

“Algorithmic Transparency” or “Algorithm Panic”? A Literature Review and Frontier Issues on AIGC Advertising Transparency and Consumer Responses

Xiyuan Wu*

Information Management and Financial Intelligence Management, School of Management Science and Engineering, Southwestern University of Finance and Economics, Chengdu, 611130, Sichuan

**Corresponding author: Xiyuan Wu.*

Abstract

The quick dissemination of generative artificial intelligence (AIGC) is changing the production methods of the advertising industry, and has also pushed “whether to disclose and how to disclose” to the center of regulation and the market. On a global scale, AI labeling systems have been implemented one after another, while academia still has obvious disagreements about what AIGC advertising transparency will bring—Is transparency a good medicine for restoring trust, or a hidden danger that damages advertising effectiveness? This research, based on literature from the Web of Science Core Collection and the China National Knowledge Infrastructure (CNKI), adopts a systematic review method to sort out the conceptual connotation, effect evidence, theoretical explanations, boundary conditions, and research gaps of AIGC advertising transparency. The paper ascertains that the impact of AIGC advertising transparency on consumers has a “double-edged sword” characteristic, that is to say, its net effect depends on the interaction between disclosed information and consumers’ existing expectations. The uncanny valley theory, CASA theory, the persuasion knowledge model, and attribution theory are not competing with one another; rather, they separately explain different links in the temporal chain of “expectation formation—expectation violation—defense activation—attribution judgment.” The direction and strength of the effect are also moderated by multiple factors, including consumer characteristics, brand and product factors, advertising formats, and platform contexts. In this context, this paper constructs an integrated theoretical framework and proposes future issues, such as differentiation of disclosure framing, differential effects of advertising formats, dynamic transparency, and human-AI collaborative transparency, providing a foundation for subsequent empirical research and regulatory practice.

Keywords

AIGC advertising, transparency disclosure, consumer trust, perceived eeriness, uncanny valley theory, persuasion knowledge model

1. Introduction

The high-speed spread of generative artificial intelligence (AIGC) is changing the production logic of the advertising sphere. The use of AI has already been applied by brands on a massive scale to create advertising material: both Nestlé and Mondelēz have been using DALL-E to generate advertising images and visuals for a very long time; numerous domestic internet platforms have also built their own AI systems to develop AIGC

advertising material production solutions, which are much more efficient in creating creative materials. Using the SCP model, Jiang Zhibin and Zhou Yang [1] indicated that generative artificial intelligence is introducing stage-based disruptive innovation in the structure, behavior, and performance of the advertising industry and that the so-called mechanization of the advertising creative production process cannot be reversed.

Nevertheless, the spread of technology also comes with fresh ethical concerns. The fact that consumers are not aware that the advertising material is created by artificial intelligence may cause them to make an evaluation of a brand in the condition of being in the dark; when they find out, they are likely to respond negatively to the advertisement. This is why international regulators have started adopting AI labeling systems: the Artificial Intelligence Act of the European Union demands overt labeling of deep synthetic content [2], and the Chinese Interim Measures on the Administration of Generative Artificial Intelligence Services also explicitly demand labeling of AI-generated content [3].

However, it brings up another question as well: What is precisely the role of transparency disclosure in the advertising response of the consumers? The problem has caused intense discussion in academic circles. On the one hand, a few papers have suggested that disclosure of transparent information is likely to increase consumer confidence and credibility of advertising, whereas on the other hand, other works have indicated that AI disclosure may lower advertising views and cause the uncanny valley feeling among consumers. This double-edged sword effect also places advertisers and platforms in a dilemma.

In the context of the response, this paper suggests three fundamental research questions: Firstly, what form of effect may the transparency disclosure of AIGC advertising really bring to the consumer? Secondly, what is the mechanism behind it? Thirdly, under what circumstances is the impact of transparency disclosure amplified or diminished? In order to address these questions, this paper uses the traditional approach to conducting a systematic literature review to find, filter, and code pertinent articles available in the Web of Science Core Collection and CNKI database during the period of January 2015-August 2026 (the procedure is explained in part two). Overall, 44 articles are finally considered, and a synthesis of findings is made based on conceptual definitions, effect evidence, theoretical explanations, and boundary conditions. There is an intention to eliminate controversies, develop an explanatory theory, determine gaps in research, and offer a theoretical basis for future empirical studies.

2. Literature Search and Screening Process

2.1 Search Strategy

The Web of Science Core Collection (SSCI) was searched to find information about English literature, having a time gap between January 2015 and August 2026. The initial date of 2015 was selected since deep-learning-based content creation technologies (like generative adversarial networks) roughly started to be applied commercially in this timeframe, and the theoretical background sources that this paper is based on (including uncanny valley theory and the persuasion knowledge model) are not restricted within that time frame and were researched independently. The search formula was: TS = (“AI-generated” OR “generative AI” OR “AIGC” OR “artificial intelligence” OR “virtual influencer” OR “synthetic media”) AND TS = (“adverts” OR “marketing” OR “branding”) AND TS = (“disclos” OR “transparen” OR “label” OR “identif” OR “trust”). In Chinese literature, the CNKI journal database and the collected journals database were searched based on the search formula: Subject = (AIGC advertising + generative artificial intelligence advertising + AI-generated advertising + virtual digital human + virtual influencer) AND Subject = (transparency + disclosure + labeling + identification + trust).

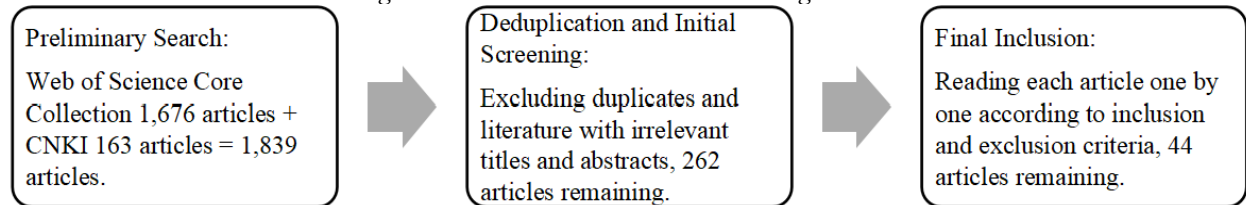
2.2 Inclusion, Exclusion Criteria, and Screening Results

The inclusion criteria are: (1) scholarly articles that have been published in a peer-reviewed journal, or white papers, standards, and regulation documents published by authoritative entities; (2) the research topic can cover transparency, disclosure, labeling, or identification related issues of AIGC/generative AI/virtual digital humans in advertising or marketing communication; (3) the research focuses on consumer-based response measures (attitude towards advertisements, attitudes towards brands, trust, purchase intent, behavior, etc.). The exclusion criteria are: (1) technical implementation literature or optimization of the algorithms literature; (2) literature dealing with generating AI content without mentioning advertising marketing contexts

or consumer reactions (including AI news writing and AI educational tools); (3) dissertations, conference abstracts, newspaper articles and commentaries; (4) literature where even after two search rounds the full text has not yet been located.

