Proceedings of ICADSS 2026 Symposium: Contemporary Visual Arts, AI Creativity, and Interactive Media: Theory and Practice
DOI: 10.54254/2753-7064/2026.BJ36227

Beyond the Transparency Paradox: Contextual AI Disclosure and User Trust on Chinese Digital Platforms

Yiyu Pan

NYU Steinhardt School of Culture, Education, and Human Development, New York University, New York, USA
yp2544@nyu.edu

Abstract. This study explores how varying degrees of disclosure of artificial intelligence information affect users' trust and behavioral intentions on Chinese digital platforms. Based on recent controlled experiments, eye-tracking research, and statistics, this paper finds that the disclosure of artificial intelligence information can have a double-edged effect, depending on the specific situation. In utilitarian situations such as advertising and product search, information disclosure tends to improve user trust and perceived credibility. In social and hedonic situations, information disclosure tends to reduce user engagement and weaken the connection between users and content. Mechanism analysis shows that the driving factors of this pattern include: first, users' attribution bias toward errors labeled as AI-generated. Second, the increase in effectiveness and discomfort after information disclosure, and finally, the non-linear relationship between the level of information disclosure detail and trust. These mechanisms vary depending on the platform type, product type, and the user's artificial intelligence literacy. Based on these findings, this study proposes a three-part governance framework: a tiered information disclosure system, traceable records that meet the current regulatory requirements, and a human-in-the-loop design for high-sensitivity interactions. The framework challenges a common assumption that more information is better in artificial intelligence information disclosure. On the contrary, it adopts a different approach - a situation-oriented approach. The goal is to achieve a balance among transparency, user trust, and user participation on digital platforms.

Keywords: AI Transparency, Algorithm Aversion, Contextual Disclosure, User Engagement, Digital Platform Governance

1. Introduction

1.1. Research background

China's digital platforms are producing content at an amazing speed. Generative artificial intelligence (AI) plays an important role in this growth. According to the 53rd Statistical Report on China's Internet Development, released by the China Internet Network Information Center, as of December 2023, China's Internet users numbered 1.092 billion, and the Internet penetration rate was 77.5%. Short video users reached 1.053 billion, accounting for 96.4% of all netizens [1]. Such a

DOI: 10.54254/2753-7064/2026.BJ36227

large user group provides a wide audience for generative AI tools. According to a report released by the State Council Information Office of China, the Cyberspace Administration of China has approved the launch of 346 generative AI services by April 2025, most of which were added within a year [2]. This shows that AI-generated content is gradually becoming part of users' daily platform use.

At the same time, platforms face a difficult choice about how to mark AI-generated content. If the platform conceals AI participation, it may mislead users. However, if the platform strongly marks all AI-assisted content, users may gain a sense of distrust. Researchers call this tendency algorithmic aversion. Dietvorst, Simmons, and Massey define it as even if the machine's judgment is accurate, people tend to avoid it [3].

This contradiction ultimately comes down to trust, which, in turn, affects whether users continue to use the platform or make purchases. Therefore, the way the platform discloses the participation of AI is of great commercial and social significance. At the business level, unclear disclosure rules will reduce user participation and damage revenue. At the social level, it deprives the public of the right to understand how information is generated. For these reasons, this paper examines how varying degrees of AI disclosure (from light assistance to full generation) affect users' trust and behavioral intentions on digital platforms.

1.2. Literature review

Many studies help explain this problem area. Among them, Dietvorst, Simmons, and Massey put forward the concept of algorithm aversion. They found that people's trust in an algorithm decreases faster than their trust in a human after seeing either one make the same mistake, even when the algorithm's overall performance is better than the human's [3]. Longoni, Bonezzi, and Morewedge found that this resistance is stronger in subjective and experience-based areas because users believe that artificial intelligence lacks the emotional understanding required in these domains [4]. Laupichler, Knoth, Schleiss, and Raupach tested how different labels ("AI-generated," "human-written," or "human-AI mixing") affect the scorer's judgment of the same text. They found that content labeled as AI-generated received lower credibility and practicality scores, although "all texts are generally highly positive" [5]. Glikson and Woolley reviewed numerous empirical studies on trust in artificial intelligence and found that users' understanding of the working principles of artificial intelligence systems (algorithm literacy) shapes how information disclosure affects trust [6]. Do, Pham, Liao, Huang, and Wu tested disclosure in AI-generated ads and found that transparency disclosure labels positively affected consumer trust and the perceived credibility of the ads [7].

Most of the above studies regard algorithm aversion or transparency as a broad concept. Many studies only focus on a single scenario, such as advertising or medical services. In contrast, few studies use fine-grained behavioral data, such as eye-tracking results, that cover the entire range from AI assistance to complete AI generation, and few studies compare how this range affects trust differently across platforms or product types. This makes the platform team lack clear evidence to build tiered disclosure rules suitable for different types of content. Most existing studies also treat disclosure as a fixed label rather than a design option that can be adjusted in wording and detail. Therefore, the platform team rarely knows how small changes in disclosure design might change users' responses. This article aims to fill that gap. Rather than focusing on a single platform or product type, it compares the disclosure effect across multiple dimensions simultaneously.

DOI: 10.54254/2753-7064/2026.BJ36227

1.3. Analytical framework: a multi-method approach

To close this gap, this article builds a four-step framework. First, this paper uses data from experimental and eye-tracking research to map the level of AI disclosure using user behavior indicators. Second, it discusses the mechanisms behind the decline in trust, namely attribution bias, loss of perceptual control, and the transparency threshold effect. Third, it examines the differential performance of these mechanisms across distinctive platforms, product types, and user groups. Finally, based on these findings, the paper proposes three governance tools: a tiered disclosure system, a traceable record-keeping mechanism, and a human-in-the-loop approach to platform governance.

2. Empirical evidence: mapping disclosure to behavioral outcomes

To distinguish emotional responses from measurable results, this section reviews controlled studies that track how disclosure of artificial intelligence information affects the behavior of specific users.

2.1. Utility vs. experience: disclosure effects in transactional domains

In advertising and utilitarian settings, disclosure usually helps to improve trust. A study of 300 active digital users found that a transparency-disclosure label positively impacts trust and users' perceived credibility of artificial intelligence-generated advertisements [7]. However, an experiment using a Chinese consumer panel found that in hedonic e-commerce, the situation is exactly the opposite. Disclosing AI curation can improve the purchase intention for hedonic products, with "the AI-versus-human lift confined to the hedonic cell," while this effect in utilitarian e-commerce was smaller and less consistent. A specific case in this series of studies shows these two product categories side by side. Study 1 recruited 228 members of a Chinese consumer panel and showed each participant a recommended product carrying either an "AI-curated for you" badge or a "Curated by our editors" badge. For a wireless mouse, the average purchase intention under the two badges was almost the same ($M = 5.03$ for each). For a music subscription playlist, the average purchase intention under the AI-curated badge ($M = 5.47$) was higher than under the human-curated badge ($M = 4.83$) [8].

2.2. Authenticity dilemma: disclosure impacts on social and hedonic engagement

In emotional and short-video settings, the pattern becomes more mixed. A study on marketing content generated by artificial intelligence found that AI disclosure strengthens engagement for functional content through a cognitive pathway, but weakens engagement for hedonic content through an emotional pathway. The study describes this as a paradoxical moderating effect of disclosure across content types [9]. A field experiment with 60 high school students found that a simple AI disclosure label increased attention and trust toward an AI chat agent, but for news content, the same simple label "reduced trust and sharing," while a more detailed disclosure "lowered engagement overall" [10]. A concrete case from the same field experiment illustrates the difference between the two content types for the same students. When the students interacted with an AI chat agent, a simple disclosure label improved their trust in the agent, but when they read the news and comments generated by artificial intelligence, the same label reduced their trust and the frequency of content sharing. A more detailed label even further reduced their engagement [10].

