-1047

Publication history: Received on 01 March 2026; revised on 06 April 2026; accepted on 09 April 2026

Article DOI: https://doi.org/10.30574/wjarr.2026.30.1.0903

Abstract

Objective: This study examines how algorithmic recommendation environments reshape the dissemination mechanisms of health information disorder and explores why existing governance strategies have become increasingly ineffective.

Methods: Drawing on the perspectives of philosophy of technology and the political economy of communication, the paper adopts a theoretical and literature-based analytical approach to construct a three-dimensional framework of technology, commerce, and cognition.

Results: The study shows that health information disorder in algorithmic environments is generated through three interrelated mechanisms: probabilistic hallucination in large-model generation, systematic attention exploitation driven by platform logic, and multimodal hallucinatory effects that weaken users' critical judgment, especially among older adults. These changes indicate that the traditional concept of the "information cocoon" is no longer sufficient to explain the complexity of current health information risks. Instead, a new form of risk, termed **cognitive hallucination**, has emerged. The findings further reveal that governance approaches centered on content blocking face temporal, spatial, and structural limitations.

Conclusion: Effective governance should move beyond ex post deletion and toward front-end intervention. The concept of cognitive friction provides a useful theoretical basis for redesigning governance strategies and advancing age-friendly digital governance in algorithm-mediated health communication.

Keywords: Algorithmic recommendation; Health information disorder; Cognitive hallucination; Cognitive friction; Age-friendly digital governance

1. Introduction

The dissemination of health information is undergoing a profound transformation driven by technological mediation. China's Healthy China 2030 strategy regards health information as a foundational strategic resource, yet the pace of technological evolution has far outstripped the adaptive capacity of existing governance systems. Traditional health communication research has largely focused on the authority of the information source (who says it) and the accuracy of the content (what is said). By contrast, algorithmic recommendation environments have restructured the power relations of communication: platform algorithms have become the ultimate arbiters of information visibility, and their

* Corresponding author: Yuexing WU

optimization goal of engagement maximization has created a structural tension with the public-oriented goals of health communication, namely the pursuit of the public good [1].

An even more serious challenge comes from the widespread adoption of Artificial Intelligence Generated Content (AIGC). The probabilistic generative mechanism of large language models is, in essence, a form of statistical conjecture: their outputs are characterized by syntactic coherence coupled with factual uncertainty [2]. When such "algorithmic hallucinations" are presented in multimodal forms—including text, images, voice, and digital humans—and are then precisely delivered to older adults experiencing cognitive decline, the risk described by the traditional theory of the "information cocoon" as merely a narrowing of informational horizons is upgraded into a risk of cognitive hallucination. In other words, audiences are exposed not only to a single and limited range of information sources, but also to an erosion of their fundamental capacity to judge the truthfulness of information [3–4].

This paper seeks to address three key questions: How have the dissemination mechanisms of health information disorder undergone a qualitative transformation in algorithmic recommendation environments? Why have existing governance frameworks proven ineffective? And what theoretical resources can support governance innovation?

Cass Sunstein's concept of the information cocoon describes a phenomenon in which users, amid vast quantities of information, selectively consume homogeneous content, thereby narrowing their horizons. Chinese scholars have extended this line of thought through such concepts as the "echo chamber effect" and the "filter bubble," emphasizing that algorithms exacerbate these tendencies. Research by Mo Zuying and Pan Daqing further shows that the information cocoon effect can weaken users' ability to identify false information [5]. However, these theories presuppose users' agency (active selection) and lucidity (an awareness of homogenization despite continued choice), making them inadequate for explaining the current characteristics of health information risks.

First, the shift is from activity to passivity. In algorithmic recommendation environments, information acquisition has moved from active searching to passive feeding. Users no longer need to search for information; instead, algorithms deliver content precisely on the basis of user profiling. The weaver of the "cocoon" is thus no longer the user, but the platform algorithm.

Second, the shift is from lucidity to hallucination. Traditional cocoon theory assumes that users retain the ability to recognize the boundaries of information. Yet the multimodal realism of AIGC—through photorealistic images, cloned voices, and digital humans—endows false information with the quality of hyperreality: it appears more perfect, more persuasive, and more credible than reality itself, thereby drawing audiences into a hallucinatory state in which they accept falsehood as truth [2–4].

Third, the shift is from narrowing to poisoning. A cocoon merely limits vision, whereas algorithmic amplification enables specific false information to acquire viral dissemination power. In this sense, the "cocoon" is transformed into a "poisoned chamber."

From the perspective of the political economy of communication, the commercial essence of algorithmic recommendation becomes clear. In the platform economy, attention is commodified, and user dwell time becomes the central performance indicator. Through the continuous extraction, analysis, and prediction of behavioral data, algorithms effectively realize the covert shaping of user behavior.

