DOI: 10.71317/jgst.2.11.2026.1089

Journal of Global Social Transformation

Journal homepage: https://rjsaonline.org/index.php/JGST

AI Marketing Transparency and Brand Trust: How AI Literacy and Perceived Manipulative Intent Shape Consumer Responses in Pakistan

Mohammad Aminul Islam[1], Muhammad Umair Khalid[2], Sheikh Muhammad Ali[3], Mian Haseeb Mushtaq[4]

[1] Department of Management, Faculty of Business Studies, University of Dhaka, Dhaka 1000, Bangladesh.
Email: md.aminulislam8899@gmail.com

[2] Department of Business Studies, Bahria Business School (Karachi), Bahria University, Karachi Campus, Pakistan.
Email: umairkhalid.bukc@bahria.edu.pk

[3] Department of Business Studies, Bahria Business School (Karachi), Bahria University, Karachi Campus, Pakistan.
Email: smali.bukc@bahria.edu.pk

[4] Department of Software Engineering, Lahore Garrison University, Pakistan.
Email: haseebmuhstaq@gmail.com

ARTICLE INFO

Submission:
October 12, 2026
Revised:
October 31, 2026
Accepted:
November 11, 2026
Available Online:
November 21, 2026

Keywords:
AI marketing transparency; AI literacy; perceived manipulative intent; brand trust; purchase intention; Pakistan; moderated mediation

Corresponding author:
Mian Haseeb Mushtaq
Email: haseebmuhstaq@gmail.com

ABSTRACT

With AI's increasing role in digital marketing, a brand's transparency about its use of AI could act as an informational signal of accountability. However, the interpretation consumers make of such transparency might vary with their knowledge of AI and their perception of persuasive intent. This study examines the relationship between perceived AI marketing transparency, brand trust, and purchase intention and the role of perceived manipulative intent as a mechanism and AI literacy as a boundary condition. We surveyed 387 matched Pakistani consumers, in two waves. AI literacy was operationalised with the core 18-item AI literacy dimensions of the MAILS instrument (use and apply AI, know and understand AI, detect AI, and AI ethics), whereas the remaining MAILS dimensions were treated as separate constructs. We tested the hypotheses using regression-based simple mediation (PROCESS Model 4) and first-stage moderated mediation (PROCESS Model 7). Perceived AI marketing transparency was positively associated with brand trust ($\beta = 0.24$, $p < 0.001$). Brand trust carried a significant indirect association between perceived AI marketing transparency and purchase intention (indirect effect = 0.06, 95% CI [0.02, 0.11]); the direct association was not statistically significant ($c' = 0.10$, $p = 0.081$). Perceived manipulative intent was negatively associated with brand trust ($b = -0.25$, $p < 0.001$). AI literacy positively moderated the association between perceived AI marketing transparency and perceived manipulative intent ($B = 0.13$, $p = 0.031$), such that this association was positive and significant only at high AI literacy ($B = 0.25$, $p < 0.001$). The conditional indirect association was statistically distinguishable from zero only at high AI literacy (effect = −0.06, 95% CI [−0.13, −0.01]). The index of moderated mediation was statistically significant at the 5% level (index = −0.03, 95% CI [−0.07, −0.001]) but the CI was close to the null value. The results are consistent with both theoretically opposite mechanisms suggested: an accountability-related association between perceived transparency and brand trust, and a conditional suspicion-related indirect association via perceived manipulative intent, which were analysed under the umbrella of two different empirical models. The interaction was small and the moderated-mediation evidence was statistically weak and should be considered a preliminary rather than a general finding of the negative effects on transparency. One of the contributions of the study is distinguishing perceived transparency from its interpretation, and examining core AI literacy competencies in addition to a focused measure of persuasion knowledge activation in the context of AI marketing and signaling research. Finally, practical implications highlight the importance of AI disclosure being credible and context-specific, not necessarily a strategy that builds trust.

Introduction

In Pakistan, the digital commerce scene has its own trajectory, challenging the conventional story of AI's empowerment of consumers. According to the Stay Secure study by Visa (2026), 82% of Pakistani consumers have utilized AI tools for price comparison, checking product reviews, and discovering gifts. while 93% feel that AI-powered tools make shopping faster and easier. However, only 42% would trust an AI agent to finish the checkout process and 55% had a financial scam in the last year. The 55% is the percentage of respondents who were a victim of a financial scam in the last year, overall, and not particularly related to AI-powered shopping. Consumers aren't turning their backs on AI, they're using AI to search and systematically taking away transaction AI trust.

These statistics are projections of the debate in progress. They find AI shopping is everywhere and all around. Transactional trust is limited. But, not so much among consumers who distrust brands' disclosures of AI. This inference must be derived from evidence of how consumers interpret disclosure when they see it. This is the question that the present study tackles from a marketing perspective, because its effectiveness depends not only on AI transparency, but on how consumers perceive it when it is communicated as a brand. Transparency, usually linked with accountability and openness, can be interpreted in different ways by consumers and be a factor influencing consumer trust. Transparency can convey responsible use of technology, and at the same time provide information that customers can assess as to its persuasive or manipulative purpose in AI-driven marketing. These interpretations do not necessarily have to be applied in the same way. A consumer that is more AI-literate might have more awareness of persuasive mechanisms related to AI and therefore might have different interpretations of the AI-related brand communication. This interpretation difference could be a "better" pathway of suspicion, but only an empirical investigation can determine this. The present study investigates this relationship by considering AI literacy as a moderator in this relationship, and by examining the conditional indirect relationship between perceived AI marketing transparency and perceived manipulative intent and brand trust.

Marketing research in Pakistan positions brand trust as a central pathway through which AI-related perceptions relate to consumer responses in a digital market characterized by a trust deficit. Yet the empirical record on whether perceived transparency is associated with trust is unsettled. Pakistan-based evidence indicates that AI-related marketing communications can activate consumers' persuasion knowledge and prompt critical evaluation of persuasive intent (Qadri et al., 2025), while other Pakistan research reports that AI disclosure transparency is positively associated with purchase intention through consumer trust and brand perception (Ahmed et al., 2026). But Ali and Sunela (2026) found that trust did not significantly mediate the transparency-purchase intention relationship, suggesting that the transparency's pathway is weaker, more complex, or conditional than originally assumed. Internationally, a meta-analysis of 21 experimental effects (N = 13,582) identified a general negative effect of AI disclosure on consumer responses, with significant heterogeneity across task contexts (Kim, 2026). While AI disclosure and perceived AI marketing transparency are not identical constructs, this evidence emphasises the significance of exploring the conditions under which consumers interpret AI-related brand communication.

While the literature on AI-enabled marketing and consumer trust is growing, the link between perceived AI marketing transparency and consumers' attribution of manipulative intent is not well understood. Empirically little is known about whether consumers' understanding of brand disclosures on AI is affected by AI literacy and whether these understandings are related to brand trust. Existing research has examined transparency, trust, AI anxiety, and digital literacy, but these constructs have rarely been integrated into a framework that distinguishes the favorable trust-related interpretation of transparency from its potentially unfavorable suspicion-related interpretation. To fill this gap, this study aims to explore perceived AI marketing transparency as a consumer evaluation and the perceived manipulative intent as another consumer evaluation, as well as explore the role of AI literacy as a boundary condition in its relationship.

Notably, the study does not include an analysis of the impact of AI disclosure manipulated in a laboratory setting or AI disclosure as objectively established practices. It explores the level of transparency that consumers feel about a stimulus brand's utilization of AI. The results are thus not necessarily causal, in the sense of establishing that a disclosure intervention causes a particular change in the association between consumers' perceptions and evaluations. The results, therefore, focus on associations among consumers' perceptions and evaluations, and not on the causal impact of particular disclosure interventions. This distinction matters in a marketing context because perceived transparency may reflect both the information a brand communicates and consumers' interpretations of that information.

The study makes three contributions to AI marketing research. First, it identifies perceived manipulative intent as a suspicion-related pathway linking perceived AI marketing transparency with brand trust, while examining AI literacy as a moderator of the

transparency–manipulative-intent association. This provides a more differentiated account of how consumers interpret AI-related brand communication than treating transparency as either uniformly beneficial or uniformly harmful. Second, it advances a two separately estimated pathways: perceived transparency is associated with a favorable trust pathway and, conditionally, with a suspicion-related pathway, rather than being uniformly beneficial or harmful. This goes beyond adding a moderator to an existing mediation model, because it specifies two theoretically distinct interpretive routes that are examined through separate empirical models. Third, it distinguishes core AI literacy competencies from a focused measure of persuasion knowledge activation, and offers exploratory evidence that directly measured persuasion knowledge activation may be a stronger boundary condition than general AI literacy. This distinction is treated cautiously because the two constructs overlap empirically and conceptually.

Theoretical Background and Hypotheses

Signaling Theory and Perceived AI Marketing Transparency

Signaling theory concerns the communication of information under conditions of information asymmetry (Spence, 1973). It has been well-exploited in marketing, especially for research on the theory's boundary conditions and receiver-side processes (Connelly et al., 2011) and for research on brand equity as a signaling phenomenon (Erdem & Swait, 1998). Here three analytically distinct stages are relevant to consider. Signal availability is the presence of information being transmitted. The credibility of the signal is determined by whether it is costly or verifiable and is, therefore, likely to be honest. The decoding process, or receiver interpretation, is shaped by the receiver's knowledge, prior beliefs and institutional context (Connelly et al., 2011). Perceived transparency in the use of AI in marketing can act as an informational cue indicating openness and helping to alleviate uncertainty regarding a brand's technological approach, which could be beneficial in AI-driven marketing strategies. Perceived transparency, however, is a receiver-side phenomenon; it measures consumers' understanding of what has been communicated and not a third-party verification of the credibility of the signals. A consumer can feel that a brand is transparent without having the information to make an informed determination whether the disclosure is honest, complete and expensive to generate. Therefore, it is not a given that perceived transparency of AI marketing will lead to evenly positive evaluations. Perceived transparency can also be seen as a sign of accountability, a credibility-based interpretation of the signal; or as signs of persuasive intentions or technologies that should be investigated. This helps to explore perceived transparency as an information cue whose meaning may have positive or negative connotations for consumers' evaluations. In the Pakistani context, it is more of a setting than a tested mechanism. This marriage of mass-market AI-powered shopping and the reported issues with trust in transactions and potential exposure to scams underscores the need to gauge consumer perception of AI-driven marketing information. However, even if these contextual indicators do reduce trust in institutions, they do not necessarily affect consumers' interpretations of brand transparency. No measures of institutional trust and verification mechanisms were taken. This context justifies the inclusion of perceived manipulative intent and AI literacy as individual level factors that may be relevant to the evaluations for transparency, but are not tested but as explanatory mechanisms.

The Pakistan evidence on transparency is mixed. AI disclosure transparency has been modelled as an antecedent of purchase intention through consumer trust and brand perception, with indirect pathways statistically significant in one PLS-SEM study of 350 online consumers (Ahmed et al., 2026). Yet Ali and Sunela (2026) found that trust mediated personalization but not transparency, suggesting that transparency's pathway is weaker, more complex, or conditional.

H1: Perceived AI marketing transparency is positively associated with brand trust.

Brand Trust as the Relational Mediator

Brand trust is the expectancy that a brand can be relied upon to deliver on its promises, and it is conventionally decomposed into ability, benevolence, and integrity (Mayer et al., 1995). In consumer research, brand trust has been shown to mediate the effects of brand-level evaluations on brand performance outcomes such as purchase loyalty (Chaudhuri & Holbrook, 2001), and it occupies a central mediating position in emerging-market AI marketing research. Brand trust was conceptualized through the ability, benevolence, and integrity dimensions identified in the integrative trust framework (Mayer et al., 1995), with item wording adapted to the consumer–brand context. Consumers who perceive greater transparency regarding a brand's AI use may also report greater brand trust. Because trust represents an important evaluative basis for purchase-related decisions, brand trust may account for part of the association between perceived AI marketing transparency and purchase intention. We hypothesize mediation without committing to full mediation in advance, because the Pakistan evidence is mixed.

H2: Brand trust mediates the positive association between perceived AI marketing transparency and purchase intention.

Perceived Manipulative Intent as a Distinct Marketing Evaluation

The Persuasion Knowledge Model holds that consumers develop knowledge about persuasion tactics and that this knowledge is activated when consumers recognize a persuasion attempt (Friestad & Wright, 1994). After the interpretive frame is activated, the frame becomes "Why is the sender telling me this now?". This is a change from perceived transparency as an information signal to a motive-attribution cue. The belief that a sender is trying to persuade excessively or inappropriately using tactics is a consumer response that is well known in the study of persuasion (Campbell, 1995). Presumably, accessibility of persuasion knowledge and cognitive capacity to put it into use is a condition when such inferences emerge (Campbell & Kirmani, 2000). In the case of AI marketing, we use perceived manipulative intent as our outcome.

It is important, however, to make an important clarification about what is being perceived. The perception of manipulative intention here is when respondents perceive the disclosure as a strategic communications maneuver, that is, when they experience the disclosure as a marketing tactic. It doesn't speak to the perception of the underlying AI practices of the brand (such as biases in algorithms, or the intrusiveness of data practices). These are conceptually different judgments, and the former only is reflected in the present measure. This is important for the results interpretation as it relates to the suspicions pathway rather than to consumers' beliefs on AI systems.

Additionally, the present study separates the core AI competencies from a specific measure of persuasion knowledge activation. While AI literacy is about consumers' understanding and skills around AI, persuasion knowledge is about the understanding and skills of persuasive strategies and how to identify and assess persuasive attempts. Even though they are related, they are different. The instrument used to measure AI literacy, called MAILS, identifies several scales, one of which is the 'AI self-competency (persuasion literacy)' scale, which measures awareness of AI-using persuasion (Carolus et al., 2023). In the present study, the primary AI literacy moderator is operationalized using the four core AI literacy dimensions (use and apply AI, know and understand AI, detect AI, and AI ethics), which constitute 18 items and correspond to the AI literacy construct as originally conceptualized. The Create AI, AI self-efficacy, and AI self-competency dimensions are treated as separate constructs and are not included in the primary moderator. This operationalization follows the original MAILS factor structure, in which the core AI literacy dimensions are distinguished from the psychological competency dimensions. A sensitivity analysis using the full 34-item composite is reported in Section 4.5. Perceived manipulative intent represents the outcome of these interpretive processes a context-specific evaluation of a brand's AI-related marketing communication and should not be read as a direct measure of persuasion knowledge activation.

H3: Perceived manipulative intent is negatively associated with brand trust.

We anticipate that the indirect association through perceived manipulative intent will hold conditionally rather than as main-effect mediation. Perceived transparency is unlikely to raise suspicion uniformly; it should do so for consumers for whom strategic attribution is cognitively accessible. That is the proposed role of AI literacy.

AI Literacy as a Boundary Condition

AI literacy may provide consumers with knowledge about AI capabilities, applications, which could be significant for their assessment of AI-supported marketing communication. This type of knowledge could shape the interpretation of information regarding the use of AI, but not necessarily lead to higher levels of persuasion knowledge or increased likelihood of attributing manipulative intent to a brand. AI literacy is thus seen as a potential boundary condition between perceived AI marketing transparency and perceived manipulative intent. Why could there be suspicion, not credulity, with AI literacy? There are two plausibility arguments that are relevant. First, domain knowledge makes it easier for users to understand the value of the ability of AI to target, personalize, or optimize content, increasing the salience of this ability.

Secondly, those who know the production of AI-generated and AI-optimized content may be better equipped to realize that disclosure is a marketing decision. Such arguments make sense but are not evaluated in the current design. The study consequently does not treat the mechanism as a process actually demonstrated but as an assumption that can be made based on some evidence of its activation of persuasion knowledge that is reported in Section 4.5. The core AI literacy dimensions of the MAILS instrument (use and apply AI, know and understand AI, detect AI, and AI ethics) are used as the primary AI literacy moderator in this study.

This operationalization follows the original MAILS factor structure, in which these four dimensions constitute the AI literacy construct and are distinguished from Create AI, AI self-efficacy, and AI self-competency (which includes persuasion literacy and emotion regulation). The core dimensions capture technological knowledge and competency rather than persuasion-related

awareness, which makes them a more appropriate boundary condition for testing whether general AI literacy conditions the transparency–suspicion association. A sensitivity analysis using the full 34-item composite is reported in Section 4.5.

Pakistan-specific evidence is consistent with, but does not directly test, this reasoning. Qadri et al. (2025) examine digital literacy rather than AI literacy among young Pakistani consumers in a three-wave time-lagged design, and demonstrate that digital literacy operates as a dual-type moderator: it mitigates the negative impacts of perceived manipulation at higher levels but amplifies them at lower levels, so the literacy–scepticism relationship is not uniformly protective. This finding is not specific to AI literacy, but instead provides theoretical context for the general claim that literacy is not always protective.

A global study of 13,582 participants in 21 experimental effects revealed a general negative impact of the disclosure of AI on consumer responses, albeit with significant differences depending on task contexts. The negative effect was more pronounced in objective contexts like advertising and service interactions than in subjective informational (Kim, 2026). This evidence provides support for the role of contextual conditions as moderators in AI disclosure effects, but does not directly validate AI literacy as a moderator. In the present study, however, moderation is tested on the basis of the core AI literacy dimensions and persuasion knowledge activation is measured separately to allow for the examination of the mechanism instead of assuming it.

H4: AI literacy positively moderates the association between perceived AI marketing transparency and perceived manipulative intent, such that the association becomes more positive at higher levels of AI literacy.

Conditional Indirect Association

If AI literacy moderates the association between perceived AI marketing transparency and perceived manipulative intent, and perceived manipulative intent is negatively associated with brand trust, the resulting indirect association between transparency and trust may vary across levels of AI literacy. Specifically, the indirect association is expected to become more negative as AI literacy increases. This proposition represents a conditional indirect association and does not imply that the overall transparency–trust association is moderated by AI literacy. The index of moderated mediation, not the significance of any single conditional indirect effect, is the appropriate test of whether the conditional indirect associations differ (Hayes, 2022; Igartua & Hayes, 2021; Preacher et al., 2007).

H5: The indirect association between perceived AI marketing transparency and brand trust through perceived manipulative intent is conditional on AI literacy, becoming more negative at higher levels of AI literacy.

The Dual-Pathway Model

Taken together, signaling theory and the Persuasion Knowledge Model provide a basis for conceptualizing two theoretically opposing interpretive pathways. In the trust-building pathway, perceived AI marketing transparency is positively associated with brand trust because it communicates openness and accountability, consistent with the credibility-based reading of signals in marketing (Erdem & Swait, 1998). In the suspicion-related pathway, perceived transparency is associated with perceived manipulative intent when consumers interpret the communication as a persuasive tactic; because perceived manipulative intent is negatively associated with brand trust, this pathway may counteract the favorable trust-related association. AI literacy is proposed as a boundary condition in the latter pathway. The two pathways are theoretically opposing and were examined through separate empirical models; their relative magnitude is not jointly estimated in the present design. These associations should not be read as demonstrating that transparency necessarily reduces trust among highly AI-literate consumers.

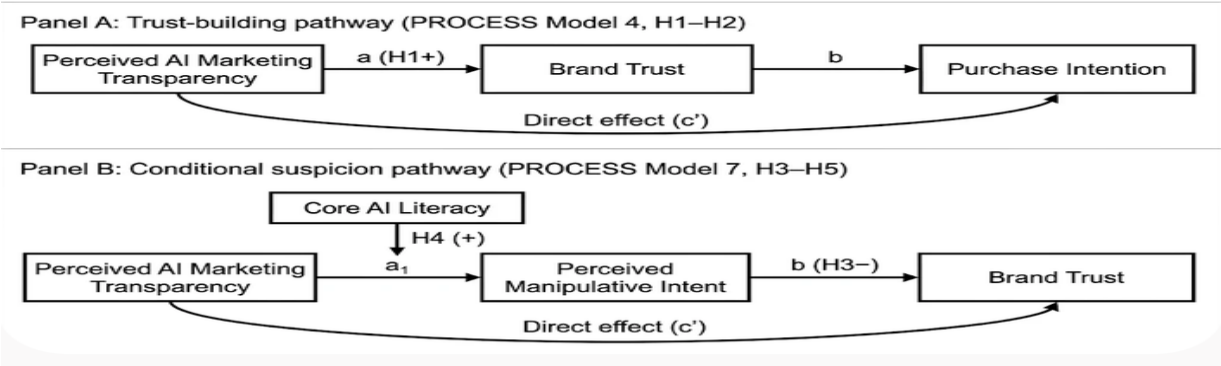

Figure 1. Dual interpretive pathways of perceived AI marketing transparency

Note. The figure does not show causal effects, but rather associations. The paths of Panel A and Panel B were estimated separately (Panel A in Model 4 and Panel B in Model 7) and therefore cannot be compared. The pathway of trust building is represented by H1, and H3, H4 and H5 are the conditional suspicion pathway. The two pathways are in theory opposing, and were explored separately by empirical models. There is no interaction between AI Literacy and Transparency which is hypothesized to predict brand trust directly in the model. The solid arrows are hypothesized to represent the paths in the model; the moderation arrow in Panel B is the interaction effect.

Method

Design, Measurement Approach, and Sample

A two-wave survey design was used with a two-week washout period in between. Wave 1 assessed AI literacy and individual characteristics of the respondents, such as socio-demographic factors and consumer experiences relevant to AI. Wave 2 captured perceived AI marketing transparency, perceived manipulative intent, brand trust and purchase intention. The temporal separation decreases, or removes, immediate consistency pressures and separates the moderator and controls from the focal perceptual evaluations. It does not imply temporal precedence since all of the constructs of focus were measured in Wave 2. Nor does time between events remove the possibility of common-method variance among those four constructs. The design is an observational design and does not tell about causal effects.

Perceived AI marketing transparency was measured as respondents' perceptions of the openness and clarity with which a focal brand communicates its use of AI in its marketing. The construct reflects perceived transparency rather than an objective audit of disclosure practices, and should not be interpreted as an assessment of disclosure compliance or independently verified disclosure quality. No standardized disclosure stimulus was manipulated. Respondents were asked to think of the brand from which they had most recently encountered AI-recommended product content on an e-commerce platform or social media, and to answer all items with that brand in mind.

To ensure that respondents had encountered relevant brand communication, a screening item at Wave 1 asked whether they had seen any information from the brand about its use of AI. Of the 519 respondents who began Wave 1, 47 were excluded because they had not seen any information from the focal brand about its use of AI. This yielded 472 Wave 1 respondents. The analytical sample therefore represents consumers who had encountered both AI-recommended content and some AI-related brand communication; it does not represent all consumers exposed to AI recommendations. This screening is acknowledged as a limitation in Section 5.4.

Data was collected using a non-probability online sampling approach. Of the 472 Wave 1 respondents, 268 were recruited through Qualtrics online panels and 204 through social media channels (Facebook and Instagram brand-community groups). The two recruitment sources may differ in ways relevant to the study variables; this is acknowledged as a limitation and examined in sensitivity analyses reported in Section 4.5. Wave 1 respondents were drawn from five urban centres Lahore (n = 131), Karachi (n = 124), Islamabad (n = 97), Rawalpindi (n = 68), and Faisalabad (n = 52). The final matched analytical sample comprised 387 respondents. The matched sample's city distribution was Lahore (n = 107), Karachi (n = 101), Islamabad (n = 79), Rawalpindi (n = 56), and Faisalabad (n = 44). This distribution did not differ significantly from that of the Wave 1 sample ($\chi^2(4) = 1.23$, $p = .87$). Eligibility required respondents to be aged 18–40 and to have encountered AI-generated or AI-recommended brand content on e-commerce platforms or social media in the preceding six months. Exposure was verified by a screening item asking respondents to identify the platform on which they had most recently encountered AI-recommