

Navigating the Trust Paradox: AI-Generated Fake Reviews, Influencer Transparency, and Consumer Purchase Intentions in Digital Markets

Ramyashree M. S. ^{1,*}, Usha N.², Sahana Shree N.³, Navita Vijay ¹, and S. Sivasankari⁴

¹*School of Commerce and Management Studies (PG), Dayananda Sagar University, Bengaluru, Karnataka, India*

²*School of Commerce and Management Studies, Dayananda Sagar University, Bengaluru, Karnataka, India*

³*Department of Commerce and Management, Dayananda Sagar College of Arts, Science and Commerce, Bengaluru, Karnataka, India*

⁴*RV Institute of Management, Bengaluru, Karnataka, India*

*Corresponding author: ramyashreemr@gmail.com

Abstract. The marketing industry is increasingly filled with content generated by artificial intelligence (AI), yet the effect of this content on consumer trust remains largely unknown. This study investigated the effects of AI-generated fake reviews, bot engagement, influencer transparency, and perceived authenticity on consumer trust, brand credibility, and purchase intention. Data from a quantitative survey of 402 respondents across age groups and product categories were analyzed with Pearson correlation, mediation analysis, moderation analysis (linear regression), exploratory factor analysis, and chi-square tests. The results point to a “transparency paradox” in influencer marketing: explicit influencer disclosure was weakly but significantly associated with lower trust, and exposure to AI-generated fake engagement had a moderate negative association with purchase intention. Perceived authenticity and consumer skepticism formed the strongest mediating construct and accounted for a large part of the link between transparency and trust. The study extends source credibility theory, dual-process cognition, and heuristic persuasion by showing that consumers recognize inauthenticity and still respond to volume-based social proof. Brands, content platforms, and regulators can draw on these results to design ethical and effective digital marketing ecosystems.

Keywords. AI-generated fake reviews; influencer transparency; consumer trust; perceived authenticity; purchase intention; brand credibility; social proof

I. INTRODUCTION

Digital marketing has changed how consumers make decisions, and they no longer decide independently. With the rise of social media platforms, influencer marketing has become one of the most powerful and most contested vehicles for building brand perception. Alongside genuine human influence, a new ecosystem of AI-generated fake reviews, bots, and algorithmically enhanced social proof is blurring the line between real recommendation and fabricated consensus [1], [2]. Regulators, brands, and consumers are all affected. According to the Advertising Standards Council of India (ASCI), the influencer marketing industry in India is expected to exceed INR 3,500 crore by 2028. With rising disclosure fatigue and synthetic engagement, the trust deficit is also growing. ASCI has strengthened its mandatory disclosure norms, but violations continue. Does transparency help build trust, or could it be counterproductive? Previous studies have addressed influencer transparency, fake reviews, and consumer trust, but separately. What is lacking is an integrated empirical examination of how these constructs interact, especially the mediating role of

perceived authenticity and consumer skepticism and the moderating role of social media usage frequency. This study addresses that gap.

The paper examines the effect of AI fake reviews and bot engagement (AIFRBE), influencer transparency and ethicality perceptions (ITEP), and perceived authenticity, consumer skepticism, and social proof (PACSSP) on consumer trust, purchase intention, and brand credibility (CTPIBC) in a sample of 402 digital consumers in India. The study makes three contributions. First, it provides empirical evidence of a “transparency paradox,” the counterintuitive finding that explicit sponsorship disclosure may slightly undermine rather than enhance trust. Second, it shows that the influencer-trust relationship runs through PACSSP, with skepticism as a strong intervening mechanism. Third, it shows that the frequency of social media use significantly moderates the authenticity-to-trust relationship, with the most frequent users responding more warily. Together, the findings extend source credibility theory, dual-process cognition frameworks, and heuristic persuasion models to AI-driven digital commerce.

II. LITERATURE REVIEW

A. Source Credibility and Consumer Trust

According to the source credibility model of Hovland, Janis, and Kelley [3], a message will not persuade if the communicator does not appear competent and trustworthy. Online, the model has been extended to influencers and online reviews as important trust signals [4], [5]. When sources are seen as credible, consumers are more likely to accept their messages and make a purchase. Credibility is easily manipulated, however, by AI-generated content, which imitates an authentic voice without the human experience behind it [2].

B. AI-Generated Fake Reviews and Bot Engagement

Fake reviews are among the most widespread forms of online fraud. Mayzlin, Dover, and Chevalier [6] provided evidence that hotels use review platforms to manipulate the standing of their competitors. Generative AI can now write fake reviews that human readers struggle to tell apart from genuine ones [2]. Bot-generated engagement creates artificial social proof that triggers heuristic judgments among consumers [7]. Consumers who suspect that reviews are fake come to distrust them, which weakens the value of reviews for consumers, platforms, and brands [1], [8].

C. Influencer Transparency and the Disclosure Paradox

Disclosure rules such as the 2019 guidance of the US Federal Trade Commission (FTC) [9] and the 2021 ASCI guidelines in India require influencers to label their sponsorships. Transparency can build trust by signaling honesty: an influencer disclosure that explains why a post is sponsored increased source and message credibility above a simple disclosure [10]. An emerging counter-narrative documents a “disclosure backfire effect”: an explicit sponsorship label activates consumers' persuasion knowledge, and they become resistant to the message [11], [12]. The present study builds on these mixed findings and assumes that transparency may lessen trust under certain conditions.

D. Perceived Authenticity, Skepticism, and Social Proof

Authenticity is the central claim of current influencer marketing [13]. Consumers judge a brand's authenticity by the consistency of its content, the alignment between its stated values and its actual behavior, and the absence of overt commercial intent [14].

Consumer skepticism, the tendency to disbelieve advertising claims [15], matters more as exposure to sponsored content grows. The elaboration likelihood model (ELM) of Petty and Cacioppo [16] predicts that consumers under high elaboration pay close attention to message quality, while consumers under low elaboration may rely on social proof heuristics such as the number of followers or reviews. This dual-process view helps explain how AI-generated engagement can create both suspicion and bandwagon effects.

III. METHODOLOGY

A. Research Design

A cross-sectional quantitative survey design was used. A structured online questionnaire was administered to 402 respondents in five age groups: below 18 ($n = 39$), 18 to 24 ($n = 165$), 25 to 34 ($n = 64$), 35 to 44 ($n = 86$), and 45 and

above ($n = 48$). Purposive and snowball sampling were used to reach active social media users who had made at least one influencer-driven purchase in the previous six months.

B. Instrument Development

Four multi-item constructs were operationalized: AIFRBE (three items), ITEP (four items), PACSSP (four items), and CTPIBC (five items). The items were adapted from validated scales in the literature and rated on a five-point scale (1 = strongly disagree, 5 = strongly agree). An exploratory factor analysis (EFA) with Promax rotation confirmed a single factor for AIFRBE, which accounted for a cumulative variance of 63.4% (eigenvalue 2.198).

C. Reliability

Internal consistency was measured with Cronbach's alpha (Table 1). CTPIBC showed the highest reliability ($\alpha = 0.840$), followed by AIFRBE ($\alpha = 0.803$). PACSSP ($\alpha = 0.745$) and ITEP ($\alpha = 0.739$) also exceeded the reliability threshold of 0.70 [17].

IV. OBJECTIVES AND HYPOTHESES

The study pursued four objectives: to examine the direct relationships among AIFRBE, ITEP, PACSSP, and CTPIBC; to test whether PACSSP mediates the effect of ITEP on CTPIBC; to assess whether social media usage frequency moderates the PACSSP-CTPIBC relationship; and to identify age- and gender-based differences in engagement with influencer-driven product categories. Six hypotheses were tested.

TABLE 1. RELIABILITY OF THE CONSTRUCTS

Construct	Items	Cronbach's α
AIFRBE	3	0.803
ITEP	4	0.739
PACSSP	4	0.745
CTPIBC	5	0.840

H1: There is a significant association between ITEP and CTPIBC.

H2: There is a significant association between AIFRBE and CTPIBC.

H3: There is a significant association between AIFRBE and PACSSP.

H4: There is a significant association between ITEP and PACSSP.

H5: There is a significant association between PACSSP and CTPIBC.

H6 (moderation): Social media usage frequency significantly moderates the association between PACSSP and CTPIBC.

V. FINDINGS AND DISCUSSION

A. Correlation Analysis (H1 to H5)

Table 2 summarizes the Pearson correlations for the five direct hypotheses, all significant at $p < 0.001$.

H1 (ITEP to CTPIBC). Influencer transparency and consumer trust showed a statistically significant but weak negative relationship ($r = -0.167$, $p < 0.001$). The direction was unexpected, but the association is significant, so H1 is

supported. The finding does not confirm the assumption that disclosure builds trust and points instead to a “transparency paradox.” When influencers signal that a relationship is commercial, consumers activate their persuasion knowledge [18], which suppresses rather than enhances trust. This agrees with the disclosure backfire literature [12] and implies that formulaic disclosures without a relatable narrative may come across as inauthentic rather than honest.

TABLE 2. HYPOTHESIS TESTING: PEARSON CORRELATION SUMMARY

Hypothesis and path	r	Strength
H1: ITEP to CTPIBC	-0.167	Weak, negative
H2: AIFRBE to CTPIBC	-0.283	Moderate, negative
H3: AIFRBE to PACSSP	0.294	Moderate, positive
H4: ITEP to PACSSP	0.195	Weak, positive
H5: PACSSP to CTPIBC	-0.456	Moderate to strong, negative

Note: Primary data, $n = 402$; $p < 0.001$ for all five correlations; H1 to H5 all supported.

H2 (AIFRBE to CTPIBC). A moderate negative correlation between AIFRBE and CTPIBC ($r = -0.283$, $p < 0.001$) shows that exposure to AI-generated fake reviews and bot engagement is associated with lower consumer trust and purchase intention. Under source credibility theory, artificially engineered engagement lowers the perceived expertise and trustworthiness of a source [3]. AI-generated text has become more linguistically polished than human writing, which strains consumers' ability to detect it. A diffuse distrust still persists and damages both the review platform and the brand under review [1].

H3 (AIFRBE to PACSSP). AI-generated fake engagement and PACSSP showed a moderate positive correlation ($r = 0.294$, $p < 0.001$), perhaps the most surprising result. It reflects the dual-process model of Petty and Cacioppo [16]: consumers may recognize that the metrics are inflated and still treat them as signals of popularity. The volume of engagement, even when believed to be fake, can activate bandwagon heuristics [7] and produce a reverse social proof effect. This shows how heuristics can be exploited in online environments.

H4 (ITEP to PACSSP). ITEP and PACSSP showed a weak positive association ($r = 0.195$, $p < 0.001$), which suggests that influencer transparency raises perceived authenticity slightly. This is directionally consistent with Audrezet, de Kerviler, and Guidry Moulard [13], but the small effect size implies that transparency alone is not enough. Consistency with lived experience, a consistent content style, and narrative coherence are probably stronger determinants of perceived authenticity, and without them disclosures may contribute little to trust.

H5 (PACSSP to CTPIBC). PACSSP and CTPIBC showed the strongest bivariate association in the model ($r = -0.456$, $p < 0.001$), with higher PACSSP scores accompanying significantly lower consumer trust and purchase intention. Consumers increasingly doubt curated commercial content, which makes them far less likely to trust or buy. Skepticism, once treated as a limited moderating variable, here functions as a strong mediating variable, and brands that do not address it structurally will continue to face trust deficits [15].

B. Mediation Analysis

Maximum likelihood mediation analysis showed that PACSSP fully mediated the effect of ITEP on CTPIBC. The indirect effect was significant and negative ($\beta = -0.086$). The direct path from ITEP to CTPIBC ($\beta = -0.081$, $p = 0.072$) lost significance once the mediator was included, while the total effect remained significant ($\beta = -0.167$, $p < 0.001$). In other words, influencer transparency raises consumer skepticism, which shapes perceived authenticity and in turn lowers trust. Brands and influencers should address skepticism at its source rather than assume that disclosure will resolve the issue on its own.

C. Moderation Analysis (H6)

Hierarchical linear regression (Model 1: $R^2 = 0.150$, adjusted $R^2 = 0.137$, $F(6, 395) = 11.609$, $p < 0.001$) indicated that the frequency of social media use significantly moderates the PACSSP-CTPIBC relationship. Respondents who used social media sometimes ($\beta = 2.240$, $p < 0.001$), often ($\beta = 3.710$, $p < 0.001$), or very frequently ($\beta = 2.322$, $p < 0.001$) had higher baseline trust than infrequent users. The picture is mixed, though. The interaction terms (PACSSP $\times$ sometimes: $\beta = -0.501$, $p = 0.002$; PACSSP $\times$ often: $\beta = -0.819$, $p < 0.001$; PACSSP $\times$ very frequently: $\beta = -0.463$, $p = 0.004$) confirm that the trust-reducing effect of PACSSP is stronger for more frequent users. Frequent social media users are not naive. They watch for performative authenticity and penalize brands faster when they sense dissonance.

D. Chi-Square Tests: Age and Gender Differences

Marketers should not overlook age and gender in influencer-driven engagement. Fashion was, unsurprisingly, strongly associated with the 18 to 24 age group ($\chi^2 = 30.410$, $p < 0.001$), while the 45 and above group was, more surprisingly, most strongly associated with Health and Fitness ($\chi^2 = 17.790$, $p = 0.001$). Home and Kitchen showed a similar age dependency ($\chi^2 = 15.506$, $p = 0.004$). By gender, women gravitated toward Beauty and Personal Care and men toward Health and Fitness, while Home and Kitchen split along conventional lines; all of these differences were statistically significant.

VI. IMPLICATIONS

A. For Brands and Marketers

Brands need to go beyond ticking compliance boxes and make transparency part of their stories, so that disclosure reads as sincere rather than perfunctory. Working with influencers whose followers are genuinely engaged is now essential, because inflated bot-driven numbers damage credibility.

B. For Influencers and Content Creators

For influencers, authenticity lies in the message, not the form. Audiences are more sophisticated, quicker to detect performance, and much less tolerant of lost trust. Disclosures should be honest and sincere, and in the long run a sponsored post alone does not build trust.

C. For Policymakers and Platforms

Policymakers and platforms need to move beyond binary disclosure rules, both under ASCI and among its global counterparts. ASCI should take an ecosystem-level approach, including AI detection tools that flag fake reviews and an escalation process for repeat offenders at the platform level. Digital platforms should also be accountable for the algorithms that amplify fake engagement on them.

D. Theoretical Implications

The study extends source credibility theory to AI-powered influencer marketing and gives dual-process cognition theory an empirical basis in the context of fake engagement. The full mediation through PACSSP adds to the understanding of trust formation and identifies skepticism as an important mediator. The results also show that frequency of social media use moderates the authenticity effects.

VII. CONCLUSION

The meeting of AI and influencer marketing produces a complex set of trust relationships. Based on survey data from 402 digital consumers, this study found that the connections among transparency, fake engagement, authenticity, and trust are neither linear nor simple. Influencer disclosure was associated with slightly lower, not higher, trust, and AI-powered fake engagement was associated with lower trust while still boosting social proof through heuristic routes. Consumer skepticism is an important factor in how influencer behavior translates into consumer trust and purchase. The picture that emerges is of a consumer population that is more skeptical but still susceptible: more aware, yet still responsive to volume-based social cues. Practitioners should treat authenticity as a long-term strategic investment rather than a tactic. Where AI can generate large volumes of convincing false content, disclosure-based regimes are inadequate, and the changing trust structures of AI-driven markets need continued study. Future work should test these conclusions with longitudinal studies, cross-platform comparisons, and experiments that manipulate and isolate the mechanisms. Cross-cultural studies would show whether the transparency paradox and the dual-process effects observed here are specific to the Indian digital context or hold more generally.

AUTHOR CONTRIBUTIONS

All authors contributed to the conception and preparation of this work and approved the final version of the manuscript.

FUNDING

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

DECLARATION OF GENERATIVE AI USE

The authors declare that no generative artificial intelligence tools were used in the preparation of this manuscript. The authors take full responsibility for the content of the publication.

DATA AVAILABILITY

The data supporting the findings of this study are available from the corresponding author upon reasonable request.

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

REFERENCES

- [1] Yuanyuan Wu, Eric W. T. Ngai, Pengkun Wu, and Chong Wu, "Fake online reviews: Literature review, synthesis, and directions for future research," *Decis. Support Syst.*, vol. 132, Art. no. 113280, 2020, doi: 10.1016/j.dss.2020.113280.
- [2] Joni Salminen, Chandrashekhara Kandpal, Ahmed Mohamed Kamel, Soon-gyo Jung, and Bernard J. Jansen, "Creating and detecting fake reviews of online products," *J. Retail. Consum. Serv.*, vol. 64, Art. no. 102771, 2022, doi: 10.1016/j.jretconser.2021.102771.
- [3] Carl I. Hovland, Irving L. Janis, and Harold H. Kelley, *Communication and Persuasion: Psychological Studies of Opinion Change*. New Haven, CT, USA: Yale Univ. Press, 1953.
- [4] Roobina Ohanian, "Construction and validation of a scale to measure celebrity endorsers' perceived expertise, trustworthiness, and attractiveness," *J. Advert.*, vol. 19, no. 3, pp. 39-52, 1990, doi: 10.1080/00913367.1990.10673191.
- [5] Christy M. K. Cheung and Dimple R. Thadani, "The impact of electronic word-of-mouth communication: A literature analysis and integrative model," *Decis. Support Syst.*, vol. 54, no. 1, pp. 461-470, 2012, doi: 10.1016/j.dss.2012.06.008.
- [6] Dina Mayzlin, Yaniv Dover, and Judith Chevalier, "Promotional reviews: An empirical investigation of online review manipulation," *Amer. Econ. Rev.*, vol. 104, no. 8, pp. 2421-2455, 2014, doi: 10.1257/aer.104.8.2421.
- [7] Robert B. Cialdini, *Influence: Science and Practice*, 4th ed. Boston, MA, USA: Allyn and Bacon, 2001, ISBN 0-321-01147-3.
- [8] Wasim Ahmad and Jin Sun, "Modeling consumer distrust of online hotel reviews," *Int. J. Hosp. Manag.*, vol. 71, pp. 77-90, 2018, doi: 10.1016/j.ijhm.2017.12.005.
- [9] Federal Trade Commission, "Disclosures 101 for social media influencers," FTC, Washington, DC, USA, Nov. 2019. [Online]. Available: <https://www.ftc.gov/business-guidance/resources/disclosures-101-social-media-influencers>
- [10] Carolina Stubb, Anna-Greta Nyström, and Jonas Colliander, "Influencer marketing: The impact of disclosing sponsorship compensation justification on sponsored content effectiveness," *J. Commun. Manag.*, vol. 23, no. 2, pp. 109-122, 2019, doi: 10.1108/JCOM-11-2018-0119.
- [11] Margaret C. Campbell and Amna Kirmani, "Consumers' use of persuasion knowledge: The effects of accessibility and cognitive capacity on perceptions of an influence agent," *J. Consum. Res.*, vol. 27, no. 1, pp. 69-83, 2000, doi: 10.1086/314309.
- [12] Sophie C. Boerman, Lotte M. Willemsen, and Eva P. Van Der Aa, "'This post is sponsored': Effects of sponsorship disclosure on persuasion knowledge and electronic word of mouth in the context of Facebook," *J. Interact. Mark.*, vol. 38, no. 1, pp. 82-92, 2017, doi: 10.1016/j.intmar.2016.12.002.
- [13] Alice Audrezet, Gwarlann de Kerviler, and Julie Guidry Moulard, "Authenticity under threat: When social media influencers need to go beyond self-presentation," *J. Bus. Res.*, vol. 117, pp. 557-569, 2020, doi: 10.1016/j.jbusres.2018.07.008.
- [14] Mike Schallehn, Christoph Burmann, and Nicola Riley, "Brand authenticity: Model development and empirical testing," *J. Prod. Brand Manag.*, vol. 23, no. 3, pp. 192-199, 2014, doi: 10.1108/JPB-06-2013-0339.
- [15] Carl Obermiller and Eric R. Spangenberg, "Development of a scale to measure consumer skepticism toward advertising," *J. Consum. Psychol.*, vol. 7, no. 2, pp. 159-186, 1998, doi: 10.1207/s15327663jcp0702_03.
- [16] Richard E. Petty and John T. Cacioppo, "The elaboration likelihood model of persuasion," in *Advances in Experimental Social Psychology*, vol. 19. New York, NY, USA: Academic Press, 1986, pp. 123-205, doi: 10.1016/S0065-2601(08)60214-2.
- [17] Jum C. Nunnally, *Psychometric Theory*, 2nd ed. New York, NY, USA: McGraw-Hill, 1978, ISBN 0-07-047465-6.
- [18] Marian Friestad and Peter Wright, "The persuasion knowledge model: How people cope with persuasion attempts," *J. Consum. Res.*, vol. 21, no. 1, pp. 1-31, 1994, doi: 10.1086/209380.