In the field of health information, this logic manifests as the monetization of anxiety. Existing studies have found that older adults' adoption of false health information in contexts such as online live-streaming platforms is often jointly shaped by platform-specific situational design, emotional mobilization, and interactive mechanisms [6]. Meanwhile, health information anxiety further influences users' paths of information avoidance, acceptance, and judgment [7]. In other words, health topics concern fear of illness and survival, making them highly emotionally arousing and thus naturally compatible with the optimization logic of algorithms. Platforms indiscriminately amplify highly engaging content, while false information is often far more emotionally provocative than evidence-based medical information. A headline such as "Shocking! This food causes cancer" will predictably attract more clicks than "A meta-analysis of cohort studies on the carcinogenic risk of a particular vegetable." Algorithms thereby function as selective amplifiers of false information.

The philosophical foundations of large language models deserve close scrutiny. GPT-like models are based on autoregressive architectures and generate text by predicting "the next most probable word"; their essence lies in statistical pattern matching rather than factual reasoning [2]. This mechanism naturally entails the risk of hallucination,

that is, the confident fabrication of nonexistent facts, fictitious studies, and forged authorities. Within generative simulacral environments, this risk further evolves into cognitive distortion and socially shared hallucination [3].

More importantly, AIGC technologies have industrialized the production of false information and enabled its multimodal packaging. Black-market and gray-market actors exploit generation-optimization technologies to mass-produce seemingly professional health content, distribute it across platforms through proxy IPs and networks of fake accounts, and ultimately contaminate the training data of large models. This creates a vicious cycle of hallucination–pollution–rehallucination [2].

Existing literature on AIGC-related health risks has focused primarily on the explicit problem of deepfakes, while paying insufficient attention to what may be termed everyday hallucination—namely, AI-generated health advice that appears plausible yet contains subtle factual errors. Reviews of the themes and core elements of false health information research also suggest that the complexity of health misinformation has already surpassed the explanatory scope of a single "rumor governance" paradigm, calling instead for integrated analysis across the levels of information generation, dissemination pathways, and user cognition [8].

2. Materials and Methods

Building on the above literature, this paper constructs a three-dimensional analytical framework of technology–commerce–cognition to explain the dissemination mechanisms of health information disorder in algorithmic recommendation environments.

Hallucination in large models is not a technical defect (bug), but rather an architectural feature (feature). Its roots lie in the following aspects.

First, there is statistical bias in training data. In online environments, false health information often achieves greater dissemination volume than evidence-based medical content. What models learn, therefore, is not what is most accurate, but what is most prevalent [2][8].

Second, there is a lack of fact-anchoring mechanisms. Models cannot access authoritative medical databases in real time to verify their claims. Their "knowledge" is bounded by the cutoff date of the training data, and they are unable to distinguish truth from falsehood within that data. Existing studies have pointed out that if health information identification relies solely on surface-level textual features, while lacking prior knowledge about publishers, social perception data, and authoritative knowledge anchoring, it is highly prone to misjudgment or omission [9].

Third, there is the fragility of multimodal alignment. Cross-modal generation across text, images, and speech depends on statistical correlation rather than semantic understanding, making it easy to produce hallucinatory combinations in which the image and text are inconsistent, yet still appear professional and credible [2–3].

There exists an irreconcilable conflict between the optimization goals of platform algorithms and the public goals of health communication, as shown in Table 1.

Table 1 Conflicts Between Platform Logic and the Public Goals of Health Communication in Algorithmic Recommendation Environments

Dimension	Platform Algorithm Goals	Public Goals of Health Communication	Manifestation of Conflict
Time	Maximize dwell time	Minimize decision-making time (to obtain accurate information efficiently)	Fragmentation; endless scrolling
Emotion	Maximize arousal (fear, anger, hope)	Neutrality, prudence, evidence-based communication	Clickbait; sensationalism
Interaction	Maximize sharing and commenting	Minimize the risk of misinformation dissemination	Encouraging the forwarding of unverified information
Personalization	Maximize matching precision	Minimize informational bias	Reinforcement of information cocoons; narrowing of horizons

This structural conflict deprives platforms of sufficient incentive to strictly review health information. Research shows that in older adults' adoption of false health information, platform contexts, interactive discourse strategies, and emotional cues can significantly strengthen users' acceptance tendencies [6]; moreover, health information anxiety further reshapes their information-processing and behavioral pathways [7]. This means that the high interactivity of false information can be directly converted into platform profits, whereas the costs of correction—such as fact-checking, user education, and reputational risk—are often borne by society as a whole.

