Information Research - Vol. 31 No. iConf (2026)

See, trust, and interact: how AI disclosure shapes high school students' trust

Nuo Chen, Zhiyuan Lai, Yichu Liu, Jia Li, Rui Wang, and Pu Yan

DOI: *https://doi.org/10.47989/ir31iConf64165*

Abstract

Introduction. The rise of AI-generated content challenges adolescents' ability to evaluate information and calibrate trust. This study explores how AI disclosure influences high school students' attention, trust, and interaction with AI-generated news and comments.

Method. A field experiment was conducted at a county-level high school in Henan with 60 students. Participants were randomly assigned to a control group (no disclosure) or one of two experimental groups (simple vs. detailed disclosure), enabling examination of group-level effects. Data collection combined eye-tracking, post-test questionnaires, and interviews.

Analysis. Eye-tracking metrics and survey data were analysed quantitatively to examine the main and moderating effects of AI disclosure, while interview transcripts were thematically coded to provide qualitative insights.

Results. Simple disclosure increased attention and trust in AI bots but reduced trust and sharing for news content. Detailed disclosure lowered engagement overall, slightly reducing trust in conversational settings and strongly reducing news-sharing. Individual differences moderated these effects: light internet users benefited most from simple labels, whereas heavy users showed stronger gains from detailed explanations in AI trust and technical understanding.

Conclusion. AI disclosure produces context-dependent effects. Effective design should align label complexity with content type and user experience to provide guidance for ethical AI integration in education and social media.

Introduction

The rapid development of artificial intelligence has permeated nearly every aspect of our lives, profoundly shaping the digital landscape. The rise of AI-Generated Content (AIGC) is one of the most transformative advancements, with AI-generated images, videos, comments, and news reports now filling our online world. Therefore, it has become crucial to accurately understand and measure user trust in AIGC, and the various factors that shape it.

While the academic community has increasingly focused on human-AI trust and interaction, most studies have been conducted with users from Western, urban or university populations, leaving a significant gap in our understanding of how trust is formed and impacted in other cultural and social contexts. This cultural divergence is underscored by the 2025 Edelman Trust Barometer, which found that 87% of the Chinese public trusts AI compared to just 32% in the U.S.. Our study aims to address this gap by focusing on a previously under-examined but crucial demographic: high school students in Chinese county towns. As China's most populous inland province with nearly 100 million inhabitants, Henan provides a representative case; it maintains a high total GDP while its per-capita disposable income (~42,027 RMB) remains in the lower-middle tier, offering a micro-model of the vast digital ecosystems within Chinese county-level regions. Specifically, these students operate within high-stakes academic environments where AI is gradually adopted as a pragmatic efficiency tool for learning, yet they often lack systematic AI literacy—which fundamentally shapes their perceptions of and interactions with AI-generated content.

For high school students, AI-generated content (AIGC) has become a double-edged sword: a powerful tool for creativity and learning, but also a source of significant risk. Unlabelled and unchecked AI-generated deepfakes pose a unique threat to students, as the spread of such content is known to compromise a student's ability to discern truth, impacting their critical thinking skills and even their mental well-being (Hancock & Bailenson, 2021). Furthermore, despite a global push for AI transparency, findings on the impact of AI disclosure on user trust are mixed, as some studies suggest a positive relationship (Lee & Cha, 2024) while others report no clear impact (Rossner et al., 2024). However, little is known about how AI disclosure strategies impact users' cognitive processing of trust and their intention to interact. This is especially important in China, which introduced its own AIGC labeling regulations in March, 2025, with implementation taking effect in September.

Furthermore, despite a global push for AI transparency, findings on the impact of AI disclosure on user trust are mixed, as some studies suggest a positive relationship (Lee & Cha, 2024) while others report no clear impact (Rossner et al., 2024). However, little is known about how AI disclosure strategies impact users' cognitive processing of trust and their intention to interact. This is especially important in China, which introduced its own AIGC labeling regulations in March, 2025, with implementation taking effect in September.

To address these significant gaps, we conducted a controlled field experiment in a county-level high school in Henan, China in June, 2025. We employed a multi-modal approach to investigate how AI disclosure strategies influence high school students' trust in and interactions with AIGC. Our methodology combined objective behavioral data from eye-tracking with subjective self-reports from questionnaires, providing a comprehensive view of how students cognitively process trust in real-time.

Based on our review of the literature, we propose the following research questions and hypotheses:

RQ1: How do different AI disclosure strategies affect high school students' visual attention, explicit trust in AI-generated content, and their intention to interact with it?

> H1a: A simple AI disclosure strategy will positively influence high school students' visual attention to AI-generated content, their trust in it, and their intention to interact with it.

H1b: A technical and cautionary AI disclosure strategy will negatively influence high school students' visual attention to AI-generated content, their trust in it, and their intention to interact with it.

H1c: AI disclosure strategies will enhance high school students' AI literacy.

RQ2: How do differences in AIGC format (e.g., bots comments vs. images) and content type influence high school students' trust and interaction intentions?

H2a: The effect of AI disclosure strategies on user trust and interaction intention will differ across AIGC formats (e.g., AI bot comments vs. AI-generated images).

H2b: Within the same AIGC format, the perceived topic and truthfulness of AI-generated content will influence high school students' trust and interaction intention.

RQ3: Do individual differences among high school students moderate their trust in and interaction intention with AIGC?

H3a: Students' individual demographic factors (e.g., age and gender) and academic performance will moderate their trust in and interaction intention with AIGC.

H3b: Students' prior digital experience and skills (e.g., internet usage experience and digital literacy) will moderate their trust in and interaction intention with AIGC.

The findings of this research offer both theoretical and practical contributions. Theoretically, our study will provide empirical evidence for human-AI interaction models within a previously unexplored cultural and demographic context. Practically, the results will offer actionable insights for technology developers, educators, and policymakers to create ethical and trustworthy AI systems that promote positive learning outcomes and protect the well-being of young users.

Literature review

Understanding trust in human-AI interaction

Trust is a foundational element in human-technology interaction, widely defined as a user's willingness to be vulnerable to a system's actions based on positive expectations (Mayer et al., 1995). This definition highlights core elements of vulnerability, risk, and reliance, which are central to how users engage with and rely on AI-driven systems (Ueno et al., 2022).

Recent research has explored the evolving, bidirectional nature of this relationship and the various factors that influence it (Jarrahi, 2018; Liao & MacDonald, 2021). Trust is a dynamic process, evolving from initial judgments to sustainable confidence built through repeated interactions (Gao & Waechter, 2017). This trajectory is influenced by a range of factors, including individual differences, situational contexts, information transparency and system-related elements such as perceived competence and predictability (Hoff & Bashir, 2015; Naiseh et al., 2021; Oleson et al., 2011). Specially, Lewis and Marsh (2022) built on the three-factor model of human-machine trust (Schaefer et al., 2016) to propose an integrated model for human-AI trust, suggesting that trust is determined by four key factors: the AI's capability, the predictability of its behavior, its perceived honesty and integrity, and its benevolence or willingness to meet human needs.

HCI studies further highlight the significance of interactional cues, ranging from explanations and feedback (Ehsan et al., 2019) to anthropomorphic design elements (de Visser et al., 2018), in shaping users' trust judgments. The ideal state for human-AI collaboration is trust calibration, which ensures a user's trust level is appropriately aligned with the AI's actual capabilities (Robinette et al., 2016). Both under-trust (failing to use a capable AI) and over-trust (relying on a flawed AI) can lead to negative outcomes in interaction (Lee & See, 2004). Ultimately, trust level has a significant impact on interaction behavior. Low trust can lead to reduced usage, while increasing trust can significantly boost user satisfaction with AI interaction and raise users' expectation threshold for the AI (Shin, 2021).

As AI technologies and AI-generated content become increasingly integrated into educational settings, researchers have also begun to analyze students' trust in AI and the factors that influence it. Students generally exhibit a moderate or neutral level of trust in AI, encompassing both GenAI and digital assistants (Samonte et al., 2023; Amoozadeh et al., 2024; Ramirez et al., 2024; Kozak & Fel, 2024). Sociodemographically, trust varies by grade, gender, nationality, and religion (Kozak & Fel, 2024; Amoozadeh et al., 2024). Male students typically display higher trust than females, and students in higher years of study and those with higher frequencies of religious practices tend to show greater trust (Kozak & Fel, 2024). Regarding perception and interaction, students generally trust AI's functionality and usefulness (Ramirez et al., 2024; Kozak & Fel, 2024), yet concerns about biases, errors, and the necessity for human supervision contribute to distrust (Bochniarz et al., 2022; Amoozadeh et al., 2024; Samonte et al., 2023). Many students also mistakenly perceive GenAI as transparent, indicating a need for calibrated trust (Amoozadeh et al., 2024). Furthermore, confidence in one's AI literacy positively correlates with trust and a more positive attitude towards AI (Ramirez et al., 2024; Kozak & Fel, 2024).

While research has explored this topic generally, it has largely overlooked marginalized and underrepresented populations such as rural adolescents—a group uniquely vulnerable due to high social media use and still-developing cognitive schemas. This study addresses these key gaps by focusing on this specific population and investigating how specific disclosure strategies affect the formation of trust in the context of AIGC.

AI transparency and disclosure strategies

Transparency is a cornerstone of trustworthy human-AI interaction, defined broadly as the clear communication of information in regulations like the European Union’s General Data Protection Regulation (GDPR) (European Commission, 2018).Within the AI context, this concept has evolved from a focus on an AI’s ability to explain its actions (Kim & Hinds, 2006) to a broader principle encompassing the communication of an AI’s goals and status (Chen & Barnes, 2014). More specifically, algorithmic transparency has been defined as the degree to which users can understand an AI's predictions or decisions (Shin & Park, 2019). This study adopts a disclosure-centric view of AI transparency, focusing on the explicit communication of information to users. This includes revealing technical details, rights descriptions, and privacy policies (Chan, 2023), with a particular emphasis on the disclosure of AI contribution and warning labels, which are most common in social media contexts (Lund et al., 2023).

While early research suggested a simple, positive relationship between transparency and user trust (Ribeiro et al., 2016; Linegang et al., 2006; Sinha & Swearingen, 2002), more recent findings reveal a complex and often contradictory picture. Excessive transparency can lead to information overload and decreased trust (Kizilcec, 2016), and in some cases, can even reduce a user's willingness to engage with the system (Poursabzi-Sangdeh et al., 2021; Schmidt et al., 2020). These complexities are particularly evident in the AIGC domain. While some studies show that detailed disclosure can increase user trust (Lee & Cha, 2024; Sunnie, 2024), others demonstrate a more negative response. For example, disclosing an AI's identity may reduce a user's likelihood of making a purchase (Luo et al., 2019) or have no effect on trust in news content (Rossner et al., 2024). These conflicting results suggest that the impact of AI disclosure is not universal and is likely dependent on the content category and topic of the AIGC material.

The effectiveness of transparency is highly dependent on its implementation, which is complicated by a significant gap between the ideals of algorithmic transparency and its practical application in user interfaces. This is compounded by the fact that users, particularly on social media, often struggle to identify or correctly interpret credibility indicators, leading to a state of opaque transparency (Chang et al., 2025). A critical factor in this complexity is the type of disclosure. A study on AI-generated social media images found that while labels reduced user trust, their specific wording mattered. A simple "AI-generated" label had a less negative impact on trust than a label implying the content was "manipulated" or "false" (Wittenberg et al., 2024). This shows that a disclosure about an AI's contribution is perceived differently than one that implies a negative quality.

While prior research has demonstrated that algorithmic transparency can meaningfully shift user perceptions, it remains unclear how specific disclosure strategies operate within the complex and fast-paced social media environment in AI era. Addressing this gap is crucial not only for refining theoretical models of how users trust digital information but also for informing the design of platforms that foster appropriate user trust and engagement with AI-generated content.

Challenges of measuring human-AI trust

Quantifying trust in human-AI interaction is a significant challenge in empirical research. The most common approach is self-report, relying on questionnaires and structured scales. These instruments, often adapted from established interpersonal trust scales like the one from McKnight et al. (2002), have evolved to capture the unique nuances of AI. More recently, researchers have developed new scales involving perceived usefulness (Choudhury & Shamszare, 2024), perceived risk (Liao & MacDonald, 2021), workload (Choudhury & Shamszare, 2023) and AI literacy (Hwang et al., 2023) specifically for generative AI. Among these, the Multi-Dimensional Measure of Trust (MDMT) is a widely cited scale that quantifies user trust across four key dimensions—reliability, competence, ethics, and integrity—and offers a comprehensive view of how users perceive an AI system (Ullman & Malle, 2019). Researchers also use qualitative methods like interviews to gain rich insights into how users form trust (Tossell et al., 2024). However, these self-report methods are known to interfere with user interaction and are prone to subjective bias, making it difficult to capture real-time changes in trust dynamics (Bindewald et al., 2018; Tossell et al., 2024).

To address the limitations of self-report, researchers have increasingly turned to observable behavioral measures as a proxy for trust. While the academic definitions of trust and trust behavior can be nuanced (Mayer et al.,1995; Thielmann & Hilbig, 2014), empirical studies often focus on two key indicators: compliance (the degree to which a user follows an AI's advice) and reliance (the extent to which a user depends on an AI system). For example, Compliance is often measured by observing whether users override automated commands in high-risk scenarios, such as taking over control of an autonomous vehicle (Bindewald et al., 2018; Ajenaghughrure et al., 2020), or by whether they correct errors after a task is completed. Reliance is quantified by the proportion of time a user allows an AI to operate autonomously (Gremillion et al., 2016) or the speed at which a user takes control from a system when they perceive a risk (Molnar et al., 2018). Specially, in the context of user-generated content (UGC) or artificial intelligence, scholars have used behaviors of content engagement like liking, sharing, or posting content to infer trust (Jiang et al., 2024).

Despite the utility of these methods, the existing trust measurement landscape presents a dual challenge of methodological innovation and real-world applicability. While both self-report and behavioral methods provide valuable insights (Balakrishnan & Dwivedi, 2021; Chattaraman et al., 2019), they are often confined to controlled lab settings and struggle to capture trust in complex, naturalistic environments (Akash et al., 2018). Furthermore, psychophysiological measures remain largely underutilized in the AIGC domain. This study addresses these gaps by employing a multi-modal data collection approach that integrates objective eye-tracking for real-time behavioral analysis with a post-task questionnaire to assess subjective trust. This methodology enables the creation of a more valid and reliable trust assessment system, aiming to bridge the divide between theoretical frameworks and practical applications by providing a more comprehensive understanding of human-AI trust dynamics.

Method

Research design

This research employs a mixed-methods design to comprehensively investigate the impact of AI information disclosure on human-AI trust and interaction. The workflow is demonstrated in Figure 1.

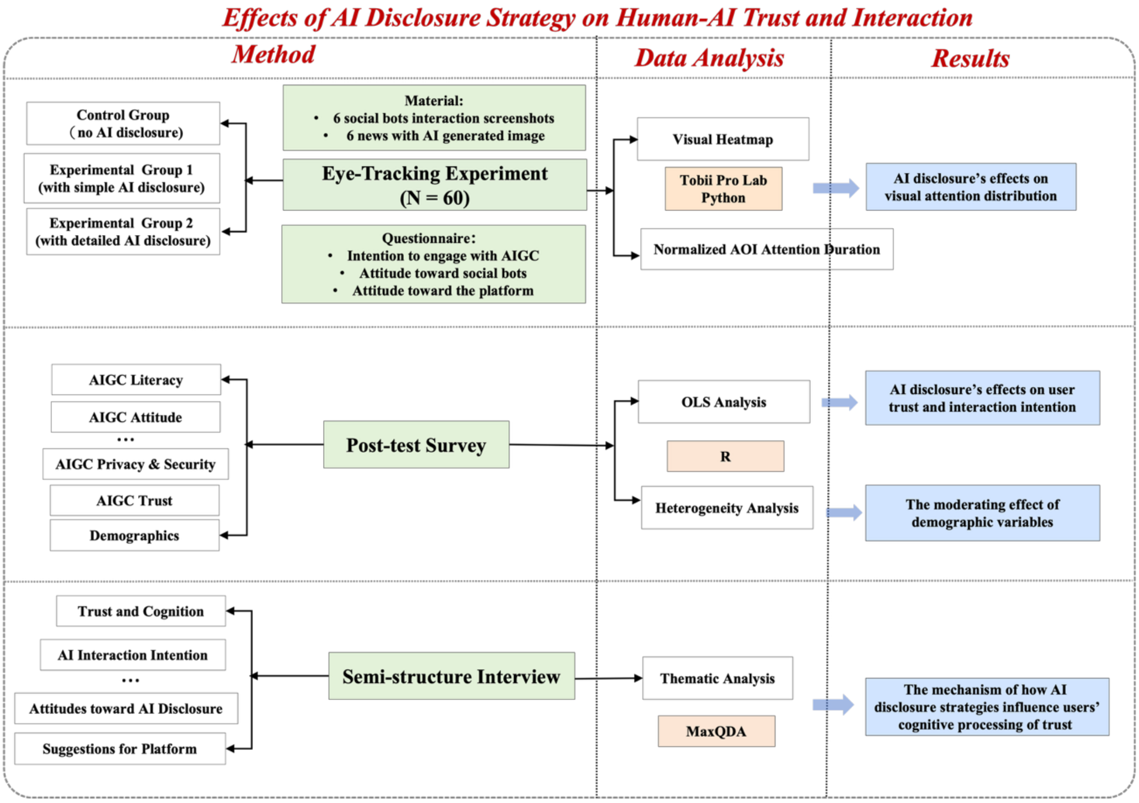

Effects of AI Disclosure Strategy on Human-AI Trust and Interaction
Method
Data Analysis
Results
Control Group (no AI disclosure)
Experimental Group 1 (with simple AI disclosure)
Experimental Group 2 (with detailed AI disclosure)
Material:
• 6 social bots interaction screenshots
• 6 news with AI generated image
Eye-Tracking Experiment (N = 60)
Questionnaire:
• Intention to engage with AIGC
• Attitude toward social bots
• Attitude toward the platform
Visual Heatmap
Tobii Pro Lab Python
Normalized AOI Attention Duration
AI disclosure's effects on visual attention distribution
AIGC Literacy
AIGC Attitude
…
AIGC Privacy & Security
AIGC Trust
Demographics
Post-test Survey
OLS Analysis
R
Heterogeneity Analysis
AI disclosure's effects on user trust and interaction intention
The moderating effect of demographic variables
Trust and Cognition
AI Interaction Intention
…
Attitudes toward AI Disclosure
Suggestions for Platform
Semi-structure Interview
Thematic Analysis
MaxQDA
The mechanism of how AI disclosure strategies influence users' cognitive processing of trust

Figure 1. Research design workflow

Participants

A total of 60 high school students from a county-level high school in Henan, China, took part in the study. To ensure a broad and representative sample, students were recruited using a stratified sampling method, with one student selected from each class across the first and second high school years. This approach ensured a diverse range of academic subjects and performance levels were reflected in the sample.

The sample consisted of 23 females and 37 males, with ages ranging from 15 to 18 (M = 16.85 years). As digital natives, participants reported an average of 8.17 years of internet usage experience. While their overall online experience was relatively rich, their weekly discretionary internet time varied significantly, ranging from less than 1 hour to over 10 hours.

Mixed-method eye-tracking experiment

Eye-tracking experimental design

To investigate the research questions, we adopted a mixed-design experimental approach. The between-subjects variable was the AI disclosure strategy, with 60 participants randomly assigned to one of three conditions: an unlabelled control group and two experimental groups with distinct AI labels.

The experiment also featured within-subjects manipulations of the material types and their content. All participants viewed a total of 12 AI-generated content (AIGC) materials, divided equally between two types. The first type included AI-driven social bot replies and the second type featured AI-generated images as news post visuals.

Within-subjects manipulation

Our experiment materials were designed across two primary formats with manipulations:

(1) AI-driven social bot replies: These materials included a social media user's original post and an AI bot's reply to it. Post topics were equally distributed between two types: personal emotional discussions and social event discussions. This allowed us to observe if the topic of the AI-generated content affected user trust and interaction intentions.

(2) AI-generated images: These images were integrated into social media news posts. For the AI-generated images, the corresponding news text was equally divided between real news and fake news to examine if the veracity of the information had an impact on how students engaged with the AI-generated visuals.

Between-subjects manipulation

The core of this experiment was the systematic manipulation of AI disclosure strategies. While disclosure is regarded as generally beneficial for user trust and interaction with technological systems (Linegang et al., 2006; Sinha & Swearingen, 2002; Wang & Benbasat, 2007), excessive or poorly designed disclosure may lead to negative effects (Kizilcec, 2016; Poursabzi-Sangdeh et al., 2021; Schmidt et al., 2020). Besides, recent research confirms that varying AI disclosure strategies on social media platforms influence users' trust in AI-generated content (Wittenberg et al., 2024).

To investigate this nuanced impact, participants were randomly assigned to one of three groups with distinct disclosure treatments. These strategies were designed to reflect real-world practices, based on a summary of AI disclosure in global and local social media platforms. The control group viewed all content without any AI disclosure, while the two experimental groups were exposed to different types of disclosures: one received a simple disclosure, and the other revealed more technical details.

For AI social bots materials, the color and font size of the AI label were kept constant, while its content and format was manipulated to simulate different points of disclosure:

(1) Control group: No AI labels.

(2) Experimental Group 1 (simple disclosure): A simple disclosure that identified the AI bot account as 'AI *social bots*,' with the label positioned directly after the account's username in the comments section.

(3) Experimental group 2 (detailed disclosure): This group built upon the simple disclosure from Experimental group 1 by adding a second, more detailed layer of disclosure. When participants hovered over the simple 'AI *social bots*' label, a text box would appear, revealing: '*Powered by a large language model to enable automated responses.*'

For the AI-generated image materials, while the color, font size, location, and format of the AI label were kept constant, the two types of disclosure differed in both their technical specificity and their advisory nature.:

(1) Control group: No AI labels.

(2) Experimental group 1 (simple disclosure): A simple AI label was positioned at the bottom of the news post, stating: 'Image is AI-generated.'

(3) Experimental group 2 (detailed disclosure): The label for this group was placed in the same location. However, it was more detailed and designed to be cautionary, reading: '*Image is generated by AI that learned from a large number of images. Please carefully discern.*'

Procedures

The eye-tracking experiment consisted of three main phases.

Eye-tracking task

During the eye-tracking task, participants viewed the materials on a computer monitor while their gaze was recorded using a Tobii Pro Spark eye tracker. The device captured gaze order, fixation duration, saccade count, and pupil diameter, enabling a moment-by-moment assessment of visual attention. Unlike self-report measures that rely on conscious responses, eye tracking provides objective, real-time indicators of cognitive processing and meaning construction (Rayner, 1998).

Participants viewed 12 AIGC materials presented in two sets of six. After viewing the first set of AI bot replies, they completed a questionnaire assessing attitudes toward the bot accounts and the platform. They then viewed a second set of AI-generated news posts and completed a follow-up questionnaire measuring trust in the content and interaction intentions.

Post-test survey

Following the eye-tracking task, participants filled out a post-test questionnaire measuring their AI literacy, AI skills, AI trust, AI attitudes, and sociodemographic information. The questionnaires utilized in this study were designed to measure a comprehensive set of variables (Table 1). They incorporated established scales and self-developed measures to assess key dimensions such as digital skills (van Deursen & van Dijk, 2010), AIGC literacy (Lintner, 2024; Lee & Park, 2024), and AI privacy and security (Kozyreva et al., 2021; Zhou & Lu, 2025). A central measure was AIGC trust, which was adapted from the literature to evaluate five dimensions: competence, anthropomorphism, integrity, transparency, and benevolence (De Freitas et al., 2023; Hu et al., 2021; Koo et al., 2015; Lee & See, 2004). Beyond, we included a dedicated survey item to analyze whether AI labels affect trust in human-original or human-AI co-created content (Wittenberg et al., 2024).

Construct	Variables	Reference
Digital Skills (TotalSkills)	Different network usage skills	van Deursen & van Dijk, 2010
Privacy Awareness (TotalPrivacy)	Sensitivity to the privacy of different types of personal information	OxIS-Questionnaire-2019 (Modified)
AI Access (AI_ACCESS)	Familiarity with and usage of different AI-generated content platforms and applications	OxIS-Questionnaire-2019 (Modified) and self-developed scale
AIGC Literacy (AIGC_LITERACY)	Technical Proficiency	Lintner, 2024; Lee & Park, 2024
	Communication Proficiency	
	Creative Application	
	AIGC usage experience	
AIGC Trust (AIGC_TRUST)	Competence	Hu et al., 2021
	Anthropomorphism	De Freitas et al., 2023
	Integrity	Hu et al., 2021
	Transparency	Lee & see, 2004; Koo et al., 2015
	Benevolence	Hu et al., 2021
AIGC Trust 2(AIGC_TRUST2)	Views on different types of AI-generated content	Wittenberg et al., 2024
AI Privacy and Security (TotalSecurity)	Personal privacy and security	Kozyreva et al., 2021
	Societal responsibility and security	Zhou & Lu, 2025
	Societal regulation	

Construct	Variables	Reference
DEMOGRAPHICS	Age, gender, education (grade, scores), household income, parents' occupation and educational background, internet usage experience, average weekly internet usage time, etc.	CFPS, 2024

Table 1. The structure of the post-test survey.

Semi-structured interview

The final phase involved a semi-structured interview, which focused on the user's overall experience during the experiment and their attitudes and trust toward AIGC. Participants were asked to elaborate on their prior exposure to and attitudes toward AI labels online. Additionally, we explored their expectations for platforms, with the goal of providing insights for both platform and policy design.

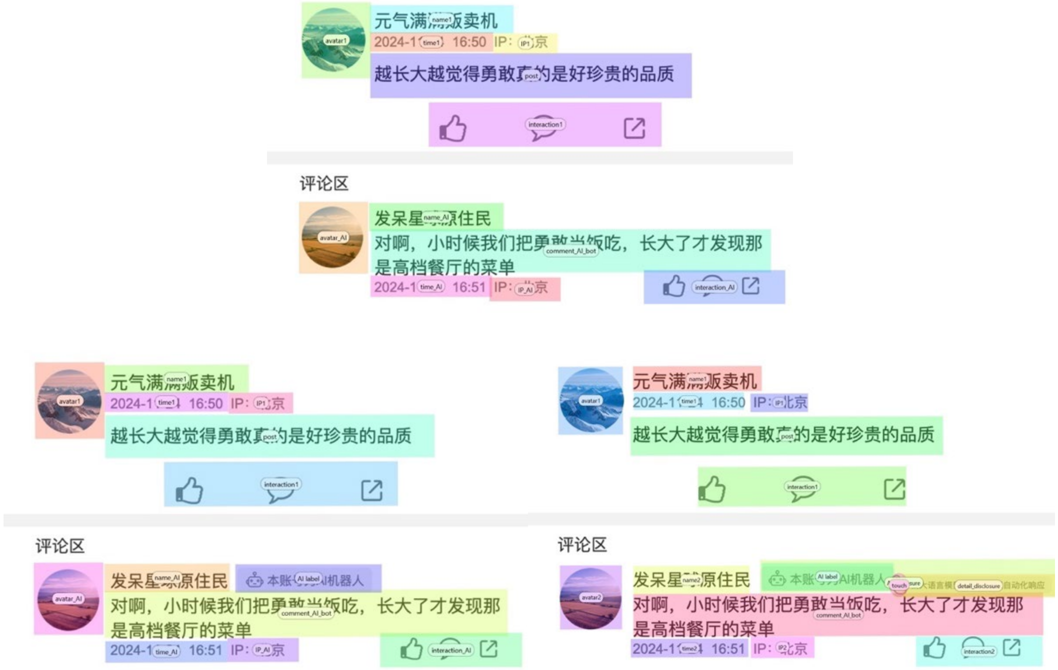
Data analysis

Quantitative analysis: eye-tracking measures

Before analysing the eye-tracking data, we performed a quality control to ensure its reliability. We excluded one participant whose eye-tracking error was outside the acceptable threshold. Additionally, during areas of interests (AOIs) processing, we removed three data points due to insufficient fixation durations or device measurement errors. This strict protocol resulted in a final data retention rate of 97.9%.

Our analysis focused on two levels of eye-tracking data to assess the impact of the AI disclosure intervention. First, we examined overall visual attention using the spatial distribution of eye-movement paths across regions of interest. Eye-tracking heat maps were also generated to visualize these patterns, which helped us understand which types of content and specific areas of the materials garnered the most attention.

Next, we conducted a more fine-grained AOI-based analysis. Using Tobii Pro Lab, we defined AOIs for key interface elements in both AI bot replies and AI-generated news posts, including the avatar, username, timestamp, text, and interaction buttons of the original posts and comments. Most critically, we defined a dedicated AOI for the AI disclosure label itself. Normalized fixation duration within this AOI served as a direct and objective measure of participants' attention to the disclosure. The defined AOIs are illustrated in Figures 2 and 3.



元气满满贩卖机
2024-1 16:50 IP: 京
越长大越觉得勇敢真的是好珍贵的品质

评论区

发呆星原住民
对啊，小时候我们把勇敢当饭吃，长大了才发现那是高档餐厅的菜单
2024-1 16:51 IP: 京

元气满满贩卖机
2024-1 16:50 IP: 京
越长大越觉得勇敢真的是好珍贵的品质

评论区

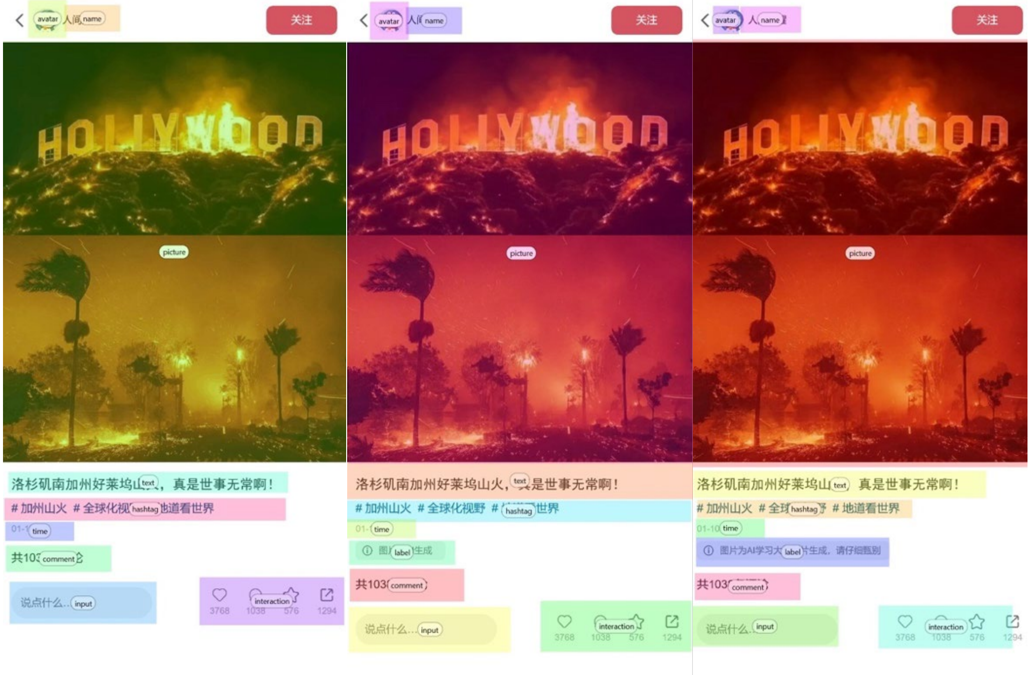
发呆星原住民 本账号为AI机器人
对啊，小时候我们把勇敢当饭吃，长大了才发现那是高档餐厅的菜单
2024-1 16:51 IP: 京

元气满满贩卖机
2024-1 16:50 IP: 北京
越长大越觉得勇敢真的是好珍贵的品质

评论区

发呆星原住民 本账号为AI机器人 大语言模 自动化响应
对啊，小时候我们把勇敢当饭吃，长大了才发现那是高档餐厅的菜单
2024-1 16:51 IP: 北京

Figure 2. AI label design and AOI divisions for the AI social bots material category, shown for the control group, experimental group 1, and experimental group 2. （The example material shows a user stating, ‘*As we grow older, we increasingly realize that courage is a truly precious quality.*’ The social bot replies, ‘*Exactly. When we were young, we treated courage as everyday sustenance; only when we grow up do we realize it belongs on a fine-dining menu.*’)



洛杉矶南加州好莱坞山火，真是世事无常啊！
加州山火 # 全球化视野 # 地道看世界
01-10
ⓘ 图片为AI学习大 生成，请仔细甄别
共103
说点什么...

Figure 3. AI label design and AOI divisions for the AI-generated news images category, shown for the control group, experimental group 1, and experimental group 2. (The example material shows a user post stating, ‘*The wildfire in the Hollywood Hills of Southern California—life is truly unpredictable. #CaliforniaWildfires #GlobalPerspective #SeeingTheWorldAuthentically*’)

Quantitative analysis: survey

Our data analysis began by ensuring the quality of our collected data. Of the 60 questionnaires, 54 were deemed valid, resulting in a high validity rate of 90%. The validity of the questionnaires ensured that each of the three groups had a sample size of at least 15 participants, a number sufficient to ensure adequate statistical power for our analysis. The data were then coded and cleaned using Excel and R for preliminary processing. We performed regression analyses with control variables to explore the effect of our intervention on the post-test variables (trust, interaction intention, and AI literacy). A moderation analysis is conducted to investigate any potential interaction effects as well. We also processed the short, in-experiment Likert-scale data to derive scores for trust and interaction with the content.

Qualitative analysis: thematic analysis of interview

For the qualitative analysis, we conducted a thematic analysis of the semi-structured interview transcripts. Thematic analysis systematically identifies and interprets patterns within qualitative data, offering contextual insight that complements quantitative findings (Braun & Clarke, 2006). Interview transcripts were first coded into meaning units, which were then analyzed to identify recurring themes and concepts. This process enabled us to synthesize participants' experiences and attitudes, providing a more nuanced understanding of their cognitive processes and expectations.

Results and analysis

Eye-tracking results

Overall visual attention distribution

To visually demonstrate how the different experimental groups allocated their attention to the materials, we generated eye-tracking heat maps that included the AOI divisions. We also calculated the average total attention duration for each material category. Figure 4 and Figure 5 illustrate the eye-tracking heat maps for each intervention.

The heat map analysis revealed that for the AI social bots materials, the presence of an AI label significantly increased participants' attention to the comments posted by the AI account. However, this effect was less pronounced in experimental group 2 compared to experimental group 1 (Figure 4). These findings suggest that users naturally dedicate more attention and cognitive resources to interactions with AI bots than human. They also indicate that the perceived identity of the poster, even for a seemingly ordinary comment, has a significant influence on user perception.

For the AI-generated news images, Experimental Group 1 dedicated more attention to the post's image and text compared to the control group. In stark contrast, this effect was absent in Experimental Group 2(Figure 5). This group's diminished attention to the AI label AOI—compared to the first experimental group—suggests a possible negative consequence of over-disclosure (Schmidt et al., 2020).

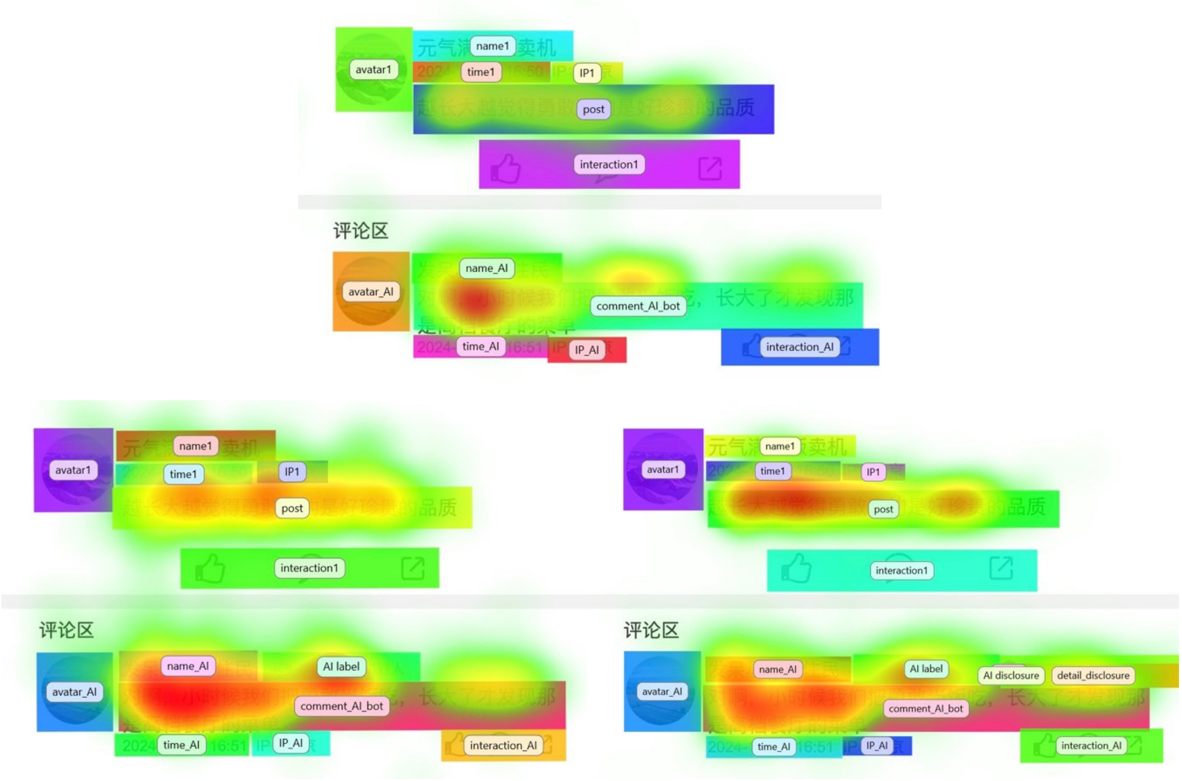


Figure 4. Eye-tracking heatmap of the AI social bots material category, shown for the control group, experimental group 1, and experimental group 2.

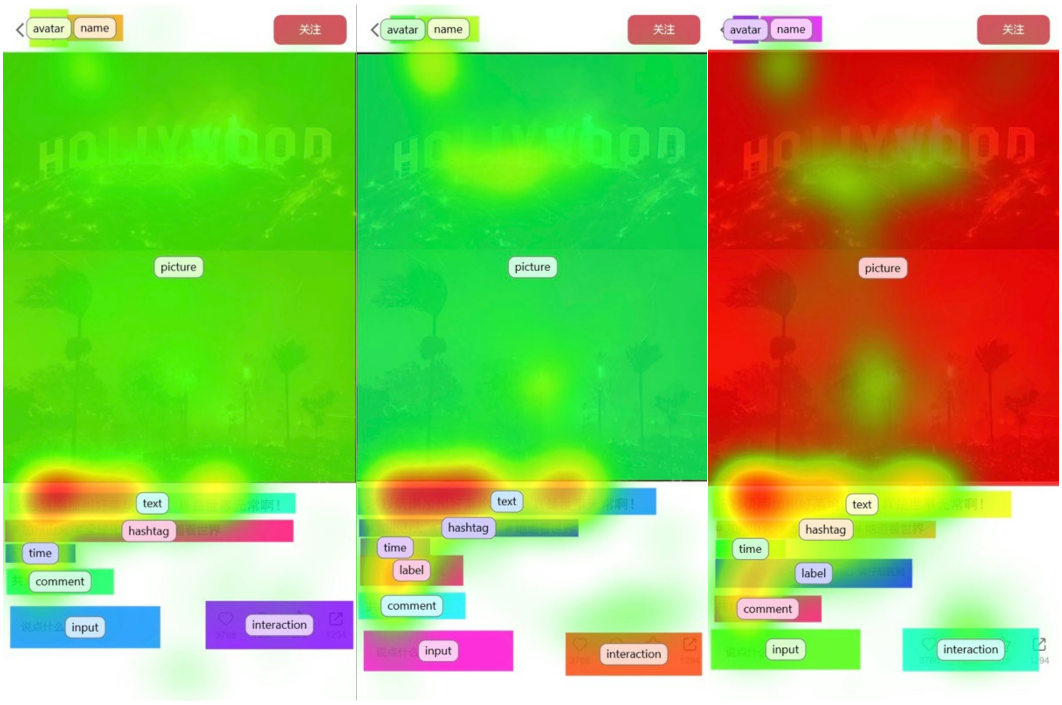


Figure 5. Eye-tracking heatmap of the AI-generated news images category, shown for the control group, experimental group 1, and experimental group 2.

In addition, we looked at the average attention time for each set of materials. We found that users spent more time looking at the AI-generated news images materials than the AI social bots materials. Besides, it is noteworthy that adding AI disclosures had different effects on the materials. For the social bot materials, we observed an inverse relationship: the more detailed and technical the AI disclosure was, the less time participants spent on the materials. However, for the AI-generated news image posts, the intervention effects were more complex. The first experimental group's intervention (experiment group 1) significantly increased the time participants spent observing the posts. Conversely, the second experimental group's intervention (experiment group 2), which included more technical details, reduced participants' attention duration. These findings are consistent with our heatmap analysis results.

Experiment Groups	Material Types	Average Total AOIs Fixation Time/ms
Control Group	AI social bots	1701.56
	AI-generated news images	2516.94
Experiment Group 1	AI social bots	1346.57
	AI-generated news images	2808.74
Experiment Group 2	AI social bots	1166.15
	AI-generated news images	2259.06

Table 2. The structure of the post-test survey.

However, given that overall attention duration is influenced by factors such as the richness of material content and the size of the AOIs, we conducted a more detailed analysis.

Differences in fixation on key areas of interest (AOIs)

In this section, we investigated the differences in AOI attention distribution across our experimental interventions and material types. We first defined a new variable: normalized AOI attention duration, calculated as the attention duration per square pixel within each AOI. This was done to mitigate the influence of an AOI's area and word count on attention distribution. We then separately analysed the AI social bots and AI-generated image materials, conducting a Kruskal-Wallis H-test to identify between-group differences for all common AOIs. This process allowed us to extract the AOIs where attention significantly changed due to the different AI disclosure strategies.

Additionally, we performed a comparative analysis of the AI label AOI attention between experimental groups 1 and 2 to examine how different disclosure strategies influenced user attention and perception. The specific analysis results are as follows:

Materials of AI social bots

For this set of materials, many of the AOIs showed significant between-group differences. These included the basic features of the original poster (such as the avatar and IP address) as well as the fundamental features of the AI bot account.

Notably, experimental group 1 paid significantly more attention to the AI bots' metadata than both the control group and experimental group 2. This suggests that an appropriate level of disclosure can increase a user's attentiveness to an AI bot interaction, a finding that aligns with the conclusions drawn from the eye-tracking heat maps. Furthermore, while the AI bot comments themselves did not show significant between-group differences, participants in experimental group 2 demonstrated a distinctively high level of attention to the content of the original post. This suggests an interest in the overall topic and content of human-AI interaction (Figure 6).

In our analysis of the AI label AOI, experimental group 2 showed a higher average fixation duration on the simple disclosure information than experimental group 1 (Figure 7). This group also paid attention to the detailed information in the hover-over window, suggesting a heightened perception of more technical and complex disclosures. However, this increased attention to the disclosure did not translate into a positive impact on attention to the AI bot's reply, highlighting the potential negative effect of over-disclosure on user interaction.

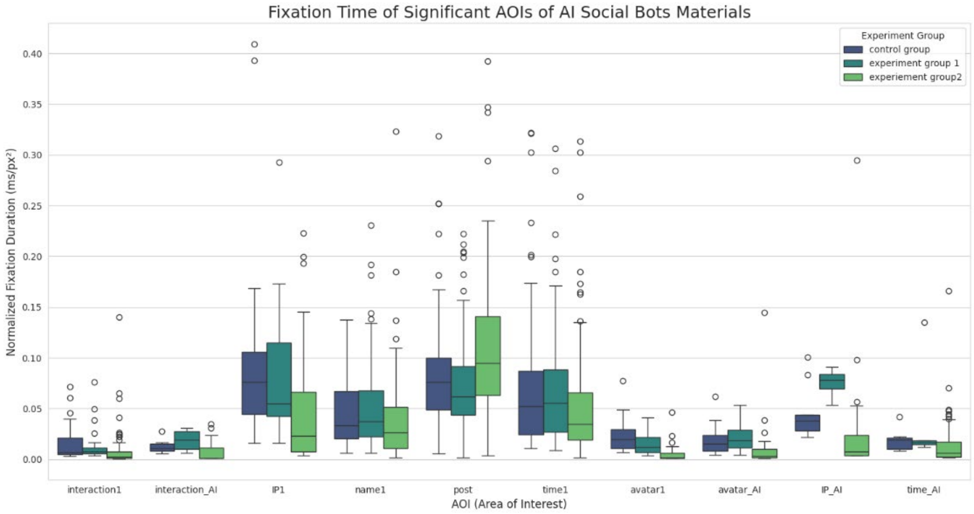


Figure 6. Fixation time of AI disclosure AOIs in experiment groups (AI social bots materials)

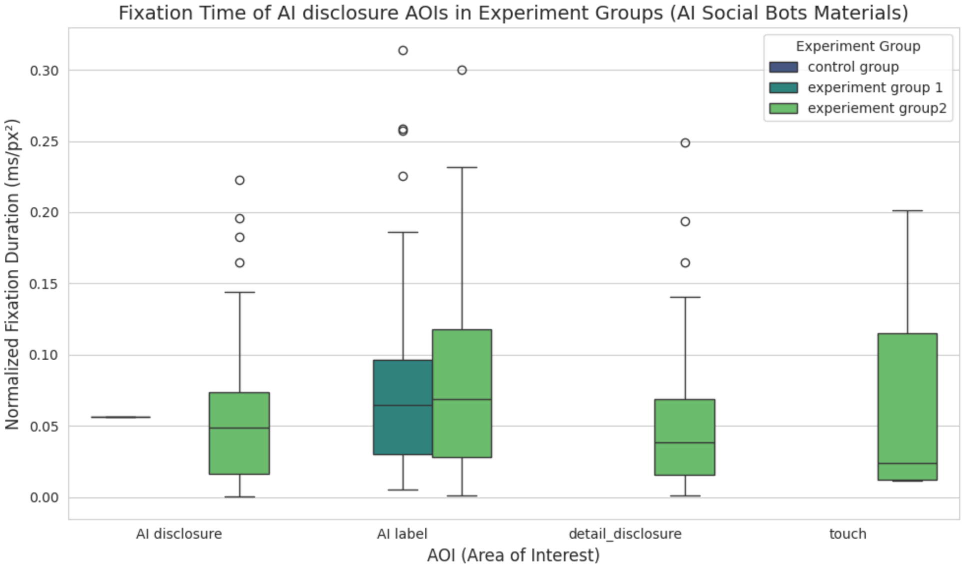


Figure 7. Fixation time of AI disclosure AOIs in experiment groups (AI social bots materials)

Materials of AI-generated news images

For news posts with AI-generated images, the disclosure in experimental group 1 significantly increased attention to both the image and the text, while the more technical and cautionary disclosure in experimental group 2 showed a weaker but still positive effect (Figure 8). Despite being informed that the images were AI-generated, participants continued to attend closely to the accompanying text, indicating a degree of fact-checking awareness when processing AIGC content.

Under the same disclosure location and format, the simple label in experimental group 1 attracted a longer average fixation duration than the technical disclosure in experimental group 2, suggesting that simpler disclosure designs may be more effective in capturing attention and communicating disclosure information (Figure 9).

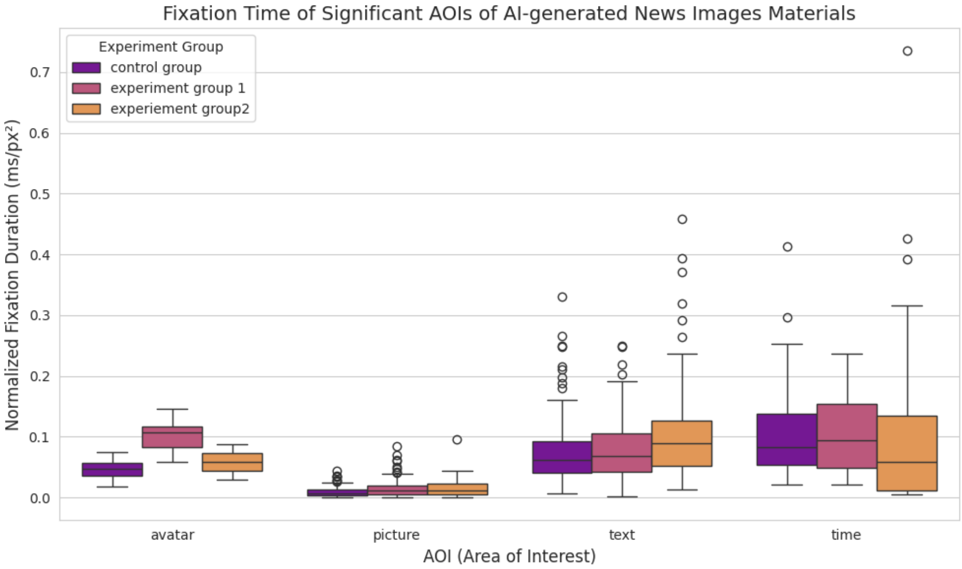


Figure 8. Fixation time of AI disclosure AOIs in experiment groups (AI-generated news images materials).

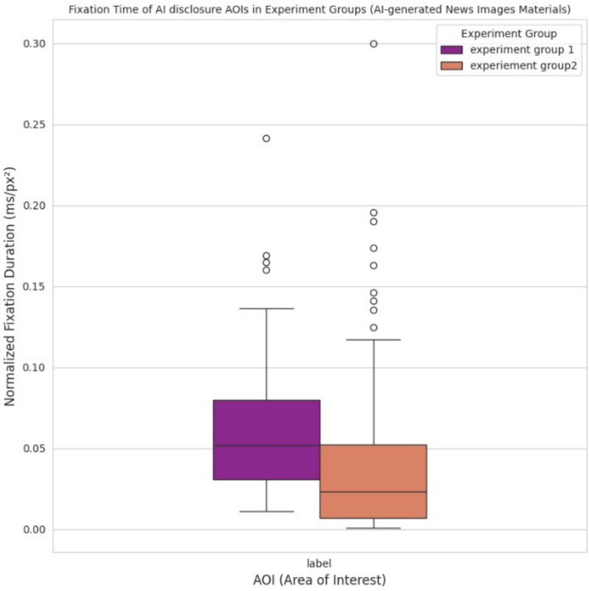


Figure 9. Fixation time of AI disclosure AOIs in experiment groups (AI-generated news images materials).

Overall, disclosing AI information effectively altered users' attention distribution across all AIGC material types. Specifically, simple AI disclosure increased user attention to and verification of AI-generated content. Conversely, over-disclosure diminished interest in both the disclosure information and the AIGC content itself. However, the influence of specific AI disclosure strategies and material types on user attention warrants further, more in-depth investigation.

Survey

Measures, sample, and balance checks

After data cleaning, the analytic sample included 54 high school students from Henan, China (mean age = 16.85; 37% female). Random assignment yielded broadly comparable groups; minor imbalances in internet/mobile experience, adjusted scores, and course track were controlled for in all regression models. All scales demonstrated acceptable to good reliability and validity, including AI trust (α = 0.82), AI literacy (α = 0.80), internet skills (α = 0.69), and AI privacy/security concerns (α = 0.75). The in-task measures for bot comments and AI-generated news images also showed high internal consistency (α = 0.85 and 0.88). The detailed descriptive statistics for all post-test variables and the comprehensive results of the scale validation are provided in Appendix A1 and A2.

Figure 10 provides a preliminary descriptive picture of how disclosure conditions shaped key outcomes. As illustrated in Figure 4-a, levels of AI trust and literacy vary across the three disclosure groups (A = control group without AI labels, B = experiment group 1 with simple disclosure, C = experiment group 2 with detailed disclosure), hinting at initial differences that are further examined in the regression models. Figures 10-b and 10-c show boxplots of interaction-related outcomes for the bots comments and news-image tasks, respectively. The group-level contrasts are visible: while labels appear to raise engagement and trust in conversational settings, they tend to lower trust and sharing in news-like contexts. These descriptive patterns offer an intuitive first look at the intervention effects, with exact statistics reported in Appendix A1–A2.

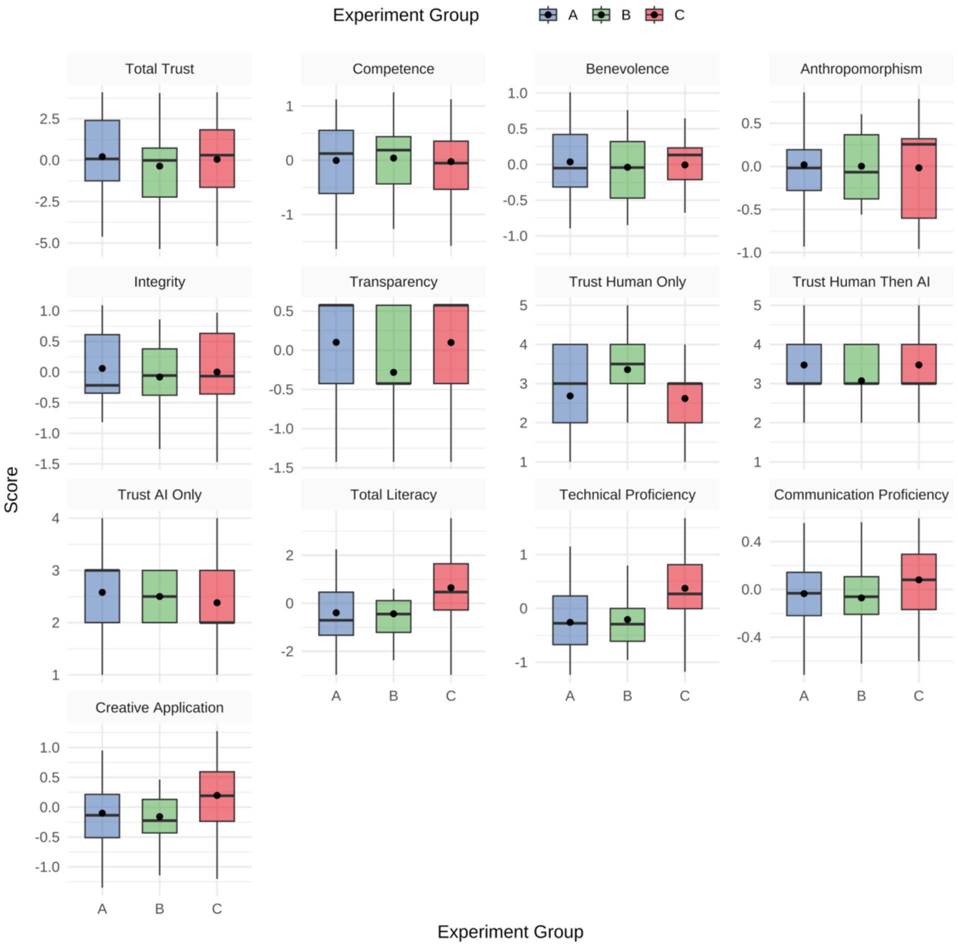


Figure 10-a. Boxplots of AI trust and literacy scores across disclosure conditions.

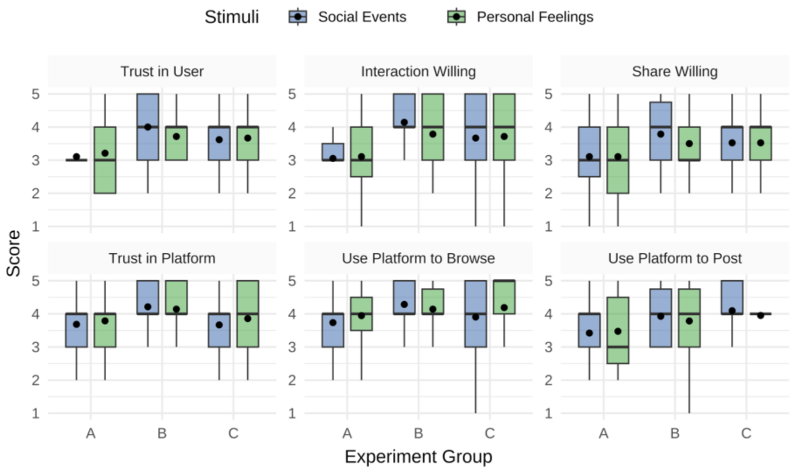


Figure 10-b. Boxplots of trust and interaction intention scores in the bots comments condition

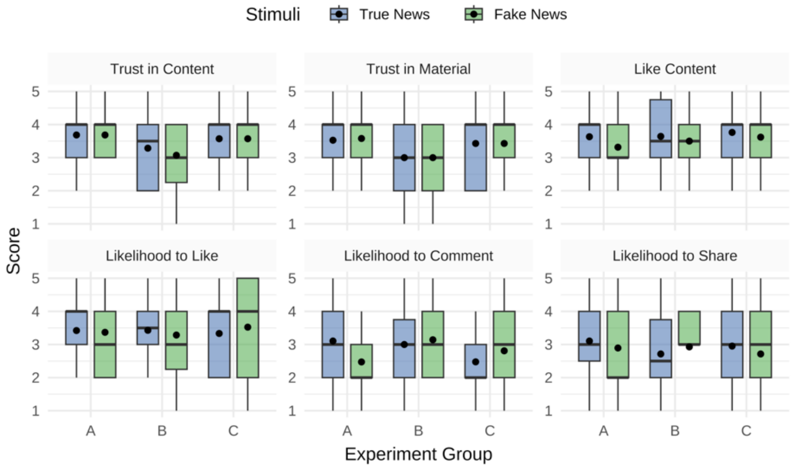


Figure 10-c. Boxplots of trust and interaction intention scores in the news-image condition.

The effect of AI disclosure on trust: labels help in conversation, caution in news

Turning to trust outcomes, OLS regression analyses of the in-task main effects reveal how different disclosure strategies specifically influenced immediate trust and interaction intentions across both social bot and news-image scenarios (detailed coefficients are provided in Appendix A3).

Figure 11-a illustrates treatment effects in the bots comments condition. Adding a simple AI label (B vs. A) significantly increased trust in the interlocutor (β = 0.654, p < .01) and willingness to interact (β = 0.836, p < .01). By contrast, the more detailed disclosure (C vs. A) produced directionally positive but statistically insignificant coefficients. Interestingly, when compared directly to B, condition C slightly reduced platform trust (β = -0.791, p < .01), suggesting that additional technical detail may trigger caution about the platform, even as interpersonal rapport remains intact.

Figure 11-b presents treatment effects in the news-image condition. Here, disclosure cues had the opposite impact: B vs. A reduced trust in the content and material (both p < .05), while C vs. A significantly lowered the intention to share (β ≈ –0.53, p < .05). In other words, labels and especially explanatory cues dampen trust and diffusion in news-like tasks. This contrast points to a context-dependent dynamic: students are permissive in conversational settings but adopt a more sceptical stance when credibility is at stake, is consistent with the observed distribution of visual attention.

Figure 11b presents treatment effects in the news-image condition. Disclosure cues had the opposite effect: relative to the control group, simple disclosure reduced trust in the content and material (both p < .05), while detailed disclosure significantly lowered sharing intention (β ≈ −0.53, p < .05). Overall, disclosure—especially explanatory cues—dampened trust and diffusion in news-like tasks. This context-dependent pattern aligns with the observed distribution of visual attention.

Survey-level results, summarized in Figure 11-c, show that disclosure did not erode generalized AI trust. Treatment effects on the post-test trust composite were modest, and there is no evidence of spillover distrust once covariates were accounted for.

Control variables also yield meaningful insights: higher academic scores predicted slightly lower platform trust (β ≈ −0.003 to −0.005, p < .05), indicating more critical stances among high achievers. Heavier internet users reported lower platform trust and reduced posting willingness (β ≈ −0.11, p < .05), consistent with greater awareness of online risks. Detailed baseline and controlled regression models for post-test outcomes are provided in Appendix A4.

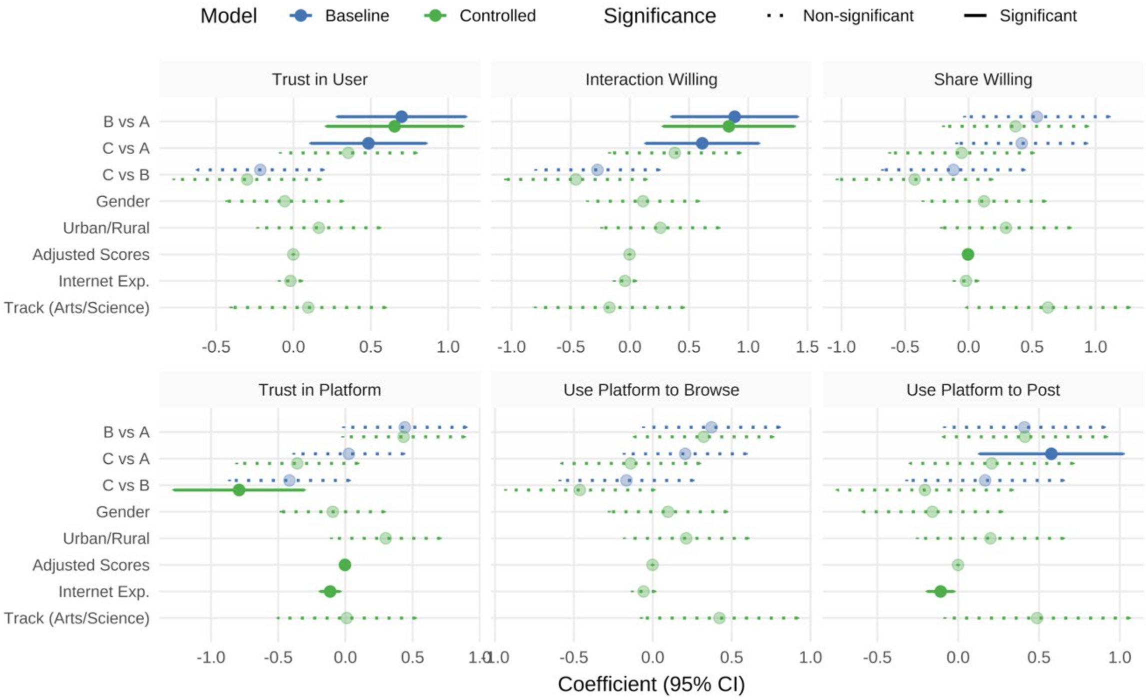


Figure 11-a. Forest plots of treatment effects (B vs. A, C vs. A, C vs. B) on trust and interaction variables in the bots comments condition.

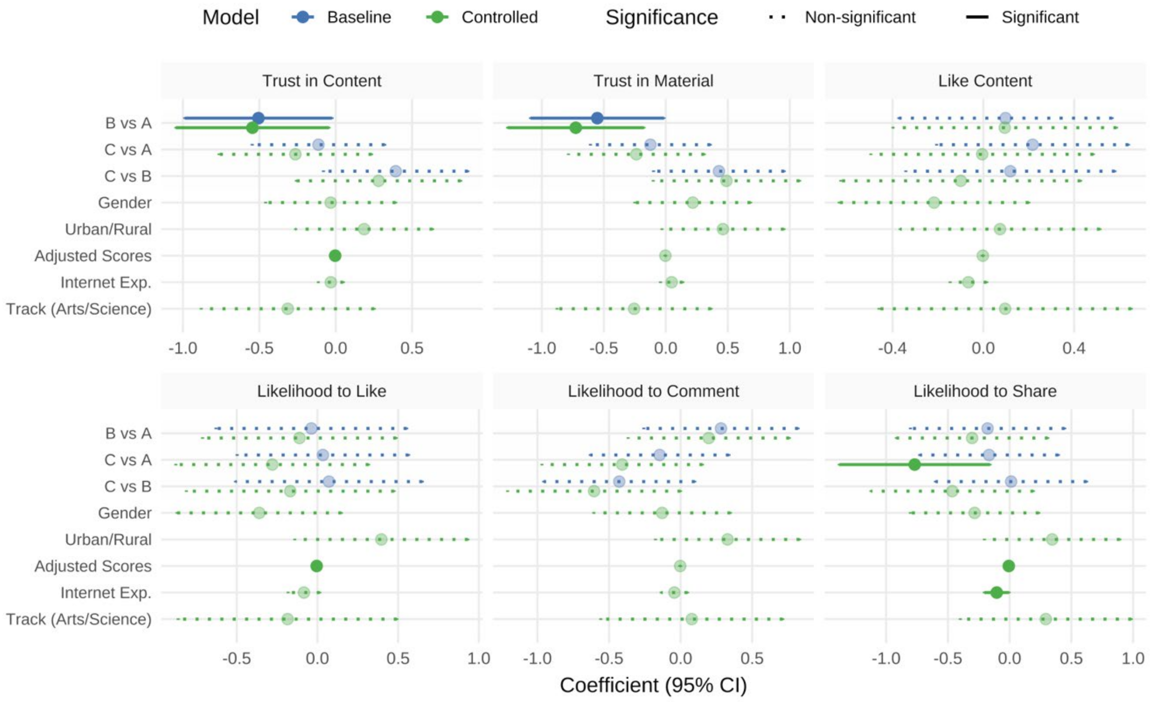


Figure 11-b. Forest plots of treatment effects (B vs. A, C vs. A, C vs. B) on trust and interaction variables in the news-image condition.

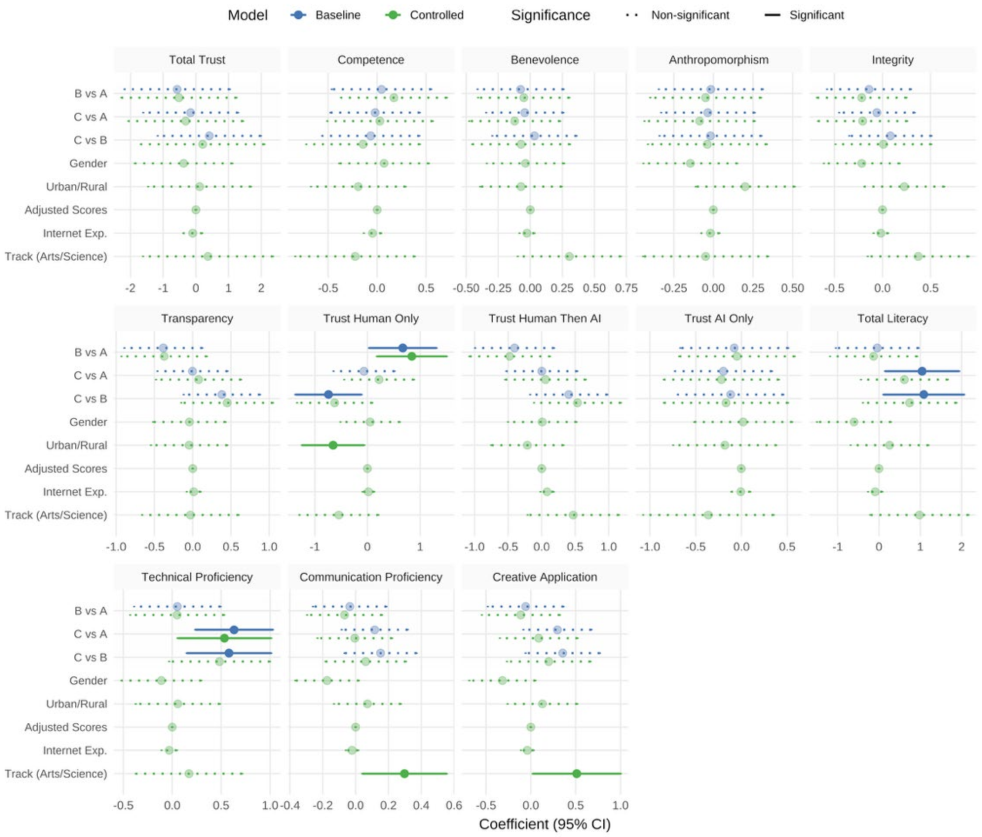


Figure 11-c. Forest plots of treatment effects (B vs. A, C vs. A, C vs. B) on survey-based trust and literacy scales.

The effect of AI disclosure on interaction intention: talk more, share less

Figure 11-a also highlights that in the bots comments condition, disclosure influenced interaction intention. B vs. A significantly increased willingness to interact (β = 0.836, p < .01). At the same time, effects on posting to the platform were mixed once covariates were included, with more experienced internet users less likely to post (β = –0.108, p < .05). Together, these results suggest that simple AI labels lower uncertainty costs and encourage private dialogue, but do not necessarily translate into greater public sharing.

In the news-image condition, Figure 11-b shows that C vs. A reduced sharing likelihood (p < .05). Effects on likes and comments were smaller and less consistent, indicating that disclosure especially tempers outward diffusion rather than all forms of engagement. Post-test results further confirm that disclosure did not generate wholesale preferences for AI-only interactions; instead, students continued to prefer human or mixed sources, consistent with their cautious stance toward news credibility.

Heterogeneity analysis: detailed explanations win with heavy internet users

Finally, moderation analyses demonstrate that internet-use time conditions the effectiveness of disclosure. Figure 5 shows that the B vs. A × internet-time interaction was negative on outcomes such as AI trust (β = -1.868, p < .05), anthropomorphism (β = -0.457, p < .01), literacy (β = -1.195, p < .05), communication proficiency (β = -0.258, p < .05) and creative application (β = -0.513, p < .05), indicating that the simple label worked best for light users but attenuated for heavy users. In direct contrasts, C vs. B × internet-time was positive and significant for overall AI trust (β = 1.387, p < .05), anthropomorphism (β = 0.264, p<.05) and technical proficiency (β = 0.371, p < .05), meaning that adding a one-sentence explanation increasingly outperformed the simple label among more experienced students. This pattern suggests that simple labels are newcomer-friendly, while brief explanations are more effective for heavy internet users. Full regression tables for these interactions are provided in Appendix A5.

Robustness and diagnostics

All OLS regressions were estimated with robust standard errors and controls for gender, urban/rural background, adjusted scores, course track, and internet/mobile experience. Multicollinearity diagnostics confirmed that all VIF values were below 5. Although Breusch–Pagan and White tests indicated heteroskedasticity for some variables, the issue was addressed by using heteroskedasticity-robust standard errors, ensuring that the reported estimates remain valid.

Reading across the findings

Taken together, these results highlight that disclosure has asymmetric consequences. For high-school students, AI labels promote conversational trust and willingness to interact, while simultaneously discouraging uncritical sharing of AI-generated news. Explanatory disclosures carry additional benefits for students with greater digital experience. For educators, this implies that disclosure design should be context- and cohort-sensitive: simple AI-generated labels are sufficient for classroom chat activities, while AI-generated plus '*how*' explanations better support verification in news-evaluation and advanced literacy tasks.

Thematic analysis

We conducted a thematic analysis of 60 interview transcripts. Beginning with open coding, we progressively grouped the data into four overarching themes regarding rural high school students' trust and interaction with AIGC.

Trust and cognition: contextual and conditional evaluation of AIGC

Interview results show that adolescents' trust in AI-generated content is highly conditional and context-dependent. In comment-based interactions, AI disclosure was commonly perceived as a transparency cue that reduced uncertainty and facilitated trust. As one respondent noted, ‘*Labeling can enhance trust between AI and people in some fields*’ (Respondent 025), consistent with the survey finding that disclosure increased trust in AI-generated comments.

In contrast, news-related scenarios elicited greater scepticism. Respondents frequently associated AI-generated news with risks of exaggeration or misinformation, and disclosure cues further amplified these concerns. As Respondent 022 stated, ‘AI-*generated news can be misleading, especially when the facts are exaggerated.*’ Here, disclosure functioned primarily as a warning signal, leading to trust withdrawal rather than analytical evaluation, indicating that trust calibration occurred without deeper understanding of AI generation processes.

Cognitively, most adolescents relied on intuitive judgments when assessing AIGC, with only a few engaging in analytical strategies such as checking sources or cross-verifying information. As Respondent 027 explained, ‘*I would look at whether the content cites any sources; otherwise, it’s hard to tell.*’ *This* heuristic processing led to two recurring errors: mistaking unlabeled AI-generated content for human-produced material and over-doubting labelled AI content. Together, these patterns help explain why AI disclosure did not significantly improve AI literacy.

AI interaction behavior: practicality as the core, with boundaries in emotional needs
Participants described their interactions with AI as primarily pragmatic and goal-oriented. Many reported using AI for information retrieval, homework assistance, and efficiency-driven tasks. In addition, some participants engaged with AI for casual entertainment or emotional relief during boredom. As one respondent noted, ‘*I was really bored, so I played with* AI'(Respondent 022). These accounts suggest that AI is integrated into everyday practices as a convenient and functional tool.

However, the interviews also revealed clear emotional boundaries in AI interaction. Explicit AI disclosure sometimes disrupted emotional engagement by emphasizing the non-human nature of the interaction. One participant remarked, ‘*There’s no real person responding, only robots, which makes the experience worse*’ (Respondent 043). In social contexts, several respondents perceived AI-generated comments as inauthentic or strategically designed to boost engagement. These perceptions undermined trust and interaction, indicating that disclosure can suppress emotional resonance without promoting deeper cognitive understanding of AI systems.

Perception of disclosure mechanisms: labeling is necessary but requires moderation
Participants consensus that AIGC needs labelling, as it guides discernment: ‘*Without labels, people won't distinguish true from false*’ (Respondent 008). They favour ‘*simple, clear text prompts*’ *for* low cognitive load and prominent placement: ‘*Labels should be in content's eye-catching spot, not page corners.*’

Overly complex disclosures divert focus from content to AI identity, heightening cognitive and emotional burdens and causing trust/interaction countereffects, matching questionnaire results.

Platform responsibility: necessity of supervision and preferences for supervision methods
Participants broadly agreed that platforms should take responsibility for supervising AI-generated content to protect authenticity and user experience. Several respondents emphasized the need to prevent exaggerated or misleading outputs, particularly in visual content. As Respondent 022 noted, ‘AI *images shouldn’t be exaggerated; otherwise, people won’t trust them.*’ Others suggested platform-level accountability mechanisms, such as reporting functions for AI accounts (Respondent 023).

While attitudes toward large-scale deployment of AI bots were generally neutral, respondents exposed to more detailed disclosures expressed stronger concerns about emotional discomfort, spam, and excessive AI presence. Respondent 049 mentioned that ‘*too many AI bots can generate negative emotions*,’ while Respondent 048 emphasized the need to ‘*avoid disputes*’ caused by AI participation. A small number of participants suggested that AI accounts should eventually be treated similarly to human users, with reduced disclosure intensity once norms stabilize. As Respondent 024 explained, '*At the beginning it's necessary, but after many years, spending too much effort on supervision may be unnecessary*.' These views reinforce adolescents' preference for moderation and further illustrate that excessive emphasis on AI identity may undermine trust without contributing to AI literacy development.

Conclusion and discussion

This study investigated how different AI disclosure mechanisms influence the cognitive processing and interaction of high school students in Chinese county towns when encountering AI-generated news and comments. By conducting a mixed-methods field experiment, we aimed to address the research gap concerning non-Western, non-urban, and non-university populations. Our multi-dimensional data, derived from eye-tracking, questionnaires, and interviews, offered a comprehensive understanding of AI disclosure strategies' impact on adolescent trust and interaction behaviors.

Summary of key findings

Firstly, AI disclosure significantly influenced attention to AI-generated content (AIGC). Eye-tracking data indicated that simple disclosure (experimental group 1) increased participants' focus on AI metadata and content in both social bot and news scenarios, whereas overly detailed disclosure (experimental group 2) reduced engagement with both the label and the content, suggesting that moderate disclosure helps avoid cognitive overload.

Secondly, the effects of disclosure on trust and interaction were context-dependent. In conversational contexts (AI social bot replies), simple labels enhanced trust and interaction willingness, supported by interview reports that labels clarified content sources and reduced uncertainty. Detailed disclosure slightly decreased platform trust. In contrast, for AI-generated news images, simple labels lowered content trust, while detailed disclosure significantly reduced sharing intention, as participants cited potential misleading risks. Overall, H1a (positive effect of simple disclosure) was supported only in conversational contexts, whereas H1b (negative effect of detailed disclosure) was evident in news contexts and partially in platform trust for social bots.

Thirdly, AI disclosure did not universally enhance AI literacy (H1c). However, individual differences moderated these effects (H3b): simple labels were more effective for light internet users, whereas brief explanations were more beneficial for heavy users in fostering AI trust and technical understanding. Academic performance and internet experience also shaped critical engagement, with higher achievers and heavier users exhibiting lower platform trust and reduced sharing.

Finally, while perceived topic and truthfulness (H2b) appeared relevant—participants often rejected AI news due to 'misleading risks'—quantitative evidence directly supporting this effect was not observed.

Discussions

Theoretical contributions

This research contributes empirical evidence to human-AI interaction models, particularly within an underexplored demographic and cultural context: high school students in Chinese county towns. By revealing the context-dependent impact of AI disclosure on adolescent trust and interaction, our findings help explain the mixed results in previous AI transparency research. The study demonstrates that disclosure's effect is not uniform, varying with content type, disclosure detail, and user individual differences. The multi-modal data collection (eye-tracking, surveys, interviews) offers a comprehensive view of cognitive processing and real-time trust dynamics, advancing methodological innovation in trust measurement.

In addition, our findings extend prior student-focused research by revealing a different pattern of individual differences. Unlike studies conducted mainly in urban or university settings (Kozak & Fel, 2024), demographic factors such as gender and grade level did not significantly moderate the effects of AI disclosure. Instead, internet use experience emerged as a key moderator shaping how disclosure influenced trust. Students with more extensive internet experience were better able to interpret relatively detailed and technical disclosure cues, demonstrating stronger capacities for AI identification and adaptive judgment. This pattern highlights a theoretically meaningful distinction in how disclosure operates among adolescents in county-level cities in China, where uneven digital exposure makes experiential factors more consequential than basic demographics.

Practical contributions

The findings provide actionable insights for technology developers, educators, and policymakers to create ethical and trustworthy AI systems.

For developers: AI disclosure design should be context- and user-sensitive. Simple AI labels are sufficient for casual chat interactions, promoting trust, while news evaluation and advanced literacy tasks require more explanatory disclosures (e.g., *'how it was generated'*) to support verification. Overly complex disclosures can be counterproductive, leading to cognitive and emotional burden.

For educators: Acknowledge students' context-dependent trust in AIGC and design teaching activities that foster critical thinking and information discernment based on content type and disclosure strategies. Moreover, participants frequently highlighted their emotional needs for AI bots. Given media reports, including from Bloomberg, that interactions with AI bots (e.g., Replika, Character AI) can lead to addictive behaviors or even encouragement of self-harm, educators must guide students in developing healthy emotional engagement with these systems, preventing excessive trust or dependency that could result in tragic outcomes.

For policymakers: Support mandatory, but concise and prominently placed AI labeling for AIGC content. Platforms should supervise AI-generated content for authenticity and user experience, considering feedback channels. A long-term perspective might involve reducing oversight as AI technology matures, avoiding excessive intervention that could lead to negative user sentiment.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grant No. 72304017), the China Association for Science and Technology (Grant No. 2023QNRC001), and the Beijing Social Science Funding (Grant No. 22XCC013). We would like to thank all research contributors and participants, including those who took part in the field experiment, as well as the anonymous reviewers for their valuable feedback. We also thank Tobii for providing the eye-tracking equipment and analysis software used in this study.

About the authors

Nuo Chen is a master's student in Information Science at the Department of Information Management, Peking University. Her research focuses on human–AI interaction, AI-mediated communication, and the social and ethical implications of generative AI. She can be contacted at nuochen@stu.pku.edu.cn

Zhiyuan Lai is an undergraduate student in the Department of Information Management at Peking University. His research centers on the socio-technical dynamics of digital technologies, with a focus on identity formation, social relationships, and inequality. He can be contacted at lzy1122@stu.pku.edu.cn

Yichu Liu is an undergraduate student at the Department of Information Management, Peking University. He is currently engaged in research projects spanning algorithmic literacy, the intelligent digital divide, and human-computer interaction. He can be contacted at liuyichu123@stu.pku.edu.cn

Jia Li is an undergraduate student in the Department of Information Management at Peking University. Her research interests include the psychological impacts of human-chatbot interaction, chatbot anthropomorphism, and the social and ethical dimensions of AI. She can be contacted at lijia23@stu.pku.edu.cn.

Rui Wang received M.S. in Information Science in 2025 from the Department of Information Management, Peking University. Her research focuses on algorithmic auditing and the transparency of generative artificial intelligence. She can be contacted at wangruiwan1203@163.com

Dr. Pu Yan is an assistant professor at the department of information management, Peking University. She focuses on Computational Social Science, Digital Inequality, and Algorithmic Governance. She holds a PhD degree on Information, Communication, and Social Science from the University of Oxford (Oxford Internet Institute). Correspondence concerning this article should be addressed to puyan@pku.edu.cn

References

Ajenaghughrure, I. B., da Costa Sousa, S. C., & Lamas, D. (2020, June 30–July 3). Risk and trust in artificial intelligence technologies: A case study of Autonomous Vehicles. 2020 13th International Conference on Human System Interaction (HSI), 118–123. https://doi.org/10.1109/HSI49210.2020.9142686

Akash, K., Hu, W.-L., Jain, N., & Reid, T. (2018). A classification model for sensing human trust in machines using EEG and GSR. ACM Transactions on Interactive Intelligent Systems, 8(4), 1–20. https://doi.org/10.1145/3132743

Amoozadeh, M., Daniels, D., Nam, D., Kumar, A., Chen, S., Hilton, M., ... & Alipour, M. A. (2024, March). Trust in generative AI among students: An exploratory study. Proceedings of the 55th ACM Technical Symposium on Computer Science Education V. 1, 67–73. https://doi.org/10.1145/3626252.3630757

Balakrishnan, J., & Dwivedi, Y. K. (2021). Role of cognitive absorption in building user trust and experience. Psychology & Marketing, 38(4), 643–668. https://doi.org/10.1002/mar.21462

Bindewald, J. M., Rusnock, C. F., & Miller, M. E. (2018). Measuring human trust behavior in human-machine teams. In D. N. Cassenti (Ed.), Advances in human factors in simulation and modeling (Vol. 591, pp. 47–58). Springer. https://doi.org/10.1007/978-3-319-60591-3_5

Bochniarz, K. T., Czerwiński, S. K., Sawicki, A., & Atroszko, P. A. (2022). Attitudes to AI among high school students: Understanding distrust towards humans will not help us understand distrust towards AI. Personality and Individual Differences, 185, Article 111299. https://doi.org/10.1016/j.paid.2021.111299

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. Qualitative research in psychology, 3(2), 77-101.

Chan, C. K. Y. (2023). A comprehensive AI policy education framework for university teaching and learning. International Journal of Educational Technology in Higher Education, 20(1), 38. https://doi.org/10.1186/s41239-023-00408-3

Chang, T., Trybala, J. J., Bassan, S., & Razi, A. (2025, May 10–15). Opaque transparency: Gaps and discrepancies in the report of social media harms. Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '25), Article 424, 1–12. https://doi.org/10.1145/3706599.3719829

Chattaraman, V., Kwon, W.-S., Gilbert, J. E., & Ross, K. (2019). Should AI-Based, conversational digital assistants employ social- or task-oriented interaction style? A task-competency and reciprocity perspective for older adults. Computers in Human Behavior, 90, 315–330. https://doi.org/10.1016/j.chb.2018.08.048

Chen, J. Y. C., & Barnes, M. J. (2014). Human–Agent teaming for multirobot control: A review of human factors issues. IEEE Transactions on Human-Machine Systems, 44(1), 13–29. https://doi.org/10.1109/THMS.2013.2293535

Choudhury, A., & Shamszare, H. (2023). Investigating the impact of user trust on the adoption and use of ChatGPT: Survey analysis. Journal of Medical Internet Research, 25, e47184. https://doi.org/10.2196/47184

Choudhury, A., & Shamszare, H. (2024). The impact of ferformance expectancy, workload, risk, and satisfaction on trust in ChatGPT: Cross-Sectional Survey Analysis. JMIR Human Factors, 11, e55399. https://doi.org/10.2196/55399

De Freitas, J., Agarwal, S., Schmitt, B., & Haslam, N. (2023). Psychological factors underlying attitudes toward AI tools. Nature Human Behaviour, 7(11), 1845–1854. https://doi.org/10.1038/s41562-023-01734-2

De Visser, E. J., Pak, R., & Shaw, T. H. (2018). From ‘automation’to ‘autonomy’: the importance of trust repair in human–machine interaction. Ergonomics, 61(10), 1409-1427. https://doi.org/10.1080/00140139.2018.1457725

Ehsan, U., Tambwekar, P., Chan, L., Harrison, B., & Riedl, M. O. (2019, March 17–20). Automated rationale generation: a technique for explainable AI and its effects on human perceptions. In Proceedings of the 24th international conference on intelligent user interfaces (pp. 263-274).

European Commission. (2018). General Data Protection Regulation (GDPR) – Legal Text. General Data Protection Regulation (GDPR). https://gdpr-info.eu/

Gao, L., & Waechter, K. A. (2017). Examining the role of initial trust in user adoption of mobile payment services: an empirical investigation. Information Systems Frontiers, 19(3), 525-548.

Gremillion, G. M., Metcalfe, J. S., Marathe, A. R., Paul, V. J., Christensen, J., Drnec, K., Haynes, B., & Atwater, C. (2016). Analysis of trust in autonomy for convoy operations. Micro- and Nanotechnology Sensors, Systems, and Applications VIII (Vol. 9836, pp. 356-365). SPIE. https://doi.org/10.1117/12.2224009

Hancock, J. T., & Bailenson, J. N. (2021). The social impact of deepfakes. Cyberpsychology, behavior, and social networking, 24(3), 149-152.

Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. Human Factors: The Journal of the Human Factors and Ergonomics Society, 57(3), 407–434. https://doi.org/10.1177/0018720814547570

Hu, P., Lu, Y., & Gong, Y. (2021). Dual humanness and trust in conversational AI: A person-centered approach. Computers in Human Behavior, 119, 106727. https://doi.org/10.1016/j.chb.2021.106727

Hwang, H. S., Zhu, L. C., & Cui, Q. (2023). Development and Validation of a Digital Literacy Scale in the Artificial Intelligence Era for College Students. KSII Transactions on Internet and Information Systems (TIIS), 17(8), 2241–2258. https://doi.org/10.3837/tiis.2023.08.016

Jarrahi, M. H. (2018). Artificial intelligence and the future of work: Human-AI symbiosis in organizational decision making. Business Horizons, 61(4), 577–586. https://doi.org/10.1016/j.bushor.2018.03.007

Jiang, X., Wu, Z., & Yu, F. (2024). Constructing consumer trust through artificial intelligence generated content. Academic Journal of Business & Management, 6(8), 263-272.

Kim, T., & Hinds, P. (2006). Who should I blame? Effects of autonomy and transparency on attributions in Human-Robot interaction. ROMAN 2006 - The 15th IEEE International Symposium on Robot and Human Interactive Communication, 80–85. https://doi.org/10.1109/ROMAN.2006.314398

Kizilcec, R. F. (2016, May 7–12). How much information?: Efffects of transparency on trust in an algorithmic interface. Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, 2390–2395. https://doi.org/10.1145/2858036.2858402

Koo, J., Kwac, J., Ju, W., Steinert, M., Leifer, L., & Nass, C. (2015). Why did my car just do that? Explaining semi-autonomous driving actions to improve driver understanding, trust, and performance. International Journal on Interactive Design and Manufacturing (IJIDeM), 9(4), 269–275. https://doi.org/10.1007/s12008-014-0227-2

Kozak, J., & Fel, S. (2024). How sociodemographic factors relate to trust in artificial intelligence among students in Poland and the United Kingdom. Scientific Reports, 14(1), 28776.

Kozyreva, A., Lorenz-Spreen, P., Hertwig, R., Lewandowsky, S., & Herzog, S. (2021). Public attitudes towards algorithmic personalization and use of personal data online: Evidence from Germany, Great Britain, and the United States. Humanities & Social Sciences Communications, 8(1). https://doi.org/10.1057/s41599-021-00787-w

Lee, C., & Cha, K. (2024). Toward the dynamic relationship between AI transparency and trust in AI: A case study on ChatGPT. International Journal of Human–Computer Interaction, 0(0), 1–18. https://doi.org/10.1080/10447318.2024.2405266 Hancock, J. T., & Bailenson, J. N. (2021). The social impact of deepfakes. Cyberpsychology, behavior, and social networking, 24(3), 149-152.

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. Human Factors, 46(1), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392

Lee, S., & Park, G. (2024). Development and validation of ChatGPT literacy scale. Current Psychology, 43(21), 18992–19004. https://doi.org/10.1007/s12144-024-05723-0

Lewis, P. R., & Marsh, S. (2022). What is it like to trust a rock? A functionalist perspective on trust and trustworthiness in artificial intelligence. Cognitive Systems Research, 72, 33–49. https://doi.org/10.1016/j.cogsys.2021.11.001

Liao, T., & MacDonald, E. F. (2021). Manipulating users' trust of autonomous products with affective priming. Journal of Mechanical Design, 143(5), 051402. https://doi.org/10.1115/1.4048640

Linegang, M. P., Stoner, H. A., Patterson, M. J., Seppelt, B. D., Hoffman, J. D., Crittendon, Z. B., & Lee, J. D. (2006). Human-Automation collaboration in dynamic mission planning: A challenge requiring an ecological approach. Proceedings of the Human Factors and

Ergonomics Society Annual Meeting, 50(23), 2482–2486. https://doi.org/10.1177/154193120605002304

Lintner, T. (2024). A systematic review of AI literacy scales. NPJ Science of Learning, 9(1), 50. https://doi.org/10.1038/s41539-024-00264-4

Lund, B. D., Wang, T., Mannuru, N. R., Nie, B., Shimray, S., & Wang, Z. (2023). ChatGPT and a new academic reality: Artificial intelligence-written research papers and the ethics of the large language models in scholarly publishing. Journal of the Association for Information Science and Technology, 74(5), 570–581. https://doi.org/10.1002/asi.24750

Luo, X., Tong, S., Fang, Z., & Qu, Z. (2019). Frontiers: Machines vs. Humans: The impact of artificial intelligence chatbot disclosure on customer purchases. Marketing Science, 38(6). https://doi.org/10.1287/mksc.2019.1192

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. The Academy of Management Review, 20(3), 709. https://doi.org/10.2307/258792

Molnar, L. J., Ryan, L. H., Pradhan, A. K., Eby, D. W., St. Louis, R. M., & Zakrajsek, J. S. (2018). Understanding trust and acceptance of automated vehicles: An exploratory simulator study of transfer of control between automated and manual driving. Transportation Research Part F: Traffic Psychology and Behaviour, 58, 319–328. https://doi.org/10.1016/j.trf.2018.06.004

Naiseh, M., Al-Mansoori, R. S., Al-Thani, D., Jiang, N., & Ali, R. (2021, October). Nudging through friction: an approach for calibrating trust in explainable AI. In 2021 8th International Conference on Behavioral and Social Computing (BESC) (pp. 1-5). IEEE.

Oleson, K. E., Billings, D. R., Kocsis, V., Chen, J. Y. C., & Hancock, P. A. (2011). Antecedents of trust in human-robot collaborations. 2011 IEEE International Multi-Disciplinary Conference on Cognitive Methods in Situation Awareness and Decision Support (CogSIMA), 175–178. https://doi.org/10.1109/COGSIMA.2011.5753439

Poursabzi-Sangdeh, F., Goldstein, D. G., Hofman, J. M., Wortman Vaughan, J. W., & Wallach, H. (2021). Manipulating and measuring model interpretability. Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, 1–52. https://doi.org/10.1145/3411764.3445315

Ramirez, J. P., Obenza, D. M., & Cuarte, R. (2024). AI trust and attitude towards AI of university students. International Journal of Multidisciplinary Studies in Higher Education, 1(1), 22-36.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. Psychological Bulletin, 124(3), 372–422. https://doi.org/10.1037/0033-2909.124.3.372

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). 'Why should I trust you?': Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1135–1144. https://doi.org/10.1145/2939672.2939778

Rossner, A., Cassel, M., & Huschens, M. (2024). Do users really care? Evaluating the user perception of disclosing AI-Generated content on credibility in (sports) Journalism. Proceedings of the 2024 Conference on Mensch und Computer (MuC '24), 413–418. https://doi.org/10.1145/3670653.3677490

Samonte, M. J. C., Escarillo, J. G. M., Go, K., Landrito, N. B. A., & Randhawa, J. K. (2023, August 24–26). Determining the trust level of senior high school associated with the use of AI-powered digital assistants. Proceedings of the 2023 6th International Conference on Information Management and Management Science, 54–61. https://doi.org/10.1145/3615431.3615442

Schaefer, K. E., Chen, J. Y. C., Szalma, J. L., & Hancock, P. A. (2016). A meta-analysis of factors influencing the development of trust in automation: Implications for understanding autonomy in future systems. Human Factors: The Journal of the Human Factors and Ergonomics Society, 58(3), 377–400. https://doi.org/10.1177/0018720816634228

Schmidt, P., Biessmann, F., & Teubner, T. (2020). Transparency and trust in artificial intelligence systems. Journal of Decision Systems, 29(4), 260–278. https://doi.org/10.1080/12460125.2020.1819094

Shin, D. (2021). The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI. International Journal of Human-Computer Studies, 146, Article 102551. https://doi.org/10.1016/j.ijhcs.2020.102551

Shin, D., & Park, Y. J. (2019). Role of fairness, accountability, and transparency in algorithmic affordance. Computers in Human Behavior, 98, 277–284. https://doi.org/10.1016/j.chb.2019.04.019

Sinha, R., & Swearingen, K. (2002). The role of transparency in recommender systems. CHI ’02 Extended Abstracts on Human Factors in Computing Systems, 830–831. https://doi.org/10.1145/506443.506619

Sunnie S. Y. Kim. (2024, May 11–16). Establishing appropriate trust in AI through transparency and explainability. Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '24), Article 433, 1–6. https://doi.org/10.1145/3613905.3638184

Thielmann, I., & Hilbig, B. E. (2014). Trust in me, trust in you: A social projection account of the link between personality, cooperativeness, and trustworthiness expectations. Journal of Research in Personality, 50, 61–65. https://doi.org/10.1016/j.jrp.2014.03.006

Tossell, C. C., Tenhundfeld, N. L., Momen, A., Cooley, K., & De Visser, E. J. (2024). Student Perceptions of ChatGPT Use in a College Essay Assignment: Implications for Learning, Grading, and Trust in Artificial Intelligence. IEEE Transactions on Learning Technologies, 17, 1069–1081. https://doi.org/10.1109/TLT.2024.3355015

Ueno, T., Sawa, Y., Kim, Y., Urakami, J., Oura, H., & Seaborn, K. (2022, April 29–May 5). Trust in human-AI interaction: Scoping out models, measures, and methods. CHI Conference on Human Factors in Computing Systems Extended Abstracts, Article 400, 1–7. https://doi.org/10.1145/3491101.3516390

Ullman, D., & Malle, B. F. (2019, March 11–14). Measuring gains and losses in human-robot trust: Evidence for differentiable components of trust. 2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI), 618–619. https://doi.org/10.1109/HRI.2019.8673154

van Deursen, A., & van Dijk, J. (2010). Internet skills and the digital divide. New Media & Society, 13(6), 893-911. https://doi.org/10.1177/1461444810386774

Wang, W., & Benbasat, I. (2007). Recommendation agents for electronic commerce: Effects of explanation facilities on trusting beliefs. Journal of Management Information Systems, 23(4), 217–246. https://doi.org/10.2753/MIS0742-1222230410

Wittenberg, C., Epstein, Z., Berinsky, A. J., & Rand, D. G. (2024). Labeling AI-generated content: Promises, perils, and future directions. An MIT Exploration of Generative AI. https://doi.org/10.21428/e4baedd9.0319e3a6

Zhou, T., & Lu, H. (2025). The effect of trust on user adoption of AI-generated content. The Electronic Library, 43(1), 61–76. https://doi.org/10.1108/EL-08-2024-0244

Appendix A1. Descriptive statistics (post-test factor scores)

	Mean	SDs	Min	Max
Total Skills	0	1.79	-4.14	4.16
Operation Skills	0	0.32	-0.64	0.66
Informational Skills	0	0.98	-2.57	1.43
Participation Skills	0	1.01	-1.93	2.07
Total Trust	0	2.30	-5.36	4.10
Competence	0	0.72	-1.64	1.25
Benevolence	0	0.48	-1.18	1.01
Anthropomorphism	0	0.47	-0.96	0.86
Integrity	0	0.62	-1.47	1.09
Transparency	0	0.74	-1.43	0.57
Total Literacy	0	1.52	-2.98	3.55
Technical Proficiency	0	0.70	-1.23	1.68
Communication Proficiency	0	0.32	-0.72	0.60
Creative Application	0	0.62	-1.35	1.27
Total Security	0	1.37	-2.95	2.59
Privacy	0	0.78	-1.88	1.32
Security	0	0.65	-1.28	1.27

Table A1. Descriptive statistics for post-test factor scores

Appendix A2. Scale validation

Factor	Item	Std. Loading	Std. Error	z-value	p-value	95% CI Lower	95% CI Upper
OP	skills_operational_1	0.549	0.126	4.359	0	0.302	0.797
OP	skills_operational_2	0.641	0.118	5.451	0	0.410	0.871
OP	skills_operational_3	0.633	0.118	5.357	0	0.401	0.865
INF	skills_informational_2	1.000	0.000	NA	NA	1.000	1.000
PAR	skills_participation_1	1.000	0.000	NA	NA	1.000	1.000

Table A2-1. Standardized factor loadings for the internet skills scale validation

Chi_square	df	p-value	CFI	TLI	RMSEA	SRMR
4.341	4	0.362	0.99	0.975	0.04	0.044

Table A2-2. Model fit indices for the internet skills scale validation

Factor	Item	Std. Loading	Std. Error	z-value	p-value	95% CI Lower	95% CI Upper
com	trust_competence1	0.879	0.069	12.791	0.000	0.744	1.013
com	trust_competence2	0.686	0.088	7.784	0.000	0.513	0.859
com	trust_competence3	0.430	0.123	3.493	0.000	0.189	0.671
bene	trust_benevolence1	0.556	0.102	5.479	0.000	0.357	0.755
bene	trust_benevolence2	0.338	0.112	3.008	0.003	0.118	0.558
bene	trust_benevolence3	0.650	0.093	6.972	0.000	0.467	0.833
anthro	trust_anthropomorphism1	0.338	0.135	2.508	0.012	0.074	0.601
anthro	trust_anthropomorphism3	0.671	0.141	4.773	0.000	0.395	0.946
inte	trust_integrity1	0.778	0.109	7.112	0.000	0.563	0.992
inte	trust_integrity2	0.581	0.115	5.063	0.000	0.356	0.806
trans	trust_transparency3	1.000	0.000	NA	NA	1.000	1.000

Table A2-3. Standardized factor loadings for the trust scale validation

Chi_square	df	p-value	CFI	TLI	RMSEA	SRMR
51.954	35	0.032	0.897	0.839	0.095	0.091

Table A2-4. Model fit Iidices for the trust scale validation

Factor	Item	Std. Loading	Std. Error	z-value	p-value	95% CI Lower	95% CI Upper
tp	tp1	0.736	0.112	6.587	0.000	0.517	0.955
tp	tp2	0.553	0.124	4.440	0.000	0.309	0.796
tp	tp4	0.637	0.117	5.442	0.000	0.408	0.866
cp	cp1	0.370	0.131	2.825	0.005	0.113	0.627
cp	cp3	0.664	0.095	6.992	0.000	0.478	0.850
cp	cp4	0.744	0.087	8.590	0.000	0.574	0.913
ca	ca1	0.718	0.082	8.761	0.000	0.557	0.879
ca	ca2	0.761	0.076	9.989	0.000	0.611	0.910
ca	ca3	0.532	0.110	4.848	0.000	0.317	0.748

Table A2-5. Standardized factor loadings for the AI literacy scale validation

Chi_square	df	p-value	CFI	TLI	RMSEA	SRMR
40.795	24	0.018	0.874	0.811	0.114	0.085

Table A2-6. Model fit indices for the AI literacy scale validation

Factor	Item	Std. Loading	Std. Error	z-value	p-value	95% CI Lower	95% CI Upper
privacy	security_willing	0.592	0.118	5.034	0	0.361	0.822
privacy	security_breach	0.896	0.120	7.439	0	0.660	1.132
security	security_rumor	0.760	0.093	8.132	0	0.577	0.943
security	security_fairness	0.511	0.121	4.205	0	0.273	0.749
security	security_bias	0.720	0.097	7.445	0	0.530	0.910

Table A2-7. Standardized factor loadings for the privacy and security scale validation

Chi_square	df	p-value	CFI	TLI	RMSEA	SRMR
6.517	4	0.164	0.962	0.904	0.108	0.05

Table A2-8. Model fit indices for the privacy & security scale validation

Appendix A3. In-task regressions (main effects)

	trust user	interaction willing	share willing	trust platform	use platform browse	use platform post
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
B_vs_A	0.699**	0.885**	0.538	0.442	0.372	0.410
	(0.205)	(0.261)	(0.270)	(0.202)	(0.186)	(0.256)
C_vs_A	0.485*	0.612*	0.419	0.025	0.206	0.576*
	(0.190)	(0.244)	(0.273)	(0.217)	(0.212)	(0.232)
Constant	3.158***	3.079***	3.105***	3.737***	3.842***	3.447***
	(0.128)	(0.166)	(0.195)	(0.140)	(0.144)	(0.180)
Observations	108	108	108	108	108	108
R^2	0.104	0.101	0.038	0.040	0.027	0.060
Adjusted R^2	0.086	0.084	0.019	0.022	0.009	0.042
F Statistic (df = 2; 105)	6.064**	5.921**	2.060	2.203	1.473	3.344*
Note:	*p<0.05; **p<0.01; ***p<0.001					

Table A3-1. OLS baseline regression results for in-task main effects in the bots comments condition

	trust user	interaction willing	share willing	trust platform	use platform browse	use platform post
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
B_vs_A	0.654**	0.836**	0.370	0.433	0.323	0.413
	(0.223)	(0.277)	(0.281)	(0.232)	(0.199)	(0.258)
C_vs_A	0.354	0.380	-0.055	-0.358	-0.140	0.207
	(0.246)	(0.285)	(0.301)	(0.248)	(0.253)	(0.260)
gender	-0.056	0.111	0.122	-0.094	0.096	-0.160
	(0.173)	(0.251)	(0.229)	(0.203)	(0.173)	(0.242)
urban_rural	0.164	0.258	0.294	0.300	0.212	0.200
	(0.180)	(0.262)	(0.227)	(0.224)	(0.191)	(0.242)
scores_adjusted	-0.001	-0.003	-0.004*	-0.003*	-0.002	-0.001
	(0.001)	(0.001)	(0.002)	(0.001)	(0.001)	(0.001)
internet_exp	-0.019	-0.042	-0.020	-0.114**	-0.058	-0.108*
	(0.039)	(0.043)	(0.046)	(0.036)	(0.036)	(0.039)
course_dir	0.096	-0.172	0.625	0.010	0.423	0.488
	(0.255)	(0.338)	(0.322)	(0.237)	(0.234)	(0.287)
Constant	3.917***	4.907***	5.167***	6.591***	5.035***	5.082***
	(1.062)	(1.029)	(1.170)	(0.908)	(0.855)	(0.920)
Observations	108	108	108	108	108	108
R^2	0.119	0.151	0.166	0.172	0.120	0.145
Adjusted R^2	0.057	0.091	0.108	0.114	0.058	0.085
F Statistic (df = 7; 100)	1.928	2.537*	2.847**	2.957**	1.949	2.415*
Note:	*p<0.05; **p<0.01; ***p<0.001					

Table A3-2. OLS controlled regression results for in-task main effects in the bots comments condition

	trust content	trust material	like content	likelihood like	likelihood comment	likelihood share
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
B_vs_A	-0.506*	-0.553*	0.098	-0.038	0.282	-0.179
	(0.235)	(0.272)	(0.243)	(0.294)	(0.287)	(0.303)
C_vs_A	-0.113	-0.124	0.217	0.034	-0.147	-0.167
	(0.212)	(0.241)	(0.213)	(0.269)	(0.238)	(0.295)
Constant	3.684***	3.553***	3.474***	3.395***	2.789***	3.000***
	(0.126)	(0.163)	(0.154)	(0.183)	(0.174)	(0.210)
Observations	108	108	108	108	108	108
R^2	0.041	0.039	0.010	0.001	0.024	0.004
Adjusted R^2	0.023	0.021	-0.009	-0.018	0.005	-0.015
F Statistic (df = 2; 105)	2.245	2.153	0.509	0.029	1.274	0.222
Note:	*p<0.05; **p<0.01; ***p<0.001					

Table A3-3. OLS baseline regression results for in-task main effects in the news-image condition

	trust content	trust material	like content	likelihood like	likelihood comment	likelihood share
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
B_vs_A	-0.545*	-0.726*	0.094	-0.111	0.196	-0.304
	(0.266)	(0.289)	(0.242)	(0.316)	(0.278)	(0.301)
C_vs_A	-0.264	-0.237	-0.005	-0.280	-0.408	-0.769*
	(0.253)	(0.280)	(0.281)	(0.322)	(0.277)	(0.308)
gender	-0.033	0.217	-0.217	-0.360	-0.129	-0.284
	(0.224)	(0.235)	(0.214)	(0.261)	(0.245)	(0.279)
urban_rural	0.186	0.461	0.073	0.396	0.328	0.344
	(0.236)	(0.245)	(0.211)	(0.266)	(0.255)	(0.279)
scores_adjusted	-0.003*	-0.003	-0.002	-0.005*	-0.003	-0.007***
	(0.001)	(0.001)	(0.002)	(0.002)	(0.002)	(0.002)
internet_exp	-0.032	0.047	-0.066	-0.084	-0.044	-0.105*
	(0.049)	(0.044)	(0.043)	(0.051)	(0.049)	(0.056)
course_dir	-0.314	-0.256	0.096	-0.185	0.078	0.293
	(0.331)	(0.341)	(0.305)	(0.385)	(0.324)	(0.360)
Constant	5.781***	4.228**	5.610***	6.952***	4.693***	8.014***
	(0.936)	(1.029)	(1.469)	(1.264)	(1.386)	(1.390)
Observations	108	108	108	108	108	108
R^{2}	0.097	0.125	0.053	0.102	0.070	0.163
Adjusted R^{2}	0.034	0.064	-0.013	0.039	0.005	0.105
F Statistic (df = 7; 100)	1.539	2.040	0.800	1.614	1.079	2.788*
Note:	*p<0.05; **p<0.01; ***p<0.001					

Table A3-4. OLS controlled regression results for in-task main effects in the news-image condition

	trust user	interaction willing	share willing	trust platform	use platform browse	use platform post
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
A_vs_B	-0.699**	-0.885**	-0.538	-0.442	-0.372	-0.410
	(0.205)	(0.261)	(0.270)	(0.202)	(0.186)	(0.256)
C_vs_B	-0.214	-0.274	-0.119	-0.417	-0.167	0.167
	(0.212)	(0.270)	(0.267)	(0.221)	(0.196)	(0.234)
Constant	3.857***	3.964***	3.643***	4.179***	4.214***	3.857***
	(0.160)	(0.202)	(0.186)	(0.145)	(0.119)	(0.183)
Observations	108	108	108	108	108	108
R^{2}	0.104	0.101	0.038	0.040	0.027	0.060
Adjusted R^{2}	0.086	0.084	0.019	0.022	0.009	0.042
F Statistic (df = 2; 105)	6.064**	5.921**	2.060	2.203	1.473	3.344*
Note:	*p<0.05; **p<0.01; ***p<0.001					

Table A3-5. OLS baseline regression results for in-task main effects in the bots comments condition (vs. B)

	trust user	interaction willing	share willing	trust platform	use platform browse	use platform post
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
A_vs_B	-0.654**	-0.836**	-0.370	-0.433	-0.323	-0.413
	(0.223)	(0.277)	(0.281)	(0.232)	(0.199)	(0.258)
C_vs_B	-0.300	-0.456	-0.425	-0.791**	-0.462	-0.206
	(0.289)	(0.352)	(0.326)	(0.285)	(0.267)	(0.299)
gender	-0.056	0.111	0.122	-0.094	0.096	-0.160
	(0.173)	(0.251)	(0.229)	(0.203)	(0.173)	(0.242)
urban_rural	0.164	0.258	0.294	0.300	0.212	0.200
	(0.180)	(0.262)	(0.227)	(0.224)	(0.191)	(0.242)
scores_adjusted	-0.001	-0.003	-0.004*	-0.003*	-0.002	-0.001
	(0.001)	(0.001)	(0.002)	(0.001)	(0.001)	(0.001)
internet_exp	-0.019	-0.042	-0.020	-0.114**	-0.058	-0.108*
	(0.039)	(0.043)	(0.046)	(0.036)	(0.036)	(0.039)
course_dir	0.096	-0.172	0.625	0.010	0.423	0.488
	(0.255)	(0.338)	(0.322)	(0.237)	(0.234)	(0.287)
Constant	4.571***	5.743***	5.537***	7.024***	5.358***	5.496***
	(1.081)	(1.139)	(1.226)	(0.997)	(0.899)	(0.981)
Observations	108	108	108	108	108	108
R^2	0.119	0.151	0.166	0.172	0.120	0.145
Adjusted R^2	0.057	0.091	0.108	0.114	0.058	0.085
F Statistic (df = 7; 100)	1.928	2.537*	2.847**	2.957**	1.949	2.415*
Note:	*p<0.05; **p<0.01; ***p<0.001					

Table A3-6. OLS controlled regression results for in-task main effects in the bots comments condition (vs. B)

	trust content	trust material	like content	likelihood like	likelihood comment	likelihood share
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
A_vs_B	0.506*	0.553*	-0.098	0.038	-0.282	0.179
	(0.235)	(0.272)	(0.243)	(0.294)	(0.287)	(0.303)
C_vs_B	0.393	0.429	0.119	0.071	-0.429	0.012
	(0.262)	(0.281)	(0.238)	(0.303)	(0.281)	(0.301)
Constant	3.179***	3.000***	3.571***	3.357***	3.071***	2.821***
	(0.199)	(0.217)	(0.188)	(0.230)	(0.229)	(0.218)
Observations	108	108	108	108	108	108
R^2	0.041	0.039	0.010	0.001	0.024	0.004
Adjusted R^2	0.023	0.021	-0.009	-0.018	0.005	-0.015
F Statistic (df = 2; 105)	2.245	2.153	0.509	0.029	1.274	0.222
Note:	*p<0.05; **p<0.01; ***p<0.001					

Table A3-7. OLS baseline regression results for in-task main effects in the news-image condition (vs. B)

	trust content	trust material	like content	likelihood like	likelihood comment	likelihood share
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
A_vs_B	0.545*	0.726*	-0.094	0.111	-0.196	0.304
	(0.266)	(0.289)	(0.242)	(0.316)	(0.278)	(0.301)
C_vs_B	0.281	0.489	-0.100	-0.168	-0.605	-0.465
	(0.292)	(0.324)	(0.279)	(0.362)	(0.326)	(0.300)
gender	-0.033	0.217	-0.217	-0.360	-0.129	-0.284
	(0.224)	(0.235)	(0.214)	(0.261)	(0.245)	(0.279)
urban_rural	0.186	0.461	0.073	0.396	0.328	0.344
	(0.236)	(0.245)	(0.211)	(0.266)	(0.255)	(0.279)
scores_adjusted	-0.003*	-0.003	-0.002	-0.005*	-0.003	-0.007***
	(0.001)	(0.001)	(0.002)	(0.002)	(0.002)	(0.002)
internet_exp	-0.032	0.047	-0.066	-0.084	-0.044	-0.105*
	(0.049)	(0.044)	(0.043)	(0.051)	(0.049)	(0.056)
course_dir	-0.314	-0.256	0.096	-0.185	0.078	0.293
	(0.331)	(0.341)	(0.305)	(0.385)	(0.324)	(0.360)
Constant	5.236***	3.501**	5.704***	6.841***	4.889***	7.710***
	(1.026)	(1.098)	(1.422)	(1.309)	(1.411)	(1.397)
Observations	108	108	108	108	108	108
R^2	0.097	0.125	0.053	0.102	0.070	0.163
Adjusted R^2	0.034	0.064	-0.013	0.039	0.005	0.105
F Statistic (df = 7; 100)	1.539	2.040	0.800	1.614	1.079	2.788*
Note:	*p<0.05; **p<0.01; ***p<0.001					

Table A3-8. OLS controlled regression results for in-task main effects in the news-image condition (vs. B)

Appendix A4. Post-test outcome models.

	TotalTrust	competence	benevolence	anthropomorphism	integrity	transparency	trust2humanonl	trust2humanthe	trust2aionly	TotalLiteracy	tp	cp	ca
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8	Model 9	Model 10	Model 11	Model 12	Model 13
B_vs_A	–0.570	0.045	–0.075	–0.017	–0.140	–0.383	0.673*	–0.402	–0.079	–0.042	0.052	–0.036	-0.058
	(0.827)	(0.259)	(0.178)	(0.149)	(0.220)	(0.259)	(0.337)	(0.310)	(0.271)	(0.454)	(0.207)	(0.106)	(0.197)
C_vs_A	–0.159	–0.020	–0.043	–0.035	–0.058	–0.003	–0.065	0.003	–0.198	1.045*	0.632**	0.116	0.297
	(0.736)	(0.234)	(0.157)	(0.157)	(0.197)	(0.229)	(0.298)	(0.263)	(0.301)	(0.479)	(0.209)	(0.104)	(0.198)
Constant	0.210	–0.004	0.036	0.018	0.059	0.100	2.684***	3.474***	2.579***	–0.395	–0.259	–0.036	-0.100
	(0.538)	(0.174)	(0.125)	(0.103)	(0.137)	(0.160)	(0.230)	(0.193)	(0.234)	(0.319)	(0.143)	(0.069)	(0.131)
Observations	54	54	54	54	54	54	54	54	54	54	54	54	54
R^2	0.010	0.001	0.004	0.001	0.008	0.052	0.106	0.044	0.011	0.119	0.187	0.041	0.066
Adjusted R^2	–0.029	–0.038	–0.035	–0.038	–0.031	0.015	0.071	0.006	–0.028	0.085	0.155	0.004	0.030
F Statistic (df = 2; 51)	0.250	0.034	0.100	0.027	0.199	1.395	3.011	1.167	0.273	3.448*	5.864**	1.103	1.808
Note:													*p<0.05; **p<0.01; ***p<0.001

Table A4-1. OLS baseline regression results for post-test outcomes

	TotalTrust	competence	benevolence	anthropomorphism	integrity	transparency	trust2humanonl	trust2humanthe	trust2aionly	TotalLiteracy	tp	cp	ca
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8	Model 9	Model 10	Model 11	Model 12	Model 13
B_vs_A	-0.516	0.172	-0.050	-0.049	-0.215	-0.373	0.843*	-0.476	-0.050	-0.129	0.050	-0.067	-0.112
	(0.959)	(0.304)	(0.189)	(0.163)	(0.247)	(0.287)	(0.314)	(0.350)	(0.319)	(0.463)	(0.229)	(0.104)	(0.196)
C_vs_A	-0.309	0.029	-0.121	-0.087	-0.206	0.077	0.221	0.057	-0.219	0.616	0.533*	-0.005	0.088
	(0.859)	(0.256)	(0.172)	(0.184)	(0.241)	(0.269)	(0.385)	(0.282)	(0.337)	(0.624)	(0.267)	(0.139)	(0.263)
gender	-0.371	0.071	-0.037	-0.146	-0.216	-0.044	0.053	0.008	0.019	-0.603	-0.111	-0.174	-0.318
	(0.783)	(0.221)	(0.164)	(0.144)	(0.219)	(0.215)	(0.314)	(0.231)	(0.309)	(0.432)	(0.183)	(0.098)	(0.185)
urban_rural	0.115	-0.192	-0.072	0.200	0.226	-0.047	-0.651*	-0.213	-0.182	0.258	0.057	0.072	0.129
	(0.850)	(0.251)	(0.167)	(0.150)	(0.210)	(0.263)	(0.313)	(0.240)	(0.279)	(0.556)	(0.244)	(0.115)	(0.226)
scores_adjusted	0.001	0.001	0.001	-0.001	-0.0004	0.001	0.001	0.001	-0.003	-0.001	-0.0002	-0.0004	-0.001
	(0.005)	(0.001)	(0.001)	(0.001)	(0.001)	(0.002)	(0.002)	(0.002)	(0.002)	(0.003)	(0.001)	(0.001)	(0.001)
internet_exp	-0.095	-0.046	-0.026	-0.019	-0.017	0.013	0.020	0.081	-0.008	-0.092	-0.031	-0.022	-0.039
	(0.114)	(0.036)	(0.023)	(0.022)	(0.034)	(0.039)	(0.054)	(0.037)	(0.052)	(0.098)	(0.041)	(0.020)	(0.040)
course_dir	0.371	-0.225	0.304	-0.049	0.371	-0.030	-0.542	0.471	-0.362	0.980	0.171	0.298*	0.512*
	(0.994)	(0.276)	(0.171)	(0.198)	(0.252)	(0.368)	(0.341)	(0.410)	(0.332)	(0.589)	(0.278)	(0.116)	(0.231)
Constant	0.726	0.027	0.016	0.541	0.472	-0.331	2.890	2.576	4.364**	1.708	0.232	0.543	0.934
	(3.650)	(1.150)	(0.697)	(0.786)	(0.987)	(1.148)	(1.757)	(1.271)	(1.684)	(2.078)	(0.787)	(0.538)	(0.971)
Observations	54	54	54	54	54	54	54	54	54	54	54	54	54
R^2	0.025	0.063	0.074	0.068	0.073	0.060	0.212	0.170	0.063	0.194	0.204	0.179	0.181

Adjusted R^2	–0.123	–0.080	–0.067	–0.074	–0.068	–0.083	0.092	0.044	–0.079	0.071	0.083	0.054	0.057
F Statistic (df = 7; 46)	0.168	0.439	0.522	0.480	0.515	0.420	1.766	1.347	0.445	1.577	1.685	1.431	1.456
Note:	$^{*}p<0.05$; $^{**}p<0.01$; $^{***}p<0.001$												

Table A4-2. OLS controlled regression results for post-test outcomes

	TotalTrust	competence	benevolence	anthropomorphism	integrity	transparency	trust2humanonl	trust2humanthe	trust2aionly	TotalLiteracy	tp	cp	ca
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8	Model 9	Model 10	Model 11	Model 12	Model 13
A_vs_B	0.570	–0.045	0.075	0.017	0.140	0.383	–0.673*	0.402	0.079	0.042	–0.052	0.036	0.058
	(0.827)	(0.259)	(0.178)	(0.149)	(0.220)	(0.259)	(0.337)	(0.310)	(0.271)	(0.454)	(0.207)	(0.106)	(0.197)
C_vs_B	0.411	–0.065	0.032	–0.019	0.082	0.381	–0.738*	0.405	–0.119	1.086*	0.580*	0.152	0.355
	(0.805)	(0.248)	(0.157)	(0.159)	(0.224)	(0.262)	(0.311)	(0.301)	(0.234)	(0.482)	(0.214)	(0.112)	(0.210)
Constant	–0.361	0.041	–0.039	0.001	–0.081	–0.283	3.357***	3.071***	2.500***	–0.437	–0.207	–0.072	–0.158
	(0.628)	(0.192)	(0.126)	(0.107)	(0.173)	(0.204)	(0.246)	(0.243)	(0.138)	(0.323)	(0.150)	(0.081)	(0.148)
Observations	54	54	54	54	54	54	54	54	54	54	54	54	54
R^2	0.010	0.001	0.004	0.001	0.008	0.052	0.106	0.044	0.011	0.119	0.187	0.041	0.066
Adjusted R^2	–0.029	–0.038	–0.035	–0.038	–0.031	0.015	0.071	0.006	–0.028	0.085	0.155	0.004	0.030
F Statistic (df = 2; 51)	0.250	0.034	0.100	0.027	0.199	1.395	3.011	1.167	0.273	3.448*	5.864**	1.103	1.808
Note:	$^{*}p<0.05$; $^{**}p<0.01$; $^{***}p<0.001$												

Table A4-3. OLS baseline regression results for post-test outcomes (vs. B)

	TotalTrust	competence	benevolence	anthropomorphism	integrity	transparency	trust2humanonl	trust2humanthe	trust2aionly	TotalLiteracy	tp	cp	ca
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8	Model 9	Model 10	Model 11	Model 12	Model 13
A_vs_B	0.516	–0.172	0.050	0.049	0.215	0.373	–0.843*	0.476	0.050	0.129	–0.050	0.067	0.112
	(0.959)	(0.304)	(0.189)	(0.163)	(0.247)	(0.287)	(0.314)	(0.350)	(0.319)	(0.463)	(0.229)	(0.104)	(0.196)
C_vs_B	0.207	–0.143	–0.071	–0.038	0.009	0.450	–0.622	0.533	–0.169	0.744	0.483	0.062	0.199
	(0.992)	(0.310)	(0.178)	(0.178)	(0.265)	(0.318)	(0.382)	(0.309)	(0.274)	(0.679)	(0.290)	(0.150)	(0.286)
gender	–0.371	0.071	–0.037	–0.146	–0.216	–0.044	0.053	0.008	0.019	–0.603	–0.111	–0.174	–0.318
	(0.783)	(0.221)	(0.164)	(0.144)	(0.219)	(0.215)	(0.314)	(0.231)	(0.309)	(0.432)	(0.183)	(0.098)	(0.185)
urban_rural	0.115	–0.192	–0.072	0.200	0.226	–0.047	–0.651*	–0.213	–0.182	0.258	0.057	0.072	0.129
	(0.850)	(0.251)	(0.167)	(0.150)	(0.210)	(0.263)	(0.313)	(0.240)	(0.279)	(0.556)	(0.244)	(0.115)	(0.226)
scores_adjusted	0.001	0.001	0.001	–0.001	–0.0004	0.001	0.001	0.001	–0.003	–0.001	–0.0002	–0.0004	–0.001
	(0.005)	(0.001)	(0.001)	(0.001)	(0.001)	(0.002)	(0.002)	(0.002)	(0.002)	(0.003)	(0.001)	(0.001)	(0.001)
internet_exp	–0.095	–0.046	–0.026	–0.019	–0.017	0.013	0.020	0.081	–0.008	–0.092	–0.031	–0.022	–0.039
	(0.114)	(0.036)	(0.023)	(0.022)	(0.034)	(0.039)	(0.054)	(0.037)	(0.052)	(0.098)	(0.041)	(0.020)	(0.040)
course_dir	0.371	–0.225	0.304	–0.049	0.371	–0.030	–0.542	0.471	–0.362	0.980	0.171	0.298*	0.512*
	(0.994)	(0.276)	(0.171)	(0.198)	(0.252)	(0.368)	(0.341)	(0.410)	(0.332)	(0.589)	(0.278)	(0.116)	(0.231)
Constant	0.210	0.199	–0.034	0.492	0.257	–0.704	3.733*	2.099	4.314**	1.580	0.283	0.475	0.822
	(3.879)	(1.227)	(0.728)	(0.772)	(1.037)	(1.254)	(1.719)	(1.420)	(1.630)	(2.070)	(0.778)	(0.536)	(0.968)
Observations	54	54	54	54	54	54	54	54	54	54	54	54	54
R^2	0.025	0.063	0.074	0.068	0.073	0.060	0.212	0.170	0.063	0.194	0.204	0.179	0.181

Adjusted R^2	–0.123	–0.080	–0.067	–0.074	–0.068	–0.083	0.092	0.044	–0.079	0.071	0.083	0.054	0.057
F Statistic (df = 7; 46)	0.168	0.439	0.522	0.480	0.515	0.420	1.766	1.347	0.445	1.577	1.685	1.431	1.456
Note:	$^{*}p<0.05$; $^{**}p<0.01$; $^{***}p<0.001$												

Table A4-4. OLS controlled regression results for post-test outcomes (vs. B)

Appendix A5. Heterogeneity models

	TotalTrust	competence	benevolence	anthropomorphism	integrity	transparency	trust2humanonl	trust2humanthe	trust2aionly	TotalLiteracy	tp	cp	ca
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8	Model 9	Model 10	Model 11	Model 12	Model 13
B_vs_A	6.050	1.638	0.264	1.533*	1.196	1.420	0.922	0.090	0.529	3.967*	1.514	0.813*	1.640*
	(2.199)	(0.805)	(0.598)	(0.401)	(0.730)	(0.737)	(1.080)	(1.353)	(1.009)	(1.447)	(0.825)	(0.353)	(0.633)
C_vs_A	1.308	0.455	–0.344	0.557	0.087	0.553	0.627	–0.394	0.072	2.143	0.698	0.489	0.955
	(2.432)	(0.784)	(0.575)	(0.467)	(0.695)	(0.710)	(1.033)	(1.046)	(0.997)	(1.839)	(0.864)	(0.365)	(0.714)
internet_using_time	0.662	0.093	0.002	0.232	0.191	0.144	0.227	–0.139	0.278	0.744	0.232	0.177*	0.335*
	(0.542)	(0.197)	(0.139)	(0.101)	(0.156)	(0.165)	(0.258)	(0.313)	(0.260)	(0.377)	(0.206)	(0.067)	(0.133)
gender	–0.349	0.047	–0.004	–0.150	–0.186	–0.056	0.077	0.024	0.078	–0.544	–0.048	–0.179	–0.317
	(0.853)	(0.239)	(0.179)	(0.145)	(0.231)	(0.235)	(0.324)	(0.256)	(0.306)	(0.513)	(0.228)	(0.107)	(0.207)
urban_rural	0.076	–0.218	–0.084	0.213	0.232	–0.068	–0.584	–0.277	–0.119	0.335	0.074	0.094	0.167
	(0.848)	(0.262)	(0.175)	(0.146)	(0.216)	(0.264)	(0.320)	(0.249)	(0.273)	(0.548)	(0.237)	(0.117)	(0.227)
scores_adjusted	0.002	0.001	0.001	–0.0003	0.00004	0.001	0.001	0.0004	–0.002	0.0002	0.001	–0.0002	–0.0002
	(0.005)	(0.002)	(0.001)	(0.001)	(0.001)	(0.002)	(0.002)	(0.002)	(0.002)	(0.003)	(0.001)	(0.001)	(0.001)
internet_exp	–0.116	–0.055	–0.024	–0.024	–0.019	0.005	0.026	0.077	–0.0001	–0.095	–0.029	–0.024	–0.043
	(0.123)	(0.037)	(0.025)	(0.021)	(0.033)	(0.042)	(0.053)	(0.039)	(0.048)	(0.093)	(0.039)	(0.020)	(0.038)
course_dir	0.212	–0.287	0.301	–0.071	0.360	–0.091	–0.474	0.409	–0.292	0.991	0.182	0.298*	0.511*
	(1.008)	(0.276)	(0.179)	(0.201)	(0.260)	(0.379)	(0.353)	(0.413)	(0.376)	(0.584)	(0.285)	(0.112)	(0.224)
B_vs_A:internet_using_time	–1.868*	–0.412	–0.086	–0.457**	–0.406	–0.507	–0.042	–0.143	–0.184	–1.195*	–0.424	–0.258*	–0.513*
	(0.615)	(0.223)	(0.162)	(0.112)	(0.195)	(0.210)	(0.307)	(0.407)	(0.285)	(0.425)	(0.237)	(0.093)	(0.172)

C_vs_A:internet_using_time	–0.481	–0.125	0.066	–0.193	–0.089	–0.140	–0.126	0.138	–0.093	–0.463	–0.053	–0.149	–0.261
	(0.667)	(0.219)	(0.156)	(0.127)	(0.186)	(0.191)	(0.282)	(0.329)	(0.287)	(0.511)	(0.248)	(0.099)	(0.195)
Constant	–1.796	–0.082	–0.152	–0.370	–0.467	–0.725	1.610	3.258	2.683	–1.802	–1.112	–0.188	–0.501
	(4.168)	(1.371)	(0.879)	(0.854)	(1.178)	(1.462)	(1.894)	(1.602)	(1.738)	(2.648)	(1.124)	(0.602)	(1.121)
Observations	54	54	54	54	54	54	54	54	54	54	54	54	54
R^2	0.141	0.131	0.103	0.213	0.161	0.147	0.247	0.212	0.138	0.315	0.319	0.281	0.295
Adjusted R^2	–0.058	–0.071	–0.106	0.030	–0.035	–0.052	0.072	0.028	–0.062	0.156	0.161	0.114	0.131
F Statistic (df = 10; 43)	0.707	0.651	0.492	1.162	0.822	0.738	1.411	1.154	0.691	1.981	2.014	1.681	1.798
Note:	*p<0.05; **p<0.01; ***p<0.001												

Table A5-1. Interaction effects of internet usage on treatment effects-controlled model (vs. A)

	TotalTrust	competence	benevolence	anthropomorphism	integrity	transparency	trust2humanonl	trust2humanthe	trust2aionly	TotalLiteracy	tp	cp	ca
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8	Model 9	Model 10	Model 11	Model 12	Model 13
A_vs_B	-6.050	-1.638	-0.264	-1.533*	-1.196	-1.420	-0.922	-0.090	-0.529	-3.967*	-1.514	-0.813*	-1.640*
	(2.199)	(0.805)	(0.598)	(0.401)	(0.730)	(0.737)	(1.080)	(1.353)	(1.009)	(1.447)	(0.825)	(0.353)	(0.633)
C_vs_B	-4.742	-1.183	-0.608	-0.976	-1.108	-0.866	-0.294	-0.484	-0.456	-1.824	-0.816	-0.323	-0.685
	(2.013)	(0.616)	(0.450)	(0.355)	(0.631)	(0.738)	(0.779)	(0.964)	(0.699)	(1.568)	(0.785)	(0.370)	(0.683)
internet_using_time	-1.206*	-0.319	-0.085	-0.224	-0.215	-0.362	0.185	-0.282	0.094	-0.451	-0.192	-0.082	-0.178
	(0.294)	(0.112)	(0.084)	(0.051)	(0.123)	(0.146)	(0.158)	(0.248)	(0.121)	(0.181)	(0.116)	(0.063)	(0.104)
gender	-0.349	0.047	-0.004	-0.150	-0.186	-0.056	0.077	0.024	0.078	-0.544	-0.048	-0.179	-0.317
	(0.853)	(0.239)	(0.179)	(0.145)	(0.231)	(0.235)	(0.324)	(0.256)	(0.306)	(0.513)	(0.228)	(0.107)	(0.207)
urban_rural	0.076	-0.218	-0.084	0.213	0.232	-0.068	-0.584	-0.277	-0.119	0.335	0.074	0.094	0.167
	(0.848)	(0.262)	(0.175)	(0.146)	(0.216)	(0.264)	(0.320)	(0.249)	(0.273)	(0.548)	(0.237)	(0.117)	(0.227)
scores_adjusted	0.002	0.001	0.001	-0.0003	0.00004	0.001	0.001	0.0004	-0.002	0.0002	0.001	-0.0002	-0.0002
	(0.005)	(0.002)	(0.001)	(0.001)	(0.001)	(0.002)	(0.002)	(0.002)	(0.002)	(0.003)	(0.001)	(0.001)	(0.001)
internet_exp	-0.116	-0.055	-0.024	-0.024	-0.019	0.005	0.026	0.077	-0.0001	-0.095	-0.029	-0.024	-0.043
	(0.123)	(0.037)	(0.025)	(0.021)	(0.033)	(0.042)	(0.053)	(0.039)	(0.048)	(0.093)	(0.039)	(0.020)	(0.038)
course_dir	0.212	-0.287	0.301	-0.071	0.360	-0.091	-0.474	0.409	-0.292	0.991	0.182	0.298*	0.511*
	(1.008)	(0.276)	(0.179)	(0.201)	(0.260)	(0.379)	(0.353)	(0.413)	(0.376)	(0.584)	(0.285)	(0.112)	(0.224)
A_vs_B:internet_using_time	1.868*	0.412	0.086	0.457**	0.406	0.507	0.042	0.143	0.184	1.195*	0.424	0.258*	0.513*
	(0.615)	(0.223)	(0.162)	(0.112)	(0.195)	(0.210)	(0.307)	(0.407)	(0.285)	(0.425)	(0.237)	(0.093)	(0.172)

C_vs_B:internet_using_time	1.387*	0.288	0.152	0.264*	0.316	0.366	–0.084	0.281	0.091	0.732	0.371*	0.110	0.252
	(0.453)	(0.139)	(0.104)	(0.084)	(0.152)	(0.175)	(0.182)	(0.265)	(0.165)	(0.393)	(0.188)	(0.093)	(0.172)
Constant	4.254	1.555	0.112	1.163	0.729	0.694	2.532	3.349	3.211	2.165	0.401	0.625	1.139
	(4.277)	(1.416)	(0.864)	(0.847)	(1.232)	(1.480)	(1.764)	(1.502)	(1.566)	(2.352)	(1.007)	(0.605)	(1.089)
Observations	54	54	54	54	54	54	54	54	54	54	54	54	54
R^2	0.141	0.131	0.103	0.213	0.161	0.147	0.247	0.212	0.138	0.315	0.319	0.281	0.295
Adjusted R^2	–0.058	–0.071	–0.106	0.030	–0.035	–0.052	0.072	0.028	–0.062	0.156	0.161	0.114	0.131
F Statistic (df = 10; 43)	0.707	0.651	0.492	1.162	0.822	0.738	1.411	1.154	0.691	1.981	2.014	1.681	1.798
Note:	*p<0.05; **p<0.01; ***p<0.001												

Table A5-2. Interaction effects of internet usage on treatment effects-controlled model (vs. B)

Appendix B. Interview theme analysis

Core Theme	Sub-theme	Initial Code	Data Example	Data Source
Perception and Cognitive Attitudes towards Content	AI Content Recognition Ability and Criteria	Can recognize AI-generated news	'I always felt the first three were a bit strange, the first three were the fake news ones.'	Participant A003
		Cannot recognize AI-generated images	'I didn't think those pictures looked real.'	Participant A001
		Cannot recognize AI-generated comments	'I felt those comments were pretty real, not like a robot.'	Participant A015
		Sense of unreality/inauthenticity	'I felt that the comment below wasn't the response he wanted in his heart, so it wasn't very real.'	Participant A010
	AI Content Trust Levels	Low trust in AIGC	'However, there are many feasibility issues with robots on social platforms, so my trust in some robots on social platforms is relatively low.'	Participant C049
		Moderate trust in AIGC	'If it's a 5-point scale, I'd give it 3 points.'	Participant A001
		High trust in AIGC	'For AI, 85%.'	Participant B022
	Influencing Factors of AI Trust Evaluation	Reasons for Trust	'I think this AI is well-trained, it can already simulate a real state. Because before, it might have been fed a certain set of values, and could only generate the most top-level content. But now it can generate things that blur the line, it's more like a regular person.'	Participant C043
		Reasons for Doubt	'I think the edges of the image are not well-processed. For example, I can feel there's a thin film between the mother and daughter, I don't know if it's real or if the image processing didn't make it very real, and the edges are a bit blurred, so it doesn't look like a real photo.'	Participant C059

	Acceptance of Different Types of AI Content	Accepts AI bots on social media	‘It's fine, I don't care if it's a real person or a robot, as long as the response makes me happy.’	Participant A006
		Does not accept AIGC images and news on social media	‘In terms of news, some creations might deviate from the original real scene. The original scene isn't real, and it's easy to mislead people.’	Participant C055
		Acceptance of AIGC news depends on its nature	‘It's best if robots don't appear in the news, but in all other non-serious contexts, like for entertainment, it's fine.’	Participant B026
	Interviewees' Critical Thinking and Strategies	Multi-source verification behavior	‘In that case, I would usually take a multi-tool approach to synthesize and analyze the content.’	Participant C048
		Self-reflection	‘But it might also cause me to rely too much on AI, and I might not want to think about many difficult problems after I get home.’	Participant C057
Interaction Behavior and Willingness	Interaction Willingness and Tendencies	Willing to interact with AIGC content on social media	"It's acceptable, it's acceptable."	Participant B021
		Unwilling to interact with AI bots	‘Generally, if it doesn't look for me, I won't look for it.’	Participant A019
		Preference for human interaction over AI bots	‘I'm more inclined to interact with a real person.’	Participant B038
	Interaction Motivations and Emotional Experience	Driven by curiosity	‘Because I'm more curious.’	Participant A003
		Emotional resonance/Psychological pleasure	‘Because I feel its replies are all very positive and uplifting, so I'm willing to reply.’	Participant A020
		Lack of emotional resonance with AI interaction	‘Because I might know that it's not a real person and can't stand in a person's shoes or interact with me based on a person's experience. I just feel that this might not be very necessary.’	Participant C051
Necessity and Preferences for AI Disclosure	The Necessity of Disclosure	AIGC content needs disclosure	‘For news, I think it should have a label.’	Participant B022
		AI bots do not need disclosure	‘There's no special need to differentiate them.’	Participant A013

		AI bot disclosure depends on the situation	‘In some areas, labeling AI can increase trust between AI and people, but in some ways, in personal conversations, if the AI label is shown, it may affect some communication and interaction between AI and people.’	Participant B025
	Preferred Disclosure Forms and Locations	Text prompt	‘A text prompt, I think that's pretty good.’	Participant A008
		Visual effects	‘For example, its avatar is a robot.’	Participant A017
	The Impact of Disclosure on User Behavior	Impacts trust	‘If you show that it's AI, people might verify the authenticity of the content, especially for news. If there's no label, people might be more willing to believe what they see on TV news and won't go and verify if it's true or false.’	Participant A008
		Loss of emotional value	‘If it's disclosed, it won't provide people with that sense of identity and belonging.’	Participant A004
AI Usage and Platform Responsibility	Current Usage Scenarios and Frequency	Usage Scenario: Learning/Homework	‘Using this AI to help me polish my essays.’	Participant B028
		Usage Scenario: Emotional/Entertainment	‘Using AI to generate what I want... chatting... just playing with it.’	Participant A018
		Usage Frequency: Infrequent	‘I've used it, but I don't use it very often.’	Participant B030
	Platform Responsibility and Ethical Awareness	Platform needs to fact-check AIGC content	‘If it fabricates some facts, I think the platform needs to be responsible.’	Participant A001
		Platform needs to ensure AI bots are friendly and neutral	‘Not those that are malicious.’	Participant B032
		Platform needs to provide feedback channels for AIGC	‘There should be an option for feedback... using 'report' isn't quite right.’	Participant B023

Table B-1. Interview theme analysis with codes and data examples