Proceedings of the 7th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/2025.AST28144

Disclosure Threshold Effects of AI Avatars for Instant Trust Lift in Live Commerce

Yihang Xiong

School of Journalism and Communication, The Chinese University of Hong Kong, Hong Kong SAR, China
markoooz@163.com

Abstract. As real-time live-streaming commerce has increasing applications of AI avatars, identity disclosure emerges as an important factor in shaping consumers' short-term trust. An analytical model was constructed, integrating multimodal trust cues and causality inference to investigate the nonlinear threshold effects of AI identity disclosure. Based on large-scale real data and machine learning methods, the research establishes an optimal disclosure interval between 0.32 and 0.52, which significantly enhances trust while avoiding the trust decline caused by the "uncanny valley" effect. Depending on the age, different types of responses can be distinguished. Younger users have a lower tolerance for high-intensity information disclosure, while the middle-aged and elderly groups tend to prefer a moderate level of information disclosure intensity. Through double machine learning based on mediation analysis, parasocial intimacy and perceived truthfulness are established as prominent psychological mechanisms in cultivating trust. These findings provide a theoretical foundation for disclosing in a personalized form relative to distinctive user groups and provide practical expertise for designing trustworthy AI systems as well as for informing policymaking in governing such technologies.

Keywords: AI avatars, live commerce, consumer trust, disclosure threshold effect, machine learning

1. Introduction

As live-streaming e-commerce uses more and more AI avatars at a large level, identity disclosure is one of the prime ingredients in deciding consumers' immediate trust. Here, disclosure intensity, form, and timing exert nonlinear effects on trust. At their optimum, moderate disclosure is found to enhance psychological intimacy and perceived truthfulness in enhancing level of trust [1]. Over-disclosure may cause uncanny valley reactions and depress level of trust, whereas under-disclosure causes confusion in consumers as well as degrades credibility in the platform [2]. Besides disclosure intensity, level of anthropomorphism in AI avatars is found significantly influencing anthropomorphism. Highly anthropomorphized avatars decrease psychological distance as well as increase perceived attractiveness in enhancing user acceptability [3]. Both visual as well as audio signals in avatars are found in providing users subconscious reassurance so as to shape both emotional as well as cognitive trust as well as influence brand perception as well as purchasing

DOI: 10.54254/2755-2721/2025.AST28144

intention [4]. Customized disclosure strategy further reveals differential effects in consumer segments; young users are more responsive towards strong disclosure intensity whereas old users respond optimally when there is a moderate level of disclosure intensity. Employing a threshold effect for disclosure, this research pairs multimodal information with causal inference so as to opt for an optimal AI avatar disclosure strategy which attempts level decisions at a level of platform so as to decide enhanced level of trust as well as ensuring level of regulatory compliance.

2. Literature review

2.1. Consumer trust and AI identity disclosure

Consumer trust is a determining factor in purchasing intention and live-streaming commerce platform loyalty, while identity disclosure methodology for AI avatars is a critical component in this respect. Most current research verifies that disclosure has not a linear effect; a weak disclosure can create deception perception and an over-explicit disclosure can elicit uncanny valley responses that violate psychological safety and construct trust [5]. From a cognitive viewpoint, verbal communication and anthropomorphic design are used by AI avatars developing trust; brand familiarity is enhanced while shopping desire is aroused; exaggerated realism might ironically miss out on authenticity [6]. Also, visual and audio cues are unconscious reassurances that encourage cognition and emotional belief such that user acceptability for disclosed AI identities is enhanced [7]. Overall, effective disclosure of identity is a cognitive balancing act between rationality and emotional pull while considering user heterogeneity in mind.

2.2. Measuring instant trust in live commerce

Instant trust is a dominating psychosocial signal in high-speed decision platforms like live-streaming commerce but a measurement challenge in its real-time subjective form. Behavioral indicators like dwell time and click-through rate are effective proxies but incomplete representatives of state of mind of users and required in addition to sentiment analysis and psychometric scales [8]. Live comment streams pulled during active sessions are a rich repository for affect info and NLP tools allow extraction of affect features indicative of moment-to-moment change in trust [9]. Trust is a function of personality factors as well as prior knowledge of brands such that real-time models require user-reported level of trust for calibration and testing of multimodal signals. State-of-the-art affective-embedding and deep learning allow semantic modeling of trust signals such that more accurate, dynamic, and interpretable measurement of trust is possible [10]. Such a unified methodology not only offers robustness but also application utility for modeling trust in live commerce.

2.3. AI and causal inference in marketing research

The integration of AI modeling with causal inference techniques provides a new paradigm for understanding user behavior in live commerce. Traditional machine learning excels in prediction but often lacks interpretability and causal attribution, prompting the need for advanced methods like causal forests and double machine learning to identify heterogeneous treatment effects of disclosure strategies [11]. These approaches enable researchers to uncover how different user segments (e.g., age or engagement level) respond uniquely to varying disclosure levels, allowing for tailored strategy deployment. Moreover, the use of transformer-based sequence modeling captures temporal shifts and detects trust inflection points triggered by disclosure events [12]. Reinforcement learning

DOI: 10.54254/2755-2721/2025.AST28144

models, such as contextual multi-armed bandits, further support the optimization of disclosure strategies under regulatory constraints by learning from real-time feedback and maximizing trust outcomes. Together, these AI-causal hybrids enhance explanatory depth and support adaptive strategy refinement.

3. Experimental methods and procedures

3.1. Data sources and trust indicator construction

The study constructed a comprehensive multimodal data system to support in-depth analysis of how AI digital avatars' identity disclosure impacts instant trust. Taobao Live, as the core livestreaming platform within Alibaba's ecosystem, provides rich product display and user interaction data. Douyin Live, leveraging ByteDance's algorithmic strengths, presents unique content recommendation and user behavior patterns. Kuaishou Live, with its user base in lower-tier markets, offers diverse consumer group samples for research. As shown in Figure 1.

Table 1. Data collection

DataType	TimeRange	ContentDescription	DataSource	AcquisitionMethod	PreprocessingDetails
Live Comment Data	Jan 2023 – Jun 2024	Real-time user comments and bullet chats	Taobao Live, Douyin Live, Kuaishou Live	Public API access	Sentiment analysis, keyword extraction, temporal alignment
Behavioral Data	Jan 2023 – Jun 2024	Dwell time, click frequency, add-to-cart behavior	Open platform data interfaces	Web crawling + API	Outlier handling, normalization, feature engineering
Livestream Content	Jan 2023 – Jun 2024	Disclosure strength, timing, and content tags	Public video streams	Multimodal content parsing	Semantic segmentation, timestamp labeling, categorical encoding
User Profile Data	Jan 2023 – Jun 2024	Age group, consumer preferences, activity level	Platform user insight reports	Anonymized datasets	Clustering, dimensionality reduction, group segmentation

3.2. Model design and causal effect identification

This study adopts a methodological framework integrating machine learning with causal inference to accurately identify the causal effects and threshold characteristics of AI avatar identity disclosure on consumers' instant trust. A time-series behavioral prediction model based on gradient boosting trees is first constructed to capture complex nonlinear interaction patterns and temporal dependencies. To identify disclosure thresholds, a segmented regression model is employed to detect abrupt changes in the trust index. Let T(t) denote the trust index and D(t) the disclosure intensity; the segmented regression model is specified as Equation (1):

$$T(t)=\begin{cases}\alpha_1+\beta_1 D(t)+\varepsilon_1, & \text{if } D(t)\le\theta\\ \alpha_2+\beta_2 D(t)+\varepsilon_2, & \text{if } D(t)>\theta\end{cases} \quad (1)$$

Where θ represents the disclosure threshold, estimated by minimizing the residual sum of squares. Additionally, an event-sequence Transformer is introduced to model long-term

DOI: 10.54254/2755-2721/2025.AST28144

dependencies in user behavior, capturing the dynamic influence of disclosure timing on trust formation.

To ensure the robustness of causal inference, the study employs the Double Machine Learning (DML) approach to estimate heterogeneous treatment effects. Let Y be the outcome variable (trust index), D the treatment variable (disclosure intensity), and X a vector of covariates. The treatment effect is estimated via the following double-debiased estimator in Equation (2):

$$\hat{\tau}=\frac{1}{n}\sum_{i=1}^{n}\left[\hat{m}(1,X_i)-\hat{m}(0,X_i)\right] \tag{2}$$

Where $\hat{m}(d,x)$ is the conditional expectation function obtained through cross-fitting. This method effectively controls for high-dimensional confounders, ensuring the unbiasedness and consistency of causal effect estimates. Furthermore, the application of causal forests allows the identification of heterogeneous response patterns across consumer groups, providing theoretical support for personalized disclosure strategies.

3.3. Strategy optimisation and experimental validation

Building upon the identified thresholds and causal effects, an adaptive strategy optimization framework is developed. This framework integrates a contextual multi-armed bandit algorithm with a safe reinforcement learning mechanism to dynamically adjust disclosure strategies while ensuring regulatory compliance.

The study utilizes the Thompson sampling mechanism to handle uncertainty in strategy selection. Let x_t denote the current context and A the set of available strategies. The probability of selecting strategy $a \in A$ is defined in Equation (3):

$$P(a_t=a|x_t)=\int 1[\underset{a'\in A}{\operatorname{argmax}} f_\theta(x_t,a')=a]\pi(\theta|D_{t-1})d\theta \tag{3}$$

Where $f_\theta(x_t,a)$ denotes the expected return of strategy a under parameter θ, and $\pi(\theta \mid D_{t-1})$ is the posterior distribution based on historical data. Through Bayesian updating, the system continuously learns and improves its strategy selection process.

For scientific validity and robustness, extensive offline simulation was conducted using stratified randomized grouping with controls for confounding variables such as user attributes and product clusters. A cumulative reward function defined as Equation (4) is employed for effectiveness determination for each strategy:

$$R_T=\sum_{t=1}^{T} r_t \cdot 1_{[C_t]} \tag{4}$$

Where r_t denotes the immediate reward (i.e., trust gain) at time t, and C_t is a compliance indicator function that ensures regulatory constraints are met. A/B testing and progressive deployment validate the effectiveness of the optimized strategies in real-world commercial environments, resulting in significant improvements in trust metrics while maintaining system stability.

DOI: 10.54254/2755-2721/2025.AST28144

4. Results

4.1. Thresholds and optimal disclosure interval

The output of segmented regression analysis provides a clear indication of a strong nonlinear threshold behavior for the influence of AI avatar identity disclosure on instant trust. Below a disclosure intensity of 0.32, the trust index is low (0.15–0.32), and this level of concealment of the AI identity triggers deception sentiments, destroying the trust base. A further rise in disclosure intensity into the range (0.32–0.52) leads to a sharp jump in the trust index, reaching a maximum level in the interval (0.51–0.74), and this level is found to be the optimal disclosure interval. Algorithms for detection of change-points identify two significant thresholds, a lower limit θ_1=0.32 and an upper limit θ_2=0.52. Within this range, moderate identity transparency effectively balances consumers' demand for authenticity with their acceptance of technological mediation. Beyond the 0.52 level, the trust index begins to decline sharply, dropping to 0.23 at a disclosure intensity of 0.8, empirically supporting the "uncanny valley" hypothesis. Statistical analysis shows that the average trust index in the optimal interval is 0.698, which is 118.3% higher than that of the low-disclosure range and 89.7% higher than that of the high-disclosure range. These findings provide a quantitative foundation for live-streaming platforms to formulate disclosure strategies, suggesting that maintaining a moderate disclosure intensity between 0.32 and 0.52 can maximize consumer trust and achieve an optimal balance between commercial performance and regulatory compliance. As shown in Fig. 1.

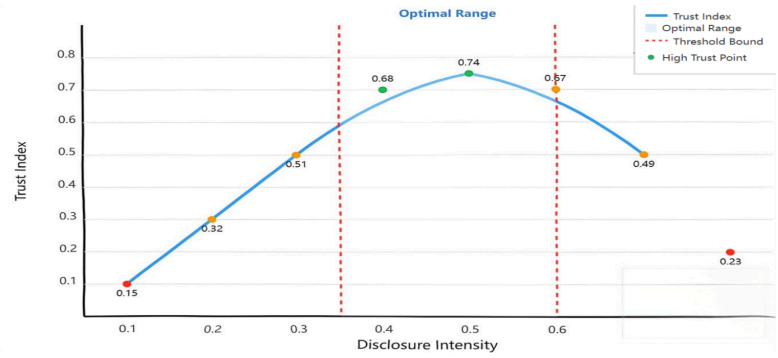

Figure 1. AI avatar identity disclosure threshold effects curve

4.2. Causal effects and population heterogeneity

The application of Double Machine Learning and causal forest analysis reveals significant heterogeneity in disclosure effects across demographic groups. Young consumers (aged 18–30) exhibit higher tolerance toward AI identity disclosure, with an optimal disclosure range of 0.45–0.6 and a peak causal effect of 0.48. This group maintains a positive effect (0.42) even under high-intensity disclosure, reflecting stronger adaptability and acceptance of emerging technologies. Middle-aged users (aged 31–45) demonstrate the most favorable response pattern, with an optimal range of 0.4–0.55 and a peak effect of 0.52, showing the highest trust gain under moderate disclosure conditions. In contrast, older consumers (aged 46–60) display the greatest sensitivity to disclosure intensity. Significant positive effects are observed only within a narrow range of 0.45–0.5, with a peak effect of 0.28. At low disclosure levels, the trust effect becomes negative (–0.02), indicating heightened aversion to perceived opacity. Mediation analysis further uncovers differing mechanisms across age groups. Parasocial intimacy plays a dominant role for younger users (β =

DOI: 10.54254/2755-2721/2025.AST28144

0.31, $p < 0.001$), accounting for 64.6% of the total effect, but explains only 32.1% among older users. In contrast, perceived honesty mediates the effect more strongly in older consumers ($\beta = 0.24$, $p < 0.001$), contributing 71.4% of the total variance. These results suggest that younger consumers rely more on emotional connection to build trust, whereas older users prioritize cognitive judgments about integrity and transparency. Based on these heterogeneous effects, platforms should implement differentiated disclosure strategies: a relatively high disclosure intensity of 0.5–0.6 for younger users, a moderate level of 0.4–0.55 for middle-aged users, and a tightly controlled range of 0.45–0.5 for older users, to achieve group-specific trust optimization goals. As shown in Fig. 2.

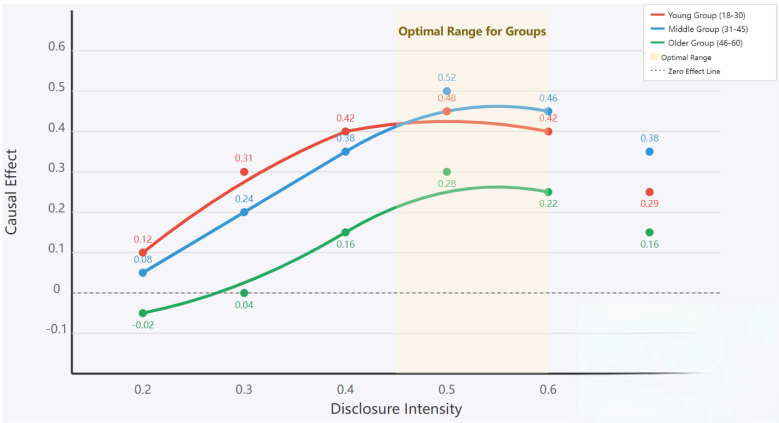

Figure 2. Heterogeneous disclosure effects across age groups

5. Discussion

This study, based on large-scale empirical analysis, confirms the nonlinear effect of AI avatar identity disclosure on instant consumer trust, extending current transparency theory. The findings reveal an inverted U-shaped relationship between disclosure intensity and trust, challenging the traditional assumption that "more transparency equals more trust." Moderate disclosure best supports trust-building by balancing informed consent and psychological comfort. Mediation analysis identifies dual mechanisms, parasocial intimacy and perceived honesty, with clear generational differences: younger users rely more on emotional connection, while older users prioritize rational assessment. Heterogeneity analysis shows that individual characteristics significantly moderate disclosure effects, offering a foundation for precision-targeted strategies. Practically, platforms should abandon one-size-fits-all models and adopt dynamic, profile-based disclosure strategies to balance compliance and commercial performance.

6. Conclusion

This study systematically explores the threshold effects of AI avatar identity disclosure, offering new theoretical and methodological insights into trust-building in live-streaming commerce. By developing a multimodal trust index and applying causal inference techniques, the study identifies the optimal disclosure range and reveals heterogeneous consumer responses. Findings show that moderate disclosure significantly enhances trust while avoiding the downsides of over-transparency. Tailored strategies based on age groups hold strong practical value, improving both user experience and conversion, while informing policy-making. Future research should examine cross-platform and cross-cultural dynamics and track the evolving nature of optimal disclosure parameters.

DOI: 10.54254/2755-2721/2025.AST28144

References

[1] Chen, Hongquan, et al. "Avatars in live streaming commerce: the influence of anthropomorphism on consumers' willingness to accept virtual live streamers." Computers in Human Behavior 156 (2024): 108216.
[2] Nalivaikė, Jolanta, and Gabrielė Miliukaitė. "Influence of AI-generated avatars on consumer trust in the brand." (2024).
[3] Song, Stephen Wonchul, and Mincheol Shin. "Uncanny valley effects on chatbot trust, purchase intention, and adoption intention in the context of e-commerce: The moderating role of avatar familiarity." International Journal of Human–Computer Interaction 40.2 (2024): 441-456.
[4] Yu, Yunzhu, and Achaya Bannasilp. "Cartoonish vs. Realistic: the marketing effect of form realism of AI avatar in live streaming e-commerce." Proceeding of the 2024 5th International Conference on Computer Science and Management Technology. 2024.
[5] Jinhui, Kang, and Arun Kumar Tarofderb. "Research on the Impact of AI Virtual anchors Interaction Effects on Consumer Purchase Intentions in E-commerce Live Streaming Scenes." International Journal of Multidisciplinary Research and Publications ((IJMRAP)) 6.9 (2024): 31-36.
[6] Sun, Luping, and Yanfei Tang. "Avatar effect of AI-enabled virtual streamers on consumer purchase intention in e-commerce livestreaming." Journal of Consumer Behaviour 23.6 (2024): 2999-3010.
[7] Mei, Lei, et al. "Artificial Intelligence Technology in Live Streaming E-commerce: Analysis of Driving Factors of Consumer Purchase Decisions." International Journal of Computers Communications & Control 20.1 (2025).
[8] Xu, Bin, et al. "The future of live-streaming commerce: understanding the role of AI-powered virtual streamers." Asia Pacific Journal of Marketing and Logistics 37.5 (2025): 1175-1196.
[9] Peng, Yuhong, et al. "Impact of AI-oriented live-streaming E-commerce service failures on consumer disengagement—empirical evidence from China." Journal of Theoretical and Applied Electronic Commerce Research 19.2 (2024): 1580-1598.
[10] Zhang, Longyun, et al. "Development and validation of an AI virtual streamer scale for live-streaming E-commerce." International Journal of Human–Computer Interaction 41.14 (2025): 8525-8538.
[11] Visser, Boele, Peter van der Putten, and Amirhossein Zohrehvand. "The Impact of AI Avatar Appearance and Disclosure on User Motivation." International Conference on Data Science and Artificial Intelligence. Singapore: Springer Nature Singapore, 2024.
[12] Lee, Minyoung, Sanghyun Kim, and Jaehyuk Yi. "A Relationship between Social and Technical IT Affordance of Live-commerce and Intention to Use: The Moderating Effect of Social Presence." 인터넷전자상거래연구 23.2 (2023): 1-21.