

The AI Disclosure Paradox: How Disclosing AI-Generated Marketing Content Shapes Consumer Trust and Decision-Making—An Integrative Review with a Focus on Chinese Consumers

Zhuoli Luo*

University of California, Irvine, USA

**Corresponding author: Zhuoli Luo.*

Abstract

The use of artificial intelligence (AI) in the creation of social media posts, product descriptions, brand copy, and advertising images is growing. Businesses and regulators must determine how AI-generated marketing should be acknowledged as it becomes more commonplace. Existing evidence does not support a simple conclusion. Although an AI label can indicate transparency, innovation, and technological proficiency, it can also diminish the authenticity, credibility, and emotional sincerity of the content. The review results indicate that the effectiveness of disclosure depends on the meaning consumers assign to the label and whether the task is suitable for AI. Functional and utilitarian contexts tend to be more readily accepted, while emotional and highly personalized contexts are more likely to provoke consumer aversion to AI. Evidence from Chinese consumers supports both the benefits of novelty and the loss of authenticity. Therefore, China's mandatory labeling environment provides an important context for future field and cross-cultural studies.

Keywords

AI-generated content, AI disclosure, consumer trust, authenticity, Chinese consumers

1. Introduction

Generative AI has moved from an experimental tool to a practical part of marketing production. Firms can now use it to develop advertising images, write product descriptions, adapt brand messages, and produce social media content at a speed and scale that would be difficult through human labor alone. This development extends the broader influence of AI on marketing strategy and consumer behavior described [1]. However, the increasing popularity of generative artificial intelligence has also brought businesses a new choice: whether, when, and how to inform consumers that artificial intelligence has participated in the generation of information.

Disclosure is often defended as a matter of transparency, yet its persuasive consequences remain unsettled. Some studies suggest that an AI label can make an advertisement appear innovative and increase engagement. Other studies have found that the same labels can reduce credibility, empathy, authenticity, or behavioral intent.

Still other studies report that such labels are almost harmless or completely harmless. These findings constitute a paradox of AI-driven information disclosure that labels may convince consumers that a company is honest, but they also remind them that the content was not created by humans.

The question is particularly relevant in China. The Measures for Labeling Artificial Intelligence-Generated and Synthesized Content, which went into effect on September 1st, 2025 [2], define implicit labels that are contained in content files as well as explicit labels that users can see. As labels become more common, customer responses become a part of the regulated platform environment as well as a voluntary marketing concern.

This review addresses three questions. The first question asks how AI disclosure influences trust, perceived authenticity, attitudes, and behavioral intentions. The second question examines the positive and negative psychological mechanisms behind these responses. And the third one considers how task type, product type, emotionality, and consumer characteristics shape disclosure effects in the Chinese context. The article argues that disclosure has no fixed persuasive value. Its effect depends on what the label signals and whether consumers believe AI is appropriate for the task.

2. Review Scope and Approach

This paper does not employ statistical meta-analysis. Instead, it provides a comprehensive literature review, integrating theoretical discussions, experimental studies, and meta-analysis results on AI-generated marketing content disclosure. The research scope includes advertising images, brand information, charitable advertising, and social media content. Key findings examined include consumer trust, credibility, authenticity, attitudes, engagement, purchase intention, sharing intention, and donation intention. Only studies specifically focusing on AI performance that help explain the relevant context are included. No new survey or experimental data were collected. This approach is advantageous because research on AI-driven information disclosure is still developing, involving diverse theories, methodologies, and market environments. However, comparing these findings requires caution, as different studies use different disclosure labels, media formats, product types, samples, and outcome metrics, making direct comparisons impossible. Furthermore, many experiments only measure consumer intention after a single ad view, without observing repeated exposure or actual behavior. Therefore, this review focuses on common findings and factors potentially influencing consumer responses, rather than calculating an overall effect size.

3. Theoretical Framework

3.1 Signaling Theory

Signaling theory concerns situations in which one party possesses information that another party cannot directly observe [3]. Consumers can learn how a marketing message was created by reading a disclosure label. Disclosure may indicate honesty, transparency, technical competence, and respect for customers' right to know when it seems voluntarily or appropriately carried out. An AI label can also attract attention by marking the content as new or innovative. According to this interpretation, transparency promotes trust and lessens information asymmetry.