The results of the search and screening are shown in Figure 1.

Figure 1: Literature Search and Screening Process



3. Definition of Core Concepts

3.1 AIGC and AIGC Advertising

The present scholarly view of the idea of AIGC is yet to be developed. According to Li Baiyang et al. [4], AIGC can be seen as a new form of producing content and a collection of technologies that are able to generate content automatically; according to the China Academy of Information and Communications Technology and the JD Exploration Research Institute [5], in the Artificial Intelligence Generated Content (AIGC) White Paper, AIGC is defined as both a kind of content, viewed in the context of content producers, as well as a way of creating content, and a group of technologies to create automated content. Jiang Zhibin and Zhou Yang [1] also noted at the industry level that the disruption caused by AIGC to the advertising industry is characterized by stages, i.e., it starts as an auxiliary tool, followed by collaborative creation, then autonomous generation, and the level of influence of any stage on the industrial organizational structure varies.

According to this review, the concept of AIGC advertising is defined by this paper as follows: an advertising form that involves generative artificial intelligence technologies (e.g., large language models, image generation models, digital human driving technologies, etc.) to automatically or semi-automatically produce advertising copy, visual materials, virtual endorsers, or full advertising creatives. The definition includes the spectrum ranging from the notion of AI-assisted human creation to AI autonomous generation, and leaves room in the conceptual space to talk about differentiated disclosure framing (the distinction in expression between AI-assisted and AI-generated) a bit later.

3.2 The Connotation of Advertising Transparency

Transparency is an idea that originated in organizational behavior and information economics and refers to the accessibility, understandability, and verifiability of information during the transmission process. Transparency in the advertising context may be separated into two elements: one is content transparency, i.e., whether the advertisement explicitly states its commercial character (e.g., an advertisement or sponsored); another is source transparency, i.e., whether the identity of the creative subject of the advertisement is clear (e.g., an AI-generated label). With regard to content transparency, Liao Bingyi and Liang Xuanxuan [6] operationalized transparency by conceptualizing it as the audience's perception of clarity regarding the commercial nature of the advertisement, whereas Peng Xin et al. [7] concentrated on the recognizability of native advertising and analyzed its interaction effect with the type of publisher. With respect to source and algorithmic transparency, Hong Jiewen and Chang Jingyi [8] viewed algorithmic transparency as the sense of the degree of algorithmic intelligibility, predictability, and controllability by the users, and discovered that "transparency" is not always equated to "trust". Also, studies of related concepts like behavioral targeting advertising transparency [9] and online service recovery transparency [10] have also been used as sources of reference to the measurement and operationalisation of AIGC advertising transparency.

Based on the above research, this paper will define AIGC advertising transparency as: the extent to which an advertisement is informative to consumers on the fact that the content is created or aided by artificial intelligence technology, with two dimensions: whether to disclose and how to disclose (the location, wordings, and salience of disclosure).

4. Research on the Relationship Between AIGC Advertising Transparency and Consumer Responses

4.1 Positive Effects of Transparency

Other researchers have indicated that AIGC advertising transparency might have a positive effect on consumer response in some circumstances, and the evidence chain has first emerged in Chinese and foreign literature. Liao Bingyi and Liang Xuanxuan [6], in a domestic study of Douyin short-video native advertising, found that high-transparency native advertising could significantly enhance the advertising attitude, product attitude, and purchase intention of audiences; perceived trust plays a positive mediating role, and brand familiarity positively moderates the relationship between transparency and perceived trust; the improvement in trust due to transparency is much higher among brands familiar to users. Peng Xin et al. [7] considered the recognizability of native advertising and discovered that recognizability can be used to enhance brand attitude by lowering cognitive resistance among consumers and, hence, “understandable advertising” is more effective at winning over consumers compared to “hidden advertising.” Internationally, Do et al. [11], in a study that involved the use of scenario questionnaires (N = 300) to measure the impact of transparency disclosure on consumer trust and credibility perception towards AI-generated advertising, found that there is a positive effect of transparency disclosure on consumer trust and credibility; Le et al. [12], in experimental research, added that the effect of AI role disclosure transparency varies depending on how consumers perceive the motive behind the company disclosure-if the disclosure is presented as a proactive indicator of the sincerity of the company, high transparency can alleviate the uncertainty of customers; Zou He et al. [13], in a real-world marketing customer acquisition experiment, demonstrated that the active revelation of AI customer service identity could actually increase consumer trust and conversion, which offers causal evidence in a real market setting on the premise that transparency creates value. Kim and Duhachek [14], in terms of psychological distance, noted that AI tends to be more influential when the advertising information is psychologically closer to the consumer, so that the disclosure of AI identity does not necessarily diminish, but can even increase, the persuasiveness in the context of so-called close-distance appeals.

4.2 Negative Effects of Transparency and the “Transparency Paradox”

Nevertheless, transparency has not necessarily led to positive outcomes, and numerous studies have identified the negative impacts of disclosure. Domestic research indicates that Gu Chenyan et al. [15] experimentally validated that model realism in AI-generated advertising influences consumer cognitive judgments and behavioral intent via perceived eeriness, and offers direct evidence of the formation of the so-called uncanny valley mechanism in the context of advertising. As per the international literature, Chen et al. [16] performed numerous experiments to explore consumer perceptions related to AI-generated advertising, and they concluded that appeal type and the social role of AI influence the resulting outcome, and if the content is perceived to be non-human created, the attitudes of consumers towards the advertisement would decrease significantly; Franke et al. [17], who investigated virtual influencers as endorsers, determined that the responses of consumers to virtual endorsers are dichotomous, where some consumers feel that it is a new form that is effective, whereas other consumers feel that the form is eerie and deceitful, and upon discovering the concealed identity of the virtual identity “disguise”, it provokes a high level of deception and brand punishment behavior. In a systematic review, Li Yan et al. [18] also outlined two negative principles of virtual influencer marketing: the first is that after awareness to the falsity is turned on, the general assessment of the authenticity of the content and credibility of the source decreases; the second is the replacement of human laborers with AI, which causes moral disgust or sympathy and subsequently leads to negative evaluation of the brand. It has been found by Crollic et al. [19] that in service recovery scenarios, the more human-like a chatbot is intended to be, the more likely it is to increase consumer anger -the greater the extent of anthropomorphism, The lesser the tolerance of customers towards its mistakes. All these results lead to the so-called transparency paradox: the disclosure itself is an honest thing, but it may happen that the truth revealed by the honesty is precisely what consumers do not wish to learn.