DOI: 10.54254/2753-7064/2026.BJ36227

2.3. Visual attention metrics: insights from eye-tracking studies

The eye tracking data adds a cognitive layer to these discoveries. In the same field experiment, disclosure changed the duration users spent watching content and the frequency of their returns to disclosure labels, indicating that users' reading patterns have shifted from open reading to a more cautious, verification-dependent style [10]. Another eye tracking study of images and videos generated by artificial intelligence on social media (with a total of 64 participants) also found that information disclosure can change users' viewing behavior. The author concluded that their findings "highlight the subtle role of artificial intelligence-generated content (AIGC) information disclosure in shaping user trust" [11]. Table 1 summarizes these patterns across five representative studies, showing that the direction of the disclosure effect largely depends on the platform and product types.

Table 1. AI disclosure effects by platform type and content type

Platform Type	Functional / Utilitarian Content	Emotional / Hedonic Content
Transactional (e-commerce, advertising)	Positive: disclosure raises trust and perceived ad credibility [7]	Positive: disclosure raises purchase intent, and the lift is larger for hedonic goods than for utilitarian ones [8]
Social (chat agents, news, short-video, images)	Positive: disclosure raises trust and engagement for chatbot interaction [9, 10]	Negative: disclosure lowers trust, sharing, and engagement for entertainment, news, and image or video content [9-11]

These models show that AI information disclosure is not a single factor. The result depends on whether the platform is transactional or social, and on whether the disclosed content is functional or emotional.

3. Mechanism analysis and moderating heterogeneity

This section explores why AI disclosure can reduce trust and under what conditions this impact will be stronger or weaker.

3.1. Unpacking the trust calculus: attribution bias and perceived control

3.1.1. Attribution bias

When users find errors in the content, they judge the responsible party. This process is called attribution. Dietvorst, Simmons, and Massey had a counterintuitive finding. Once people see that the algorithm makes a mistake, their trust in it will be lower than their trust in humans who make the same mistake, even if the algorithm's average performance is still better than that of humans [3]. This is because people expect algorithms to be almost perfect, so a single error can undermine that expectation.

Laupichler and colleagues have also found a similar pattern. Texts marked as "AI generation" have lower credibility and effectiveness scores than similar texts marked as human-written or human AI mixed, although, as they marked, "all texts received highly positive evaluations" [5]. This shows that the label itself, not just the content quality, will affect the tolerance or harshness of user judgment.

DOI: 10.54254/2753-7064/2026.BJ36227

3.1.2. Perceived control

The second mechanism is related to the user's sense of control over the AI system. Zerilli, Bhatt, and Weller reviewed the research by the human-AI collaboration team and noted that "transparency may be crucial for facilitating appropriate levels of trust in AI and thus for counteracting aversive behaviors," but they also added an important prerequisite. Transparency can only work when it provides users with choices, such as allowing users to adjust or override the system's output [12]. Without these options, information disclosure may eventually feel like a warning rather than providing options, which in turn causes discomfort.

Yu and Li tested this idea on 235 working adults. They found that AI disclosure has two simultaneous effects: on the one hand, it improves users' perception of the system's effectiveness, thereby enhancing trust; on the other hand, it increases users' discomfort, thereby weakening trust [13]. Therefore, when users are unable to cope with this discomfort, for example, by turning off AI content or requiring manual substitution, negative effects tend to dominate.

3.1.3. Threshold and non-linear effects

The third mechanism is that the impact of disclosure and transparency on trust is not direct. Yu and Li pointed out that early studies reported a "positive correlation, non-correlation, or an inverted U-shaped relationship" between AI transparency and human trust in AI [13]. Their research helps to explain the reason. While transparency improves effectiveness, it also increases user discomfort. Therefore, the net impact on trust may be positive, neutral, or negative, depending on which path has a greater impact on a specific user and setting. In other words, the platform cannot take it for granted that the more information is disclosed, the better or worse it is.

A related pattern showed up in a field study of student readers. Compared with simple labels, detailed information disclosure "lowered engagement overall" [10]. This shows that, in every case, there seems to be a threshold for how much detail a disclosure includes, and that threshold isn't the same across contexts. Once that threshold gets crossed, adding more detail no longer helps build trust. Instead, it starts to work against user engagement. This effect also explains why certain studies reach opposite conclusions, even when their disclosure designs appear similar on the surface. Even small differences in the amount of detail a label carries can push the same underlying mechanism toward either a net positive or a net negative outcome.