The signal is not automatically positive. The same label may imply that the firm invested little human effort, adopted AI mainly to save money, or cannot offer genuine emotional understanding. Thus, disclosure separates the fact being communicated from the meaning assigned to it. Consumers may appreciate the honesty of the disclosure but dislike what AI involvement suggests about the message. Signaling theory, therefore, explains why one label can produce competing inferences.

3.2 Persuasion Knowledge and Manipulative Intent

The Persuasion Knowledge Model proposes that consumers develop knowledge about marketers' goals and tactics and use it to cope with persuasion attempts [4]. An AI label can activate this knowledge by drawing attention away from the advertised object and toward the method used to construct the appeal. Consumers may ask why AI was used, whether the content represents a real experience, and whether the advertiser is trying to create an emotional reaction through synthetic material.

One possible consequence is an inference of manipulative intent. Campbell shows that attention-getting tactics can produce resistance when consumers believe the tactic serves the advertiser more than it benefits the audience [5]. In AI advertising, this inference may be especially strong when a synthetic face, voice, or story imitates human emotion. Manipulative intent is therefore best understood as a mechanism within persuasion knowledge rather than as a separate theory. Once activated, it can reduce message credibility and authenticity, which then weakens attitudes and behavioral intentions. However, information disclosure can also reduce concerns about manipulation by making the production process more transparent. The end result depends in part on whether the label is interpreted as a warning or evidence of honest communication.

3.3 Capability-Personalization Framework

The Capability-Personalization Framework helps explain why consumers may favor AI in some situation but prefer humans in others. Based on 442 effect sizes from 163 studies, Qin et al. found that AI appreciation is more likely when AI is considered more capable than humans and personalization is not necessary [6]. Negative reactions are more likely when AI seems less capable or when consumers expect individualized attention. Therefore, consumer responses depend on whether the strengths of AI match the requirements of the task.

AI might seem appropriate in marketing for data-driven targeting like factual product descriptions and standardized tasks. Charity appeals, luxury narratives, and identity-relevant messaging, on the other hand, may call for empathy, personal experience, or a specific human goal. Even excellent technical excellence might not be sufficient in some situations to meet the desire for customization. In line with task-dependent algorithm aversion, the theory thus predicts that the same revelation may be approved in an objective task and rejected in a subjective one [7].

The three viewpoints come together to provide a dual-path explanation. Information disclosure can have a positive impact through transparency, novelty, and perceptibility, or a negative impact by persuading knowledge, manipulating intent, and reducing credibility or authenticity. Context determines which route carries more weight.

4. Literature Review

4.1 Positive and Neutral Effects

Evidence for a positive pathway is clearest when disclosure adds novelty without challenging the message's core value. Shi and Jiang conducted three online experiments with Chinese participants and found that AI disclosure labels increased perceived novelty [8]. This increase improved attitudes toward the advertisement and product as well as purchase intention. The same studies also detected an opposing authenticity pathway, meaning that a favorable net response should not be interpreted as the absence of concern. Instead, novelty and authenticity operated at the same time.

Du et al. also found that AI-generated ads can increase psychological and behavioral engagement [9]. Psychological engagement helps explain behavioral engagement, and specific AI labels can influence behavioral responses such as likes, comments, and shares, although they do not significantly change psychological engagement. The study used a relatively simple ad and a well-known brand, so the findings are more convincing as evidence of a possible positive effect than as evidence that the label generally increases engagement.

Neutral findings are equally important to the paradox. Kirkby et al. tested brand texts presented as AI-generated, human-written, or undisclosed. AI disclosure did not significantly reduce perceived brand-voice authenticity or brand attitude [10]. Consumers may regard product specifications, descriptions, and familiar chatbot language as tasks that can be automated without violating expectations of genuine human experience. This result cautions against treating the disclosure label itself as a penalty. Transparency is acceptable and does not cause measurable harm to the brand when the use of AI is appropriate for the communication task.

4.2 Negative Effects

Negative effects become stronger when the message depends on trust in a real person or real suffering. Arango et al. examined AI-generated children's faces in charitable advertising [11]. Knowing that the faces were artificial lowered donation intention by reducing empathy and related emotional responses, including anticipatory guilt and perceived emotion. The problem was not merely that participants noticed a persuasion tactic. A psychological distance is created between the artificially created beneficiaries and the human needs that the advertisement originally intended to convey.