Multimodal information exerts a particularly profound cognitive influence on older audiences.

First, there is sensory overload and the suppression of critical judgment. The "hyperreal" environment composed of digital human experts, cloned or synthetic voices, and dynamic charts generates intense sensory stimulation that may exceed older adults' capacity for cognitive regulation. When the cognitive load of information processing becomes too high, audiences tend to rely on heuristic judgment—"if it looks professional, it must be true"—rather than analytic judgment—"verify the source and check the evidence" [10].

Second, there is authority transfer and trust misalignment. In traditional health communication, trust is built on institutional authority, such as hospitals and medical schools. In algorithmic recommendation environments, however, trust is reconstituted around algorithmic relevance: "recommended for you" comes to replace "certified by experts." Older adults often lack a clear understanding of algorithmic logic and may mistakenly equate "being recommended" with "being endorsed," making them more vulnerable to truth-bias effects produced by repeated exposure [10].

Third, there is the compression of the decision-making chain. The "frictionless" experience pursued by platforms compresses health decision-making from a multi-step process—information acquisition, cross-verification, risk assessment, and action decision—into a reflexive chain of seeing–believing–acting. This form of "seamless interaction" dismantles the psychological defense mechanisms that older adults should ordinarily possess, allowing risk to be delivered in a smooth, targeted, and almost invisible manner [10].

In response to the above dilemmas, this paper proposes cognitive friction as a theoretical lever for governance intervention.

Cognitive friction refers to the strategic introduction of cognitive load into information interaction so as to interrupt automated processing routines and activate users' critical thinking. In contrast to the "frictionless" experience pursued by platforms, this concept advocates a deliberate form of deceleration: through design interventions that prolong decision-making time, signal risk cues, and provide verification tools, "smooth deception" can be transformed into friction-induced vigilance. From the perspective of information processing, the key to improving users' ability to identify false health information lies in prompting them to shift from low-investment heuristic judgment to more reflective analytic processing [10].

Governance intervention based on cognitive friction can be further specified along four dimensions—time, source, verification, and sociality. The corresponding strategies and objectives are shown in Table 2.

Table 2 Dimensions, Specific Strategies, and Objectives of Cognitive Friction Governance Interventions

Dimension	**Specific Strategy**	**Objective**
Temporal friction	Impose a mandatory "cooling-off period" for health-related content (e.g., a 30-second countdown before clicking "purchase")	Interrupt impulsive decision-making
Source friction	Mandate clear disclosure of the mode of content production (AI-generated / human-created / institutionally certified)	Activate authority-based judgment
Verification friction	Provide one-click access to authoritative databases for comparison (e.g., the official website of the National Health Commission)	Offer verification tools
Social friction	Require users to confirm before sharing whether they have consulted family members or a doctor	Introduce social gatekeeping

Given the cognitive characteristics of older adults, friction design should avoid excessive friction that may lead to abandonment of use. Research indicates that middle-aged and older users' judgment paths in identifying false short videos exhibit a marked dependence on experience and heuristic tendencies; therefore, governance design should not

simply increase complexity, but rather maintain a balance between protection and usability [4]. A "selectable friction with default activation" model is recommended: users may choose to disable certain prompts, but in the default setting safety should take priority. In addition, intergenerational collaboration interfaces should be designed—for example, "one-click forwarding to children for verification"—so that family support can be incorporated into the governance network.

3. Results and Discussion

3.1. Existing governance frameworks are trapped in a threefold dilemma

First, there is the temporal dilemma: the lag of ex post deletion. The dissemination of health information follows a viral exponential curve: false information can achieve large-scale diffusion within a matter of hours, whereas fact-checking often requires a much longer cycle. By the time a debunking message is released, the original information has already produced a first-impression effect, and the algorithm has further reinforced the recommendation of similar content based on prior user interactions. Research on information dissemination and governance during public health emergencies in the era of intelligent media likewise shows that platformized and intelligent communication environments significantly amplify the problem of temporal lag in governance [1].

Second, there is the spatial dilemma: the infeasibility of cross-border enforcement. Black- and gray-market actors exploit cloud servers, proxy IPs, encrypted payments, and related tools to realize a form of de-materialized operation. Servers are located overseas, operators are based overseas, and financial flows are routed overseas, thereby creating a de facto problem of cross-border governance. Traditional territorially based regulatory logic has shown evident failure in platformized and intelligent communication environments, which is also one of the major reasons why the governance of emergency public health information repeatedly encounters real-world bottlenecks [1].

Third, there is the structural dilemma: institutional shelter for platform exemption. Under current regulatory rules, the platform liability framework of "notice-and-takedown" often appears overly passive when confronting algorithmic recommendation systems. Platform algorithms are frequently packaged as technologically neutral distribution tools rather than as editorial forces with the power to select, rank, and amplify content. In reality, however, platforms derive revenue from the traffic, advertising, and e-commerce chains generated by false health information while bearing only limited responsibility. This structural imbalance—characterized by the privatization of profits and the socialization of risks—constitutes a key institutional background for the continued expansion of health information disorder [1].

In algorithmic recommendation environments, health information disorder has evolved from the narrowing risk associated with the information cocoon into the deeper risk of deception characterized as cognitive hallucination. The probabilistic generative mechanisms of AIGC, the attention-exploitation logic of platform economies, and the sensory manipulation effects of multimodal information interact to form a systemic threat to public health.

The existing governance pathway centered on content blocking has proven inadequate across the temporal, spatial, and structural dimensions. The necessity of a paradigm shift lies in moving from back-end deletion to front-end intervention, from platform exemption to algorithmic accountability, and from information supply to cognitive empowerment.

The theory of cognitive friction provides scholarly support for this transformation, but its practical implementation depends on both platform cooperation and policy enforcement. Future research may further explore the optimal dosage of friction design—one that is effective without causing user attrition—its cross-cultural adaptability across different older populations, and its technical implementation pathways, such as browser plug-ins or platform API integration.

The main limitation of this study lies in its emphasis on mechanism analysis and theoretical construction; its empirical evidence relies primarily on secondary literature and illustrative cases. Subsequent research may employ eye-tracking experiments and cognitive psychology experiments to quantify the specific mechanisms through which multimodal hallucinatory effects operate, thereby providing more precise scientific evidence for design interventions.

4. Conclusion

This study argues that the risks of health information disorder in algorithmic recommendation environments can no longer be adequately explained by the traditional framework of the information cocoon. Instead, the emergence of AIGC and multimodal recommendation systems has produced a more profound form of risk, namely cognitive hallucination. By integrating technological, commercial, and cognitive dimensions into a unified analytical framework, this paper

highlights the structural roots of contemporary health information disorder and explains why existing content-blocking strategies have become increasingly ineffective.

The paper further contends that governance innovation should not remain limited to ex post content deletion. Rather, it should move toward front-end intervention through cognitive friction, so as to interrupt automatic persuasion, strengthen users' verification capacity, and improve age-friendly digital governance. In this sense, cognitive friction offers a potentially valuable theoretical basis for rethinking the governance of health misinformation under conditions of algorithmic mediation.

Compliance with ethical standards

Disclosure of conflict of interest

No conflict of interest to be disclosed.

Funding

This work was supported by the Hunan Provincial Social Science Foundation Project (Grant No. 25YBA235).

References

[1] Qi Z, Tong D. Information dissemination and governance in public health emergencies in the era of intelligent media. J Tianjin Norm Univ Soc Sci. 2025;(5):66-73. Chinese.

[2] Mo Z, Pan D, Liu H, et al. The problem of AIGC false information and its root causes from the perspective of information quality. Doc Inf Knowl. 2023;40(4):32-40. Chinese.

[3] Zhang D. The face of the algorithmic other: cognitive distortion and social hallucination in generative simulacra. Chongqing Sci Technol News. 2026-01-20(010). Chinese.

[4] Pan S, Zhu Y, Xie L. An analysis of the identification and judgment pathways of false short videos among middle-aged and older adults. Chin J Journal Commun. 2023;45(2):127-155. Chinese.

[5] Mo Z, Pan D. On the influence of the information cocoon effect on users' ability to identify false information. Res Libr Sci. 2023;(3):50-57. Chinese.

[6] Chen H, Yang X, Chen P, et al. A grounded analysis of older adults' adoption of false health information in the context of online live streaming. Libr Trib. 2024;44(9):152-160. Chinese.

[7] Gu D, Sun J, Ding Q, et al. Research on the influence path of health information anxiety on health information avoidance behavior: an exploration based on grounded theory. Libr Inf Serv. 2023;67(19):111-120. Chinese.

[8] Xiong H, Meng X, Ye J. A review of research themes and core elements of false health information during the COVID-19 pandemic. Libr Inf Serv. 2023;67(7):135-149. Chinese.

[9] Zhao Y, Pang H, Shi Y. Health information profiling and false health information identification: integrating social perception data and publishers' prior knowledge. Doc Inf Knowl. 2024;41(6):141-154,165. Chinese.

[10] Cao Y, Ke Q. Why do people differ in their susceptibility to false health information? An analysis based on the configuration of information processing processes. Mod Inf. 2023;43(1):40-54. Chinese.