3.2. The non-linearity of transparency: threshold effects

3.2.1. Platform type: E-commerce and advertising vs. social media

The effect of information disclosure can vary quite a bit depending on the type of platform involved. Take advertising, for example: Do and colleagues found that a disclosure label boosted both trust and perceived credibility [7]. Since users generally want information that is accurate and easy to verify, simply knowing that AI was involved ends up strengthening their trust in the platform's efficiency. Things get a bit more complicated once moving to short-form video and social platforms. Chen and colleagues found that a simple AI disclosure label improved users' trust in an AI chat agent, but for news content, that same label "reduced trust and sharing" [10]. What this suggests is that social and entertainment content leans much more heavily on a sense of authentic connection, making it more sensitive to AI disclosure than advertisements are. One possible reason comes down to motivation. Advertising and product-search interactions tend to be goal-driven. As a result, users

DOI: 10.54254/2753-7064/2026.BJ36227

primarily seek accurate information that helps them complete the task. Social and short-form video interactions work differently, though, since they are usually driven by entertainment and connection. Given that difference, an AI label can end up disrupting the sense of spontaneous, human-made content that users expect in this kind of setting.

3.2.2. Product type: functional vs. hedonic

Product type also affects how disclosure works. Ma and colleagues, working with a Chinese consumer panel, found that in hedonic e-commerce, where consumers are more inclined to form personal preferences than to meet fixed needs, disclosing AI curation increased purchase intention, whereas in utilitarian e-commerce, the effect of information disclosure was smaller and less consistent [8]. Through two controlled experiments on AI-generated marketing content, Gao, Li, and Zhao found a more detailed pattern: information disclosure strengthened the cognitive pathway that supports engagement with functional content, while it weakened the emotional pathway that supports engagement with hedonic content [9]. In summary, these findings show that disclosure works best for products where users value clear and verifiable information, and underperforms for products where users value emotional connection.

3.2.3. User segment: age, AI literacy and engagement

User background also plays a great role. Xiao and colleagues studied 720 short-video users aged 18 to 24 and found that AI disclosure had a "positive direct effect on prolonged watching intention," which can stimulate users' curiosity. But at the same time, it also increased perceived risk and reduced content trust. Moreover, the higher a user's AI literacy, the stronger both pathways became. This shows that even in the same age group, users with a better understanding of artificial intelligence systems react more strongly to information disclosure, whether this reaction is positive or negative. Since most current studies of this kind focus on young users or working-age users, scholars have little evidence on how elderly users with low AI literacy respond to the same information disclosure options. This is a gap that needs to be filled in future platform-level research. This finding also shows that platforms cannot regard all young users as a unified group. Even users of the same age may react in opposite ways to the same information disclosure design, depending on their understanding of how the underlying AI system works.

4. Governance pathways: towards a context-aware framework

Based on the problems identified in Section 3, this section proposes three governance strategies for platforms that use AI to assist in content creation.

4.1. Tiered disclosure protocols: aligning transparency with content sensitivity

Section 3.1.3 shows that disclosure details are not "the more the better". If the disclosure is inconsistent with the context, whether it is light or heavy, it may reduce user engagement [13]. Based on this finding and the finding that AI literacy can strengthen users' positive and negative reactions to the disclosed content, platforms should design a graded disclosure system rather than a single fixed label [14]. Mild AI assistance, such as grammar correction, does not require eye-catching labels because such modifications do not affect the substance of the content. On the other hand, heavy AI participation, including AI-generated drafts or images and videos, should use a clear disclosure method, consistent with the finding that disclosure improves user trust in advertising

DOI: 10.54254/2753-7064/2026.BJ36227

scenarios [7]. For hedonic content, information disclosure may reduce user engagement [9]. The platform can test short labels instead of long ones. This is based on a field study that shows that in a similar environment, detailed information disclosure reduces engagement more than simple tags [10]. Based on this, the tiered approach is reasonable. It adjusts the intensity of information disclosure based on the specific situation to reduce the risk of over-labeling low-risk content and meet the transparency requirements for high-risk content.

Platforms don't have to roll out this tiered approach all at once, either; it can be introduced step by step. A sensible place to begin is with the highest-risk content categories, such as health or financial information, since these carry the most at stake. From there, platforms can gradually extend the tiered label out to lower-risk categories, adjusting based on user feedback at every stage.

4.2. Audit trails and compliance: embedding traceability mechanisms

Under the Cyberspace Administration of China's Measures for Labeling of AI-Generated Synthetic Content, providers must add a visible label to AI-generated content and embed traceability information, like the service provider's name and a content ID, into the content's file metadata [15]. With the policy in place, platforms should maintain a clear record of exactly which AI tools generated or edited a given piece of content, when that occurred, and how much editing was done. The record-keeping serves two purposes. On the one hand, it helps platforms remain compliant with existing regulations. On the other hand, it helps protect the platform's reputation, since if a dispute comes up, for instance, users claiming that AI-generated content is misleading, the platform can quickly trace back through the content's history instead of leaning solely on public trust to settle the matter. This method also ties back to the attribution bias issue covered in Section 3.1.1. Because records are clear, platforms can correct specific errors quickly, rather than letting users take one obvious mistake and generalize it into distrust toward all AI content on the platform [3].

4.3. Preserving human agency: the human-in-the-loop imperative

Section 3.1.2 shows that perceptual control will change the way information disclosure affects trust. When users lack choices, transparency becomes a source of discomfort rather than trust [12]. Therefore, the platform should reserve the option of manual participation in scenarios where users most need authentic interpersonal interaction, such as complaint handling, health-related issues, or emotionally social content, echoing the finding that hedonic content is more sensitive to AI disclosure than functional content [8]. In practice, this means allowing users to request a manual review or reply in these types of interactions, and using the user's negative reaction as a signal to route similar future cases to human staff. The design of this human-in-the-loop approach does not remove AI from the platform. On the contrary, it applies AI to the most efficiency-driven scenarios, such as advertising and product information, while retaining manual participation in scenarios where relationship-based trust is more important [7]. This is in line with the broader view that trust in AI depends on matching technology with the right task, rather than applying it to all scenarios with the same intensity [6]. Over time, data from these human-in-the-loop cases can also help improve the tiered disclosure system in Section 4.1, establishing a feedback loop between governance design and user behavior.

DOI: 10.54254/2753-7064/2026.BJ36227

5. Conclusion

5.1. Synopsis of findings

This paper discusses how AI disclosure, ranging from light assistance to full generation, affects users' trust and behavioral intentions on digital platforms. Research shows that AI disclosure is a double-edged sword. In utilitarian and advertising scenarios, because users value clear information, disclosure can often enhance trust. However, in social and hedonistic scenarios, disclosure tends to weaken trust and engagement, because users in these settings value real interpersonal interaction more. Three mechanisms help to explain this phenomenon. First, users judge AI-labeled mistakes more harshly than they judge the same mistakes made by humans. Second, users experience a mix of greater perceived effectiveness and greater discomfort simultaneously. Third, the degree of detail in disclosure will have different impacts on trust depending on the specific situation. In addition, these effects will also differ by platform type, product type, and the user's understanding of the artificial intelligence system. Based on these findings, this paper proposes three interrelated governance strategies: a tiered disclosure system that matches content sensitivity, traceable internal records that comply with regulations, and a human-in-the-loop design for highly sensitive interactions.

5.2. Theoretical and practical implications

This research has both commercial and social value. For platform operators, research shows that a single fixed disclosure rule is unlikely to apply to all content types. On the contrary, adjusting the intensity of disclosure based on the specific situation helps balance transparency and user participation. For the broader digital environment, these findings also support the current regulatory direction on artificial intelligence-generated content labels and traceability requirements and provide a basis for designing such labels without causing unnecessary loss of user trust.

5.3. Limitations and future directions

However, this study also has limitations. It mainly relies on published research, rather than the new first-hand data. Therefore, it cannot fully test how these models work together on a single platform. In addition, some cited studies focus on young or working-age users, leaving the issue of elderly users' responses to the same disclosure design pending. Future research can address these shortcomings in the following ways: by collecting first-hand survey or interview data across a wider range of ages, and by testing the tiered disclosure design directly on the actual platform to measure real behavioral outcomes.

References

[1] China Internet Network Information Center. (2024, May 9). The 53rd statistical report on China's internet development. Retrieved July 23, 2026, from https://www.cnnic.com.cn/IDR/ReportDownloads/202405/P020240509518443205347.pdf

[2] State Council Information Office of the People's Republic of China. (2025, April 9). 346 generative AI services filed with Cyberspace Administration of China. Retrieved July 23, 2026, from http://english.scio.gov.cn/pressroom/2025-04/09/content_117814020.html

[3] 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. https://doi.org/10.1037/xge0000033

DOI: 10.54254/2753-7064/2026.BJ36227

[4] Longoni, C., Bonezzi, A., & Morewedge, C. K. (2019). Resistance to medical artificial intelligence. *Journal of Consumer Research, 46*(4), 629–650. https://doi.org/10.1093/jcr/ucz013

[5] Laupichler, M. C., Knoth, N., Schleiss, J., & Raupach, T. (2025). Algorithm aversion revisited: The role of AI literacy and attitudes towards AI in shaping perceptions of AI-generated texts. *British Journal of Educational Technology*. Advance online publication. https://doi.org/10.1111/bjet.70035

[6] Glikson, E., & Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals, 14*(2), 627–660. https://doi.org/10.5465/annals.2018.0057

[7] Do, T., Pham, T.-Q.-A., Liao, Y.-K., Huang, K.-C., & Wu, W.-Y. (2026). Transparency disclosure mechanisms for AI-generated advertising on online platforms: An empirical study of user trust and engagement. *Internet Technology Letters*. Advance online publication. https://doi.org/10.1002/itl2.70354

[8] Ma, T., Du, Y., Wang, Y., & Zhou, L. (2026). Strategic disclosure of AI curation: A boundary condition on algorithm aversion in hedonic e-commerce. *Journal of Theoretical and Applied Electronic Commerce Research, 21*(7), Article 232. https://doi.org/10.3390/jtaer21070232

[9] Gao, X., Li, W., & Zhao, Y. (2026). AI-generated marketing content value and disclosure: How functional and hedonic appeals shape customer engagement. *Frontiers in Psychology, 16*, Article 1701085. https://doi.org/10.3389/fpsyg.2025.1701085

[10] Chen, N., Lai, Z., Liu, Y., Li, J., Wang, R., & Yan, P. (2026). How AI disclosure shapes high school students' trust. *Information Research, 31*(iConf). https://doi.org/10.47989/ir31iConf64165

[11] Wang, R., Jia, B., & Yan, P. (2025). "Saying is believing": Exploring the importance of AI-generated content disclosure and user trust. *Proceedings of the Association for Information Science and Technology, 62*(1). https://doi.org/10.1002/pra2.1295

[12] Weller, A., Zerilli, J., & Bhatt, U. (2022). How transparency modulates trust in artificial intelligence. *Patterns, 3*(2), Article 100455. https://doi.org/10.1016/j.patter.2022.100455

[13] Yu, L., & Li, Y. (2022). Artificial intelligence decision-making transparency and employees' trust: The parallel multiple mediating effect of effectiveness and discomfort. *Behavioral Sciences, 12*(5), Article 127. https://doi.org/10.3390/bs12050127

[14] Xiao, Y., Du, J., Ding, Y., Zhang, M., Jiang, Y., Yang, Y., & Liu, J. (2026). AI-generated content disclosure and prolonged short-video engagement: A heuristic-systematic risk-trust model among late-adolescent and emerging-adult TikTok users. *Behavioral Sciences, 16*(7), Article 1179. https://doi.org/10.3390/bs16071179

[15] Cyberspace Administration of China, Ministry of Industry and Information Technology, Ministry of Public Security, & State Administration of Radio and Television. (2025, March 14). Measures for labeling of AI-generated synthetic content. Retrieved July 27, 2026, from https://www.cac.gov.cn/2025-03/14/c_1743654684782215.html