Baek et al. discovered a related but more cognitive process [12]. In prosocial advertising, the initial disclosure of AI led to a more negative attitude towards the advertisement, while consumers' perception of the advertisement's credibility played a mediating role, lower credibility also led to a decrease in willingness to donate. However, this negative reaction was not pronounced among participants who perceived AI as more human-like. This moderating effect suggests that consumers react not only to the category label "AI" but also to their perception of its understanding and expressive capabilities.

These findings suggest that transparency and trust should not be considered synonymous. Trust can still be compromised even if companies disclose their use of artificial intelligence if consumers perceive the source of the information as inappropriate. The ethical reasons for using artificial intelligence are also crucial. Arango et al. found that emphasizing the motivation to protect privacy improved user responses more than simply providing labels or emphasizing instrumental savings [11]. In emergency situations where real images are difficult to obtain, AI-generated images may be closer to real images. Therefore, the way the signal is interpreted can change people's perception of the motivation behind it.

4.3 Boundary Conditions

Task type has proven to be one of the more reliable boundary conditions in the relevant literature. Wu et al. proposed an important distinction here: in the eyes of consumers, AI used for ad delivery is not the same as AI used for ad creation [13]. When people believe that AI can handle complex tasks, they are significantly more willing to share ads delivered by AI than ads created by AI. The mechanism behind this difference is that people's perception of the objectivity of the task influences the machine's heuristic algorithms. Ad delivery is a data-driven, optimization task, which fits perfectly with people's existing assumptions about what machines are good at; while creative production is inherently subjective. In this area, machine involvement will instead trigger more scrutiny. In addition, it is worth noting that this effect is not consistent across all results; changes in advertising attitudes, brand attitudes, purchase intentions, and word-of-mouth intentions are not synchronous. This alone is enough to explain why the literature on this topic appears inconsistent across different studies, which do not always measure the same thing.

A similar pattern emerges when the comparison object changes from task type to product type. Shi and Jiang pointed out that utilitarian products amplify the positive "novelty" path while suppressing the negative "authenticity" path [8]. This is consistent with a common-sense observation: hedonic products rely heavily on sensory pleasure, symbolism, and emotional resonance. Therefore, compared to purely functional products, hedonic products will feel more strongly the lack of human intervention behind the content. This is quite consistent with the competence-personalization framework. Functional value can be verified to a large extent by objective standards, while hedonic value often depends on taste and experience, both of which are difficult to verify directly. In addition to task type and product type, there is another factor: the emotional intensity of the content itself. Du et al. found that emotional intensity has a negative impact on the relationship between AI-generated ads and user engagement [9]. In other words, the more an ad relies on emotional appeal, the more the source of the AI seems to have a negative impact on it, rather than becoming a neutral or accidental detail. Studies on charity advertising also show that when the persuasiveness of emotional appeals depends on empathy for real beneficiaries, such appeals appear fragile. The broader significance is not that AI cannot generate emotional content, but that there is a difference between technological authenticity and perceived sincerity.

Consumer characteristics also determine which inference becomes dominant. Promotion-focused consumers reacted more strongly to novelty and were less impacted by authenticity loss in Shi and Jiang's tests, while prevention-focused consumers placed more value on risk and credibility [8]. Prior familiarity with AI can also reduce uncertainty, but this familiarity may also make consumers more wary of synthetic manipulation.

Perceived similarity between AI and humans can mitigate negative reactions [12], while beliefs about AI capabilities can influence consumers' reactions to its performance of tasks [13]. These moderating factors answer the third research question, indicating that neither "the consumer" nor "AI information disclosure" are single, stable categories.

5. Key Findings

The first key finding is that simply disclosing information does not directly change consumers' level of trust; it merely provides users with a clue as to why AI is used and what impact AI has on information. Labels can indicate the freshness and authenticity of information, but they can also suggest a lack of human involvement or emotional understanding. Shi and Jiang observed that freshness and authenticity occurred simultaneously among Chinese participants, a phenomenon that explains the previous situation [8]. Therefore, the difference in results is not due to different consumer preferences for AI, but rather to different understandings of the same information disclosure. The Chinese regulatory environment serves as an example; the labeling measures in 2025 will cover AI-generated text, images, audio, video, and virtual scenes, requiring the use of visible labels or metadata-based labels in different scenarios [2]. In this case, while labels can indicate AI involvement, they may also indicate that the content is restricted by platform and regulatory rules. This example illustrates that the meaning of information disclosure may depend more on the context in which the label appears than on the text of the label itself.

The second finding is that matching tasks with information sources best integrates various forms of evidence. Consumers are more receptive when AI performs objective or standardized tasks, such as advertising and objective product descriptions [10, 13]; however, consumers are more resistant when information relies on empathy or life experience, such as in charity fundraising and hedonistic product scenarios [8, 11, 12]. Therefore, it is important not only whether AI is involved, but also whether consumers see AI as a legitimate source of specific information. This finding links actual evidence and capabilities with personalized frameworks. While capabilities can promote acceptance, they do not necessarily meet people's need for understanding [6]. This also explains why engagement, credibility, and purchase intention do not always change synchronously; task requirements and measurement results need to be considered together.

The third finding is that information disclosure is only useful when it helps consumers build appropriate trust. While labels like "AI" can indicate the source of information, they rarely explain why AI is used, the level of human involvement, or the security measures taken. There is evidence that information disclosure follows a common process that leads people to judge the ability, motivation and degree of human involvement of the information source. These judgments then influence people's views on the freshness, credibility, authenticity and emotional effect of the information, which in turn affect attitudes and behaviors. The specific results depend on which path has a greater impact. This comprehensive analysis adds an important distinction: transparency is not equivalent to calibrated trust. Labels such as "AI-generated," "AI-assisted," and "human-reviewed" lead to different judgments because they imply different degrees of human responsibility. This is especially important in China because mandatory labeling will make information disclosure more and more common. However, the evidence in China is currently limited to online experiments [8]. Therefore, it is necessary to directly test the understanding of AI familiarity, platform trust, regulatory trust and demographic differences, rather than relying solely on assumptions.

6. Implications and Future Research

When disclosing their AI usage, companies should align their wording with their marketing. When disclosing their use of AI, companies should tailor their wording to their marketing objectives, rather than using a generic template. When the content involves practical information, such as product specifications or price comparisons, it's best to directly highlight the speed, consistency, or accuracy advantages brought by AI. Consumers care about the reliability of the information itself, making this type of disclosure more readily accepted. However, in scenarios heavily reliant on emotional connection, such as luxury goods, entertainment content, or charitable fundraising, the same approach may not work. In these cases, consumers are less concerned with how the content was generated and more concerned with whether there was genuine oversight. Therefore, the focus of disclosure should shift to whether employees participated in the content review, editing, or final approval process. Simply labeling it "AI-generated" only describes the production method and fails to

answer the more crucial question—why AI is used? This latter point forms the basis for consumers' judgment of a company's credibility and sincerity. However, it's important to note that disclosure itself is not a panacea. If consumers feel that the emotional appeal of the advertisement has been distorted or insincere, then no matter how detailed the label or how reasonable the explanation, it will be difficult to salvage the technological transparency or fill the content gaps.

Most existing research relies on one-off experiments, where participants view an advertisement and fill out an attitude questionnaire to infer consumer reactions. The limitation of this design is that it captures more of a “one-off judgment in a laboratory,” which may not accurately reflect the actual behavioral patterns of consumers when repeatedly exposed to information on Chinese social media and e-commerce platforms. Future research would be more convincing if it shifted towards field experiments or real platform data. The details of the labels also deserve further exploration. Terms like “AI-generated,” “AI-assisted,” and “human-reviewed” suggest different levels of human involvement, which may lead to different judgments from consumers regarding ad engagement and authenticity. However, few studies systematically compare the effects of these different expressions.

Furthermore, cross-cultural comparison is a valuable but easily overlooked area. For comparisons to be truly meaningful, participants from different countries must be exposed to the same content and label formats. Only by controlling for these variables can we determine whether the observed differences stem from culture itself or from other confounding factors such as regulatory environments, platform usage habits, or translation wording.

7. Conclusion

Overall, existing literature does not offer a simple answer to whether AI-related information disclosure improves or harms consumer responses; the same label can have diametrically opposed effects on different consumers. Some consumers will see it as a sign of corporate honesty and a proactive pursuit of technological advancement, while others will question the credibility, sincerity, or purpose of the information. Signaling theory helps explain the first reaction, as disclosure conveys openness and innovation; the persuasive knowledge model explains the second reaction, showing that labels alert consumers to corporate persuasive strategies, leading to concerns about manipulation. The Capability-Personalization Framework adds an important condition: consumers are more accepting of AI when it appears suitable for a specific task and personal understanding is not a key factor. As AI-generated labels become increasingly prevalent in the Chinese online environment, these issues become even more critical; ultimately, consumers do not react solely to the label itself, but their judgments depend on what information they believe the label reveals and how much meaningful human intervention remains within that information.

References

- [1] Davenport, T., Guha, A., Grewal, D., & Bressgott, T. (2020). How artificial intelligence will change the future of marketing. *Journal of the Academy of Marketing Science*, 48, 24–42. <https://doi.org/10.1007/s11747-019-00696-0>
- [2] Cyberspace Administration of China. (2025, March 14). Measures for labeling artificial intelligence-generated and synthesized content. https://www.cac.gov.cn/2025-03/14/c_1743654684782215.htm
- [3] Connelly, B. L., Certo, S. T., Ireland, R. D., & Reutzel, C. R. (2011). Signaling theory: A review and assessment. *Journal of Management*, 37(1), 39–67. <https://doi.org/10.1177/0149206310388419>
- [4] Friestad, M., & Wright, P. (1994). The persuasion knowledge model: How people cope with persuasion attempts. *Journal of Consumer Research*, 21(1), 1–31. <https://doi.org/10.1086/209380>
- [5] Campbell, M. C. (1995). When attention-getting advertising tactics elicit consumer inferences of manipulative intent: The importance of balancing benefits and investments. *Journal of Consumer Psychology*, 4(3), 225–254. https://doi.org/10.1207/s15327663jcp0403_02

- [6] Qin, X., Zhou, X., Chen, C., Wu, D., Zhou, H., Dong, X., Cao, L., & Lu, J. G. (2025). AI aversion or appreciation? A capability–personalization framework and a meta-analytic review. *Psychological Bulletin*, 151(5), 580–599. <https://doi.org/10.1037/bul0000477>
- [7] Castelo, N., Bos, M. W., & Lehmann, D. R. (2019). Task-dependent algorithm aversion. *Journal of Marketing Research*, 56(5), 809–825. <https://doi.org/10.1177/0022243719851788>
- [8] Shi, Y., & Jiang, Z. (2026). Consumer responses to AI disclosure labels: The role of novelty and authenticity. *SAGE Open*, 16(1), 1–18. <https://doi.org/10.1177/21582440261417793>
- [9] Du, D., Zhang, Y., & Ge, J. (2023). Effect of AI generated content advertising on consumer engagement. In F. F.-H. Nah & K. Siau (Eds.), *HCI in business, government and organizations (Lecture Notes in Computer Science, Vol. 14039, pp. 121–129)*. Springer. https://doi.org/10.1007/978-3-031-36049-7_9
- [10] Kirkby, A., Baumgarth, C., & Henseler, J. (2023). To disclose or not disclose, is no longer the question—Effect of AI-disclosed brand voice on brand authenticity and attitude. *Journal of Product & Brand Management*, 32(7), 1108–1122. <https://doi.org/10.1108/JPBM-02-2022-3864>
- [11] Arango, L., Singaraju, S. P., & Niininen, O. (2023). Consumer responses to AI-generated charitable giving ads. *Journal of Advertising*, 52(4), 486–503. <https://doi.org/10.1080/00913367.2023.2183285>
- [12] Baek, T. H., Kim, J., & Kim, J. H. (2026). Effect of disclosing AI-generated content on prosocial advertising evaluation. *International Journal of Advertising*, 45(1), 171–192. <https://doi.org/10.1080/02650487.2024.2401319>
- [13] Wu, L., Dodoo, N. A., & Wen, T. J. (2025). Disclosing AI’s involvement in advertising to consumers: A task-dependent perspective. *Journal of Advertising*, 54(1), 20–38. <https://doi.org/10.1080/00913367.2024.2309929>

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.

Open Access

This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