4.3 Reconciliation of Positive and Negative Effects: A Task-Technology Fit Perspective

As indicated by the above literature, the results concerning the type of influence transparency finally exerts on the buyers are contradictory. What measures could be taken to balance the discrepancy? Task-technology

fit and algorithm aversion research provide an important line of thought. Castelo et al. [20] determined that the degree to which consumers dislike algorithms is a situational factor based on the given task: when it comes to mechanical repetition, individuals are prepared to use algorithms; when it comes to tasks that require thinking, people choose human judgment, and creative advertising lies somewhere between the two. Consumer acceptance of AI also varies across both the types of consumption, i.e. utilitarian and hedonic consumption, as Longoni and Cian [21] have further demonstrated that AI recommendations are perceived as superior in the utilitarian context, whereas in the hedonic context, human judgment is perceived as superior. The famous experiment conducted by Dietvorst et al. [22] stated that even though algorithms were better at overall performance compared to humans, after seeing an algorithm make a mistake, people would stop using it (algorithm aversion). Granulo et al. [23] note that when it comes to symbolic consumption to indicate who I am, the preference for human labor is especially strong among people. In the domestic sphere, Du Jiangang et al. [24] discovered that diverse motivations of consumers (including efficiency orientation and meaning orientation) have a systematic effect on the type of AI-generated material preferred by them. Together, these studies demonstrate that the direction of the effect of AIGC advertising transparency depends on how consumers perceive the notion of suitability of AI to complete the task. When the task is perceived as suitable, disclosure is more likely to result in positive effects, and when perceived mismatch takes place, disclosure leads to negative responses due to the presence of non-human cues. It is this viewpoint that shifts the discussion in Chapter 3 away from the question of whether there is an effect to when there is an effect, and becomes the central point in bringing together the subsequent theoretical framework.

5. Theoretical Explanations of the Mechanism

5.1 Positioning and Relationship of Four Theories: A Temporal Sequence Explanatory Chain

By examining the available literature, it can be concluded that the uncanny valley theory, CASA theory, the persuasion knowledge model, and attribution theory have been extensively employed to elucidate the impact of AIGC advertising transparency, but most studies only adhere to one theory. The interrelationship of the four theories, the causal link of the effect chain explained by each, and in what circumstances which theory would be more appropriate remains unclear. In this paper, the author will claim that these four theories are not competing or mutually exclusive; rather, these theories are linked to various points of the temporal sequence of psychological processing of consumers who interact with AI disclosure information, creating a sequential explanatory system of the form of expectations formation - expectancy violations - defense triggering - attributions: in case a consumer initially sees an advertisement, he/she uses the CASA mechanism and mind perception to create expectations concerning the source of content (formation of expectations); the unexpected display of the AI label violates this expectation and provokes a feeling similar to the uncanny valley and expectancy violation (the violation of expectations); once the expectation is violated, the persuasion knowledge of consumers is triggered, and this leads to the defensive processing stage (activation of defense); lastly, consumers evaluate disclosure intentions and content quality and develop lasting attitudes and behaviors (attribution judgment). These links can be explained independently and also have tight connections with each other due to their sequential transmission; and different contextual conditions cause certain links to become the predominant channel of the effect. The next sections give detailed explanations.

5.2 Expectation Formation: CASA Theory and Mind Perception Theory

The theoretical framework of the CASA (Computers Are Social Actors) was presented by Nass and Moon [25]. Their main assumption was that, irrespective of the fact that one knows that he or she is interacting with a machine, he or she has instinctive and subconscious social responses to computers as though he or she were interacting with other individuals. Media equation theory is the precursor to this theory- Reeves and Nass [26] discovered that when humans come across media technology, they exhibit a sort of social instinct; they subconsciously regard it as an entity with some level of social functionality, even though they may be aware that the media is basically an inanimate device. Mind perception theory explains why people do this: Gray et al. [27] reported that people view the minds of other people and machines in two ways: agency (the ability to act and plan) and experience (the ability to feel and perceive); Deng Shichang et al. [28] applied this model in the context of AI services and showed that the way consumers attribute agency and experience to AI influences

the mechanisms and promotional strategies of AI acceptance. The theory of avatar marketing, developed by Miao et al. [29], also suggests that the behavior of consumers towards virtual endorsers can be described as perceived humanization - social connection - brand outcomes.

To understand the beginning of the transparency effect in the context of AIGC advertising, this group of theories defines what happens when a consumer looks at an advertisement: the consumer subconsciously perceives the creator (or sponsor) as a social actor, and according to interpersonal interaction scripts, expects by default that a competent creator would write something intended to be read by himself. Specifically, in advertising situations, it frequently means the assessment that the creator is human, since ad creativity has been historically considered to be a human-only activity. The presence of this default expectation is the basis of all further reactions (positive or negative).

5.3 Expectation Violation: Uncanny Valley Theory and Expectancy Violation

The uncanny valley theory was developed by Masahiro Mori, a Japanese roboticist, in 1970 [30]. The basic idea is: as the human likeness of a robot is relatively low, the affinity of people towards it will increase with the increase in its human likeness; however, as human likeness increases to some critical level, people feel strong eeriness and repulsion, and it is this trough region that is called the uncanny valley (a new English translation can be found in Mori et al. [31]). There are two major theories presented by academia to account for the uncanny valley, including the lateral inhibition hypothesis, which states that it is a basic survival response of humans to external threats, and the mind perception theory, which argues that the uncanny valley comes out of a mismatch between high agency appearance and low experience or entities that are perceived to be human-like but possess no mind are the most disturbing [27]. A study by Stein et al. [32] revealed that acceptance of a digital entity depends upon what kind of interaction between appearance anthropomorphism and mental capacity is present: the entities that are extremely human-like but are low in mental capacity are the ones with the greatest likelihood of being rejected.

One must note that the uncanny valley is not activated in each communication situation. Zhang Fang and Xu Zihan [33] conducted an experiment involving a social robot and discovered that anthropomorphism had a significant positive effect on favorability in the formation of human-machine impressions; Wang Yuanxin and Han Zhuoyan [34] also reported in their study on virtual news anchors that highly simulated anchors were even more liked by audiences, and the uncanny valley did not play a significant role in that setting. This implies that the uncanny valley triggering has contextual limits: when an AI entity is seen as a useful tool rather than a disguised person, anthropomorphism evokes more sympathy than strangeness. The advertising industry research by Gu Chenyan et al. [15] established that AI-generated model realism influences consumer reactions via eeriness, and the study by Chen et al. [16] demonstrated the connection between punitive attitudes towards AI-generated advertising and the perception of eeriness mechanism. It is evident that advertising is indeed a high-trigger zone of the uncanny valley because consumers expect a high degree of authenticity with regards to the people used in advertisements (models, endorsers). Besides the uncanny valley, the concept of expectancy violation offers a parallel explanation: consumers have a default perception that advertising content sources are made by humans, and AI label disclosure violates such perceptions, resulting in negative expectancy violation, which thus reduces trust ratings. It may be stated that the uncanny valley is the explanation of discomfort in the emotional channel and expectancy violation is the explanation of dissonance in the cognitive channel, and taken together, they form the link of the so-called expectation violation.

5.4 Defense Activation: Persuasion Knowledge Model

The Persuasion Knowledge Model (PKM) was introduced by Friestad and Wright [35]. According to the model, over socialization, consumers develop a cognitive structure regarding marketing persuasion- consumers are not passive receivers of information but rather utilize persuasion knowledge in order to counteract external persuasion efforts; once persuasion knowledge is engaged, consumers assess the intentions behind the information source and subsequently begin defensive processing. Campbell and Kirmani [36] also noted that the activation and application of persuasion knowledge is situational, based on cognitive capacity and cue salience of the consumer, whereby the more salient the disclosure cue, the higher the likelihood of activating defense.

The persuasion knowledge effect of disclosure has enough empirical data in the sphere of native advertising and sponsored content: according to Wojdyski and Evans [37], where disclosure is located and what language

is applied determines how easy it is for consumers to identify the fact that the native advertising is a commercial message; the greater the recognition, the less the evaluation of the advertisement. The research by Boerman et al. [38] showed that the longer the period of sponsorship disclosure, the more active the persuasion knowledge and the worse the brand evaluation. The recognizability of native advertising has been replicated and expanded to the Chinese setting by Peng Xin et al. [7] - recognizability of native advertising influences brand attitude via cognitive resistance as a mediator. Moving this concept to the AIGC advertising: AI label disclosure will reveal the so-called source cues, which not only inform other people about the creator of the content but also indicate that they are currently being subjected to persuasion. After persuasion knowledge of consumers has been activated their assessments of truthfulness and beneficence of advertising information decrease generally and become defense reactions. This association clarifies why, after the violation of expectation, customers do not silently take it but shift into an aggressively defensive mode.

5.5 Attribution Judgment: Attribution Theory

The theory of attribution was developed by Heider [39], and it involves how people attribute occurrences to internal or external causes, and how different styles of attributions influence emotions and further behavior. In the advertising environment of AIGC, post-disclosure, there is a critical attribution task that consumers must undertake: Exactly what does AI presence in the advertising creation signify? Le et al. [12] demonstrated that the direction of disclosure motive framing influences attribution conclusions-if consumers perceive disclosure as a sincerity signal that is actively communicated by the company, they will assign high-quality content to the category of advanced technology empowering the brand and provide positive feedback; if consumers perceive disclosure as a passive compliance under regulatory pressure, they will attribute the issue to the fact that the brand has no real creativity and will give negative feedback. Hong Jiewen and Chang Jingyi [8], in the study of algorithmic transparency, identified an isomorphic effect: platforms use so-called “compliance-oriented disclosure” (piled-up terminology, passive notice to the public), and this is perceived by users as an attempt to avoid responsibility, which leads to a paradox of “increased transparency = increased suspicion.” The attribution link is actually a setting step in the process of consumer response: the cognitive conflict and emotional arousal that have been created in other links are synthesized into a steady view of the brand and advertisement, and subsequently turned into purchase intention, avoidance, or sharing behavior.

5.6 Summary of This Section: Integration of the Theoretical Chain

The division of labor and the relevant conditions of the four theories in the temporal sequence chain are presented in Table 1. Three points should be noted: Firstly, this chain is directional, i.e., if there is no social expectation created by the CASA mechanism, then the concept of “violation” cannot be raised. Secondly, the power of every connection is regulated by the context; e.g., very realistic visual advertising is more prone to activate the uncanny valley link, whereas less salient disclosure is more prone to escape the persuasion knowledge link, and framing disclosure with great sincerity is more prone to evoke positive attribution. Thirdly, this chain is not the only route - when consumers are highly aware of task-technology compatibility, the violation of expectations can directly cause positive attribution (e.g., AI is superior to people in that regard. This is exactly the theoretical origin of why the transparency effect has double-edged sword qualities.

Table 1: Division of Labor and Applicable Conditions of the Four Theories in the Temporal Sequence Chain

Chain Stage	Dominant Theory	Explanation	Representative References
Expectation Formation	CASA Theory, Mind Perception Theory	Consumers perceive AI content/endorsers as social actors and form a default expectation of “human-created” content.	[25-29]
Expectation Violation	Uncanny Valley Theory, Expectancy Violations Theory	AI disclosure violates source expectations, triggering perceived eeriness and cognitive dissonance.	[30-32]
Defense Activation	Persuasion Knowledge Model (PKM)	Disclosure activates persuasion knowledge, and consumers initiate defensive processing.	[35-38]
Attribution Judgment	Attribution Theory	Consumers attribute disclosure motives and content quality, forming stable attitudes.	[12, 39, 8]

6. Boundary Conditions and Boundary Variables

The magnitude of any given relationship within the theoretical chain described above is not constant; it is influenced by many factors. The current chapter systematically organizes some moderating variables spread throughout articles at four levels: consumer individual differences, brand and product factors, advertising format, platform, and disclosure context, and it concentrates on how advertising format (video/image/text) affects the results differently.

6.1 Consumer Individual Differences

First, AI literacy and AI knowledge. A description by Ng et al. [40] defines AI literacy as a collection of skills to critically assess AI technologies and to communicate and work together with AI, which suggests that AI literacy is a complex idea. In respect to moderating effects, Castelo et al. [20] indicated that individuals having greater cognitive reflection ability are more capable of distinguishing task types and regulating their dependence on algorithms; Du Jiangang et al. [24] indicated that consumers' motivational types (efficiency-oriented versus meaning-oriented) influence the direction of their preference towards AI-generated content. It is also important to note that the existence of an AI literacy paradox might be possible: on the one hand, enhanced literacy can assist consumers in determining whether AI and a task match, thus enhancing their evaluations, and on the other hand, it could also increase their sensitivity to the problems and manipulative nature of AI content, thus increasing negative reactions. The direction of this paradox is currently undefined and represents an important gap in the field of future research.

Second, cognitive capacity and regulatory focus. According to the classic work by Campbell and Kirmani [36] the consumption of cognitive resources is involved in the use of persuasion knowledge, and those consumers whose cognitive load is higher are less able to initiate defense mechanisms, which implies that under information overload, the effect of AI labels on persuasion knowledge can be weakened. The local evidence of prevention-focused consumers is provided by the cognitive resistance mediation reported by Peng Xin et al. [7]: people with high cognitive resistance tendencies are more responsive to the commercial nature of advertisements, and transparency adversely affects their attitudes.

The third point is perceived control. Hong Jiewen and Chang Jingyi [8] have shown that the perception of users about their control over an algorithm is one of the mediators that transform the concept of transparency into trust - if the provision of information is given, but control is not granted, it is hard to transform transparency into trust. This suggests that AIGC advertising transparency studies need to go beyond informing consumers and extend to what consumers receive after being informed (the right to know, the right to choose, and the right to exit).

6.2 Brand and Product Factors

The first two factors to consider here are brand strength and brand familiarity. Brand is one of the external clues used by consumers in assessing the advertising message: with strong brands, consumers prefer to perceive AI-generated content as a quality extension under technological empowerment, and brand equity dampens the effect of disclosure, whereas with weak brands, the same disclosure is interpreted as a sign of cost-reduction and insincerity. According to research by Liao Bingyi and Liang Xuanxuan [6], brand familiarity has a positive moderation impact on the relationship between transparency and perceived trust, whereas Kang Haiyan et al. [10] proposed that brand authenticity is also a precondition that should be met to ensure transparency plays its beneficial role, i.e., only in combination with the authentic brand narrative can consumer distrust be eliminated.

Second, the kind of product. The research conducted by Zhou et al. [41], published in the *Journal of Marketing*, indicates that there is sensory specificity to the effectiveness of virtual influencers-where the benefits are most apparent in the category of products where the experience of the senses or visuals dominates; the asymmetry between utilitarian and hedonic values of Longoni and Cian [21] supports this view; Granulo et al. [23] extended the limit even further to symbolic consumption- where products have the expressive role of who I am, the preference of consumers for human creator labour is at its peak. Mende et al. [42], in the context of service robots, reported service-side support: the negative impact of humanoid robots is increased in services with higher social presence. In general, the core of product type moderation is the distinction

between the perceived AI competence boundary: the more similar the product task is to the AI competence comfort zone, the less the negative influence of disclosure is.

6.3 Differential Effects of Advertising Format (Media Format)

The current studies generally consider AIGC advertising as a single stimulus, but in reality, various advertising formats vary significantly in realism signals, social presence, and intensity of consumer expectation, and this effect is then differentiated as the transparency effect. It is the moderating dimension that the present paper regards as the most important to discuss but poorly addressed in the existing literature. It is addressed further on by format type.

The first one: text/copy format. The AI copy is the oldest commercialized and the least anthropomorphic format, and the post-disclosure effect is predominantly cognitive rather than emotional. Deng et al. [43] have created a customized advertising copy intelligent generation system, which demonstrates that AI is more efficient to use in data-driven relevance copy; Kim and Duhachek [14] have shown that in case of a shorter psychological distance, AI persuasion is, conversely, higher. Then, AI disclosure in the form of copy is more prone to be positively attributed, e.g., as task matching, and the uncanny valley process is not activated since the text does not have a human-like appearance. Nevertheless, this hypothesis is yet to be tested empirically.

Second, static visual formats (images, posters, AI models). Visual format is a high trigger area of the uncanny valley mechanism: Gu Chenyan et al. [15] confirmed explicitly that the realism of AI-generated models impacts the responses of consumers via perceived eeriness; Chen et al. [16] state that the evaluation direction of visual advertising depends on the match between appeal type and AI role. The abundance of cues in static visuals can be enough to ensure that it passes as real until consumers realize that there is no real person in the image and thus activate awareness of falseness and dual doubts about the authenticity of the content and sincerity of the brand.

Third, video and virtual digital human dynamics. Video and digital human forms are the most complicated of all four formats because both have a high level of realistic appearance and a high social presence: on the one hand, social presence and parasocial interaction theory predict that digital humans can develop quasi-interpersonal relationships (Miao et al. [29]; Xu Zhennan et al. [44]); Cheng Siqi et al. [45] defines virtual digital humans as an experiential medium; the immersion and persuasive power of virtual AI streamers are shown to be significantly higher by Zhang et al. [46]; on the other hand, when the non-human identity of a digital human is revealed, the feeling of betrayal of passing like real people and the uncanny valley discomfort are also at their peak-the extreme responses noted down by Franke et al. [17] and the awareness-of-falsity mechanism described by Li Yan et al. [18] are best realized in this form. Meanwhile, here the contextual limit of the uncanny valley is also established: when a digital human is viewed as a brand avatar, but not a disguised person (e.g. Wang Yuanxin and Han Zhuoyan's virtual anchors [34], and Zhang Fang and Xu Zihan's social robots [33]), high realism will instead increase favorability.

Generally speaking, there are two main principles behind moderating the format of advertisements: the first is that the amount of realism cues dictates the likelihood of inducing the uncanny valley and the realization of falsity text < image < video/digital human. The second one is that the level of social presence defines the magnitude of CASA expectations- the higher the expectation, the higher the effect produced by the disclosure breaking it. This interaction structure of format x disclosure can be directly applied to a more practical context, including the question of what type of disclosure strategy is appropriate to what type of advertising format.

6.4 Platform Context and Disclosure Mode

The transparency effect also relies on the ecosystem of the platform where disclosure takes place, as well as its framing. Concerning the way of disclosure, Wojdyski and Evans [37] determined that the placement of disclosure and the language used to disclose commercial intent have significant effects on the recognition rate of commercial intent- corner small print and clear labeling as advertisement vary significantly in the level of persuasion knowledge activation; Boerman et al. [38] determined that time of disclosure also influences the development of persuasion knowledge. Concerning the framing of disclosure motives, Le et al. [12] indicated that proactive frames reduce uncertainty, whereas passive frames enhance diagnostic suspicion; the research of Hong Jiewen and Chang Jingyi [8] on the trust paradox serves as a warning on the side of the platform-mechanical, templated adherence to disclosure not only fails, but it can even backfire. Concerning overall

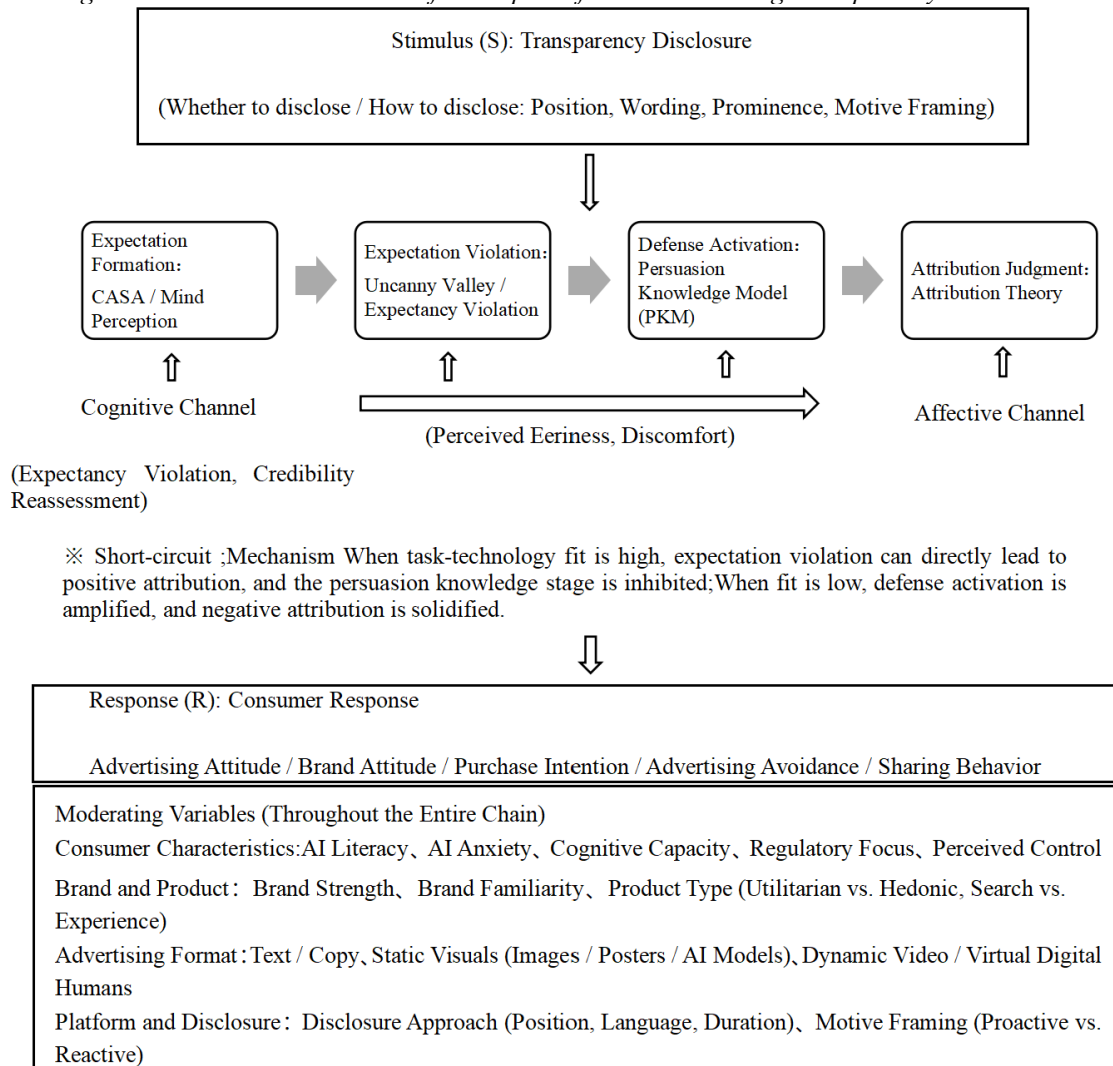
trends of algorithmic transparency, Kizilcec [47] noted that the most conducive to the calibration of trust is a moderate level of transparency, and Logg et al. [48] observed that people in general tend to appreciate the concept of the algorithms - this offers a cognitive rationale behind the idea of why transparency might be helpful. Also, Wang Yonggui et al. [9] discovered that the intermediary mechanism through which transparency impacts advertising performance through depth of information processing is moderated by advertising appeal: the positive influence of transparency is greater when it is rational appeals and lesser when it is emotional appeals - this is a very high consistency with the task-technology fit reasoning (AI is more competent at rational tasks).

7. Theoretical Integration and Research Gaps

7.1 Integrated Theoretical Framework

Based on the integration of the review conducted in Chapters 3-6, the present paper suggests a unified theoretical conceptualization of the influence of AIGC advertising transparency on consumer behavior. As per the mechanism integration diagram illustrated in Figure 2. This framework is also supported by the concept of a "temporal sequence chain" as described in Chapter 5, and it is consistent with the SOR paradigm, where transparency disclosure acts as the external stimulus (S) that has four mental processing links, i.e., expectation formation, expectation violation, defense activation, and attribution judgment (O) before eventually leading to the responses, such as advertising attitude, brand attitude, and purchase intention (R). Three interacting logics are involved in the framework:

Figure 2: Integrated Theoretical Framework of the Impact of AIGC Advertising Transparency on Consumer Responses



The first one, that is, dual-channel parallelism. The cognitive channel (expectancy violation, credibility reassessment) and the emotional channel (perceived eeriness, discomfort) work together in the context of the expectation violation. The importance of either of the two channels depends on how realistic the advertising format is: image and video formats focus on the emotional channel, whereas the copy format focuses on the cognitive channel.

And second, the chain can be short-circuited. When consumers perceive their match to technology very positively, i.e., when they have a high sense of task-technology fit, then the expectation violation directly results in positive attribution (e.g., AI is indeed superior at this), and the persuasion knowledge link is suppressed; whereas, with low fit perception, there is a magnification of defense activation, and negative attribution is consolidated. It clarifies the difference in the direction of effect.

Third, contextual moderation is a part of the whole chain. Different links in the chain are activated with different strengths by consumer literacy and motivation, brand strength, product type, advertising format, and disclosure framing, respectively, which together define the direction and magnitude of the ultimate response.

7.2 Research Gaps and Future Directions

As a result of the aforementioned review of the literature, the paper recognizes the following gaps in the research. The first one is that the study of differentiation of disclosure framing is significantly inadequate. The available studies are generally directed at the existence or nonexistence of disclosure, but do not involve empirical studies of more specific language expressions like AI-assisted, AI-generated, and AI-participated creation, nor any systematic modification of disclosure salience and location- with the upcoming adoption of regulatory labeling systems, this is the gap nearest to practice.

Second, the chain-based integration of explanatory mechanisms has not been empirically tested. As presented in this paper, the sequence of processes will be expectation formation-expectation violation-defense activation-attribution judgment, but available literature does not present a whole picture. The sequence of how CASA, expectancy violation, and persuasion knowledge are connected, and where they short-circuit needs further confirmation using evidence of the process (time series measurement, eye tracker, and physiological indicators).

Third, the different advertising format effects (video/image/text) need to be studied systematically. As has been revealed by the analysis in Chapter 6, format is the major moderating dimension of transparency, yet there exists virtually no study comparing the variations in disclosure effects and mechanism weights between all three formats based on the same experimental context.

Fourth, competing predictions can be made regarding how AI literacy will play a role. Is better literacy associated with increased capacity to recognize matches (positive moderation) or to recognize manipulation (negative moderation), and under what conditions does it exist? Better measurements to measure literacy dimensions are required (refer to the paper by Ng et al. [40]).

Fifth, the interaction between the two variables of disclosure and content quality is not explored by any of the research. Most of the available literature examines the main effect of disclosure as well as controlling the quality of content; however, with AI content being at par with or surpassing the creative quality and emotional resonance of human content, will transparency continue to have negative consequences? The question of this so-called quality threshold directly dictates the cost-benefit of disclosure systems.

Sixth, the discussion devoted entirely to the Chinese local context is not enough. The ecosystem of platforms like WeChat, Douyin, and Xiaohongshu, and the algorithm governance scenario within the framework of the Qinglang special action (Hong Jiewen and Chang Jingyi [8]) present an unprecedented institutional setting to study AIGC transparency; the choice of novelty (as opposed to authenticity) over local consumers and their association with guochao AI narrative methods can also redraw the boundary of conclusions in the Western literature.

The seventh point is the transition between static transparency and dynamic transparency. The regulatory practice changes its course towards one-time labeling (as it had been practiced earlier) and the entire-process disclosure of the label, which is generated and is traceable (such as labeling synthetic content with metadata).

It has still not been addressed how dynamic transparency alters the direction of effects after the consumer learning process.

8. Research Summary and Outlook

This paper has sorted out the recent development of research on AIGC advertising transparency and consumer responses based on a systematic literature review. The general conclusion made is the following. First, the effect of transparency disclosure on consumers has the form of a double-edged sword: it can increase the acceptance of advertising with perceived trust, novelty perception, and task-match attribution, but can also reduce purchase intention with perceived eeriness, expectancy violation, and persuasion knowledge activation. Second, the discrepancy between positive and negative outcomes can be addressed through the task-technology fit viewpoint, which is based on the central question of whether consumers think that AI is appropriate to do such a task. Third, uncanny valley theory, CASA theory, and mind perception theory, the persuasion knowledge model, and attribution theory each support one of the temporal connections in the chain of events of expectation formation - expectation violation - defense activation - attribution judgment; the four form an integrative rather than competing explanatory system. Fourth, the direction and magnitude of the effect are moderated by various factors related to consumer characteristics (AI literacy, cognitive capacity, perceived control), brands and products (brand strength, product type), advertising format (video/image/text), and platform context (disclosure mode, motive framing).

In the future, further studies may elaborate in three directions, namely, firstly, the empirical testing of the chain structure and the interaction propositions of the format and disclosure that have been put forth in this paper, secondly, the evolutionary trajectory of the static disclosure to dynamic transparency and source disclosure to process disclosure, thirdly, the strategic foundations of the advertising practice in terms of when, how, and to which type of advertising the information should be disclosed in the Chinese local regulatory environment and platform ecosystem. It is anticipated that these directions will help drive AIGC advertising transparency research out of the discussion of whether it has any impact or not at all into the discussion of the construction of the mechanism chain, and add to the consumer behavior research evidence base on the regulation of AI content labeling globally.

References

- [1] Zhibin Jiang & Yangzhou. (2023). Stage-based disruptive innovation: A study on the impact of generative artificial intelligence on the advertising industry. *Editorial Friend*, (11), 59–66.
- [2] European Parliament and Council. (2024). Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). *Official Journal of the European Union*.
- [3] Cyberspace Administration of China, National Development and Reform Commission, Ministry of Education, et al. (2023). *Interim Measures for the Management of Generative Artificial Intelligence Services*.
- [4] Baiyang Li, Yun Bai, Gang Li, et al. (2023). The technical features and morphological evolution of artificial intelligence generated content (AIGC). *Documentation, Information & Knowledge*, 40 (1), 66–74.
- [5] China Academy of Information and Communications Technology, & JD Explore Academy. (2022). *White Paper on AI-Generated Content (AIGC)*.
- [6] Bingyi Liao & Xuanxuan Liang. (2024). Research on the impact of transparency of Douyin short video native advertising on advertising effectiveness. *Journalism Research*, (11), 51–69+119.
- [7] Xin Peng, Denfeng Yang, et al. (2024). “Each to their own taste, each to their own strength”: The impact of native advertising recognizability and publisher type on consumers’ brand attitudes. *Journal of Marketing Science*, 4 (4), 93–111.
- [8] Jiewen Hong & Jingying Chang. (2025). Transparency means trust? User perception and trust paradox in platform algorithm transparency practices. *Journalism and Communication Review*, (9), 31–45.

- [9] Yonggui Wang, Dongmei Liu & Li Yan. (2022). The impact of behavioral targeting advertising transparency on advertising effectiveness. *Business and Management Journal*, 44 (7), 159–174.
- [10] Haiyan Kang, & Demin Wan. (2025). The impact of online service recovery transparency on potential consumers' distrust: From the perspectives of brand authenticity and self-protection motivation. *Enterprise Economy*, 44 (9), 36–45.
- [11] Do, T. Y., Pham, T. Q. A., Liao, Y. K., & Huang, K. C. (2026). Transparency disclosure mechanisms for AI-generated advertising on online platforms. *Internet Technology Letters*, 9 (4), e70354.
- [12] Le, K. B. Q., Khan, H., Li, F., & Kunz, W. H. (2026). More than saying “it’s AI”: How role disclosure transparency in AI-generated ads influences persuasion. *Psychology & Marketing*, 43 (10), 2532–2548.
- [13] He Zou, Jia Gai, Yu Feng, et al. (2025). The impact of disclosing AI customer service identity in customer acquisition: Evidence from a field experiment. *Journal of Marketing Science*, 5 (3), 124–144.
- [14] Kim, T. W., & Duhachek, A. (2020). Artificial intelligence and persuasion: A construal-level account. *Psychological Science*, 31 (4), 363–380.
- [15] Chen Yangu, Siyu Changxin, et al. (2026). Cyber model vs. simulated image: The impact of model realism in AI-generated advertising on consumer cognition and behavioral intentions. *Journal of Marketing Science*, 6 (1), 134–152.
- [16] Chen, Y., Wang, H., Rao Hill, S., & Li, B. (2024). Consumer attitudes toward AI-generated ads: Appeal types, self-efficacy and AI’s social role. *Journal of Business Research*, 185, 114867.
- [17] Franke, C., Groeppel-Klein, A., & Müller, K. (2023). Consumers’ responses to virtual influencers as advertising endorsers: Novel and effective or uncanny and deceiving? *Journal of Advertising*, 52 (4), 523–539.
- [18] Yan Li, Wei Chen, & Ruijuan Wu. (2025). Marketing effect of virtual influencers and its mechanisms in the context of AI technology. *Advances in Psychological Science*, 33 (8), 1425–1442.
- [19] Crolic, C., Thomaz, F., Hadi, R., & Stephen, A. T. (2022). Blame the bot: Anthropomorphism and anger in customer–chatbot interactions. *Journal of Marketing*, 86 (1), 132–148.
- [20] Castelo, N., Bos, M. W., & Lehmann, D. R. (2019). Task-dependent algorithm aversion. *Journal of Marketing Research*, 56 (5), 809–825.
- [21] Longoni, C., & Cian, L. (2022). Artificial intelligence in utilitarian vs. hedonic contexts: The “word-of-machine” effect. *Journal of Marketing*, 86 (1), 91–108.
- [22] Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, 144 (1), 114–126.
- [23] Granulo, A., Fuchs, C., & Puntoni, S. (2021). Preference for human (vs. robotic) labor is stronger in symbolic consumption contexts. *Journal of Consumer Psychology*, 31 (1), 72–80.
- [24] Jiangang Du, Zhangjian Wu, et al. (2025). Embracing AI generation: The impact of consumption motivation on preferences for AI-generated content. *Journal of Management Science*, 38 (1), 130–144.
- [25] Nass, C., & Moon, Y. (2000). Machines and mindlessness: Social responses to computers. *Journal of Social Issues*, 56 (1), 81–103.
- [26] Reeves, B., & Nass, C. (1996). *The media equation: How people treat computers, television, and new media like real people and places*. Cambridge University Press.
- [27] Gray, H. M., Gray, K., & Wegner, D. M. (2007). Dimensions of mind perception. *Science*, 315 (5812), 619.
- [28] Shichang Deng, Qi Xu, et al. (2022). User acceptance mechanism and usage promotion strategy of AI services based on mind perception theory. *Advances in Psychological Science*, 30(4), 723–737.
- [29] Miao, F., Kozlenkova, I. V., Wang, H., Xie, T., & Palmatier, R. W. (2022). An emerging theory of avatar marketing. *Journal of Marketing*, 86 (1), 67–90.

- [30] Mori, M. (1970). The uncanny valley. *Energy*, 7 (4), 33–35.
- [31] Mori, M., MacDorman, K. F., & Kageki, N. (2012). The uncanny valley. *IEEE Robotics & Automation Magazine*, 19 (2), 98–100.
- [32] Stein, J. P., Appel, M., Jost, A., & Ohler, P. (2020). Matter over mind? How the acceptance of digital entities depends on their appearance, mental prowess, and the interaction between both. *International Journal of Human-Computer Studies*, 142, 102463.
- [33] Fang Zhang & Zihan Xu. (2024). How to perceive AI interlocutors: An experimental study on the mechanism of anthropomorphism of disembodied conversational social robots' impact on impression formation. *Journalism and Communication*, (5), 46–57+96.
- [34] Yuanxin Wang & Zhuoyan Han. (2023). Enhancing interactive experience through realism: A study on the relationship between the realism of virtual news anchors and user acceptance. *Journal of Beijing University of Posts and Telecommunications (Social Sciences Edition)*, 25 (6), 11–19.
- [35] Friestad, M., & Wright, P. (1994). The persuasion knowledge model: How people cope with persuasion attempts. *Journal of Consumer Research*, 21 (1), 1–31.
- [36] Campbell, M. C., & Kirmani, A. (2000). Consumers' use of persuasion knowledge: The effects of accessibility and cognitive capacity on perceptions of an influence agent. *Journal of Consumer Research*, 27 (1), 69–83.
- [37] Wojdyski, B. W., & Evans, N. J. (2016). Going native: Effects of disclosure position and language on the recognition and evaluation of online native advertising. *Journal of Advertising*, 45 (2), 157–168.
- [38] Boerman, S. C., van Reijmersdal, E. A., & Neijens, P. C. (2012). Sponsorship disclosure: Effects of duration on persuasion knowledge and brand responses. *Journal of Communication*, 62 (6), 1047–1064.
- [39] Heider, F. (1958). *The psychology of interpersonal relations*. Wiley.
- [40] Ng, D. T. K., Leung, J. K. L., Chu, S. K. W., & Qiao, M. S. (2021). Conceptualizing AI literacy: An exploratory review. *Computers and Education: Artificial Intelligence*, 2, 100041.
- [41] Zhou, X., Yan, X., & Jiang, Y. (2024). Making sense? The sensory-specific nature of virtual influencer effectiveness. *Journal of Marketing*, 88 (4), 84–106.
- [42] Mende, M., Scott, M. L., van Doorn, J., Grewal, D., & Shanks, I. (2019). Service robots rising: How humanoid robots influence service experiences and elicit compensatory consumer responses. *Journal of Marketing Research*, 56 (4), 535–556.
- [43] Deng, S., Tan, C. W., Wang, W., & Pan, Y. (2019). Smart generation system of personalized advertising copy and its application to advertising practice and research. *Journal of Advertising*, 48 (4), 356–365.
- [44] Zhenan Xu, Qian Xu & Guilin Liu. (2023). The construction of personal brand virtual digital humans: Strategies of emotional practice. *Journal of Marketing Science*, 3 (4), 123–143.
- [45] Siqi Cheng Guoming Yu, et al. (2022). Virtual digital humans: An experiential medium—An analysis of the connection mechanism and media attributes of virtual digital humans. *Journalism and Communication*, (7), 12–23.
- [46] Zhang, X., Shi, Y., Li, T., Guan, Y., & Cui, X. (2024). How do virtual AI streamers influence viewers' livestream shopping behavior? The effects of persuasive factors and the mediating role of arousal. *Information Systems Frontiers*, 26, 1803–1834.
- [47] Kizilcec, R. F. (2016). How much information? Effects of transparency on trust in an algorithmic interface. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (pp. 2390–2395). Association for Computing Machinery. <https://doi.org/10.1145/2858036.2858402>
- [48] Logg, J. M., Minson, J. A., & Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes*, 151, 90–103.

Funding

This research received no external funding.

Conflicts of Interest

The authors declare no conflict of interest.

Acknowledgment

This paper is an output of the science project.

Copyrights

Copyright for this article is retained by the author (s), with first publication rights granted to the journal. This is an open-access article distributed under the terms and conditions of the Creative Commons Attribution license (<http://creativecommons.org/licenses/by/4.0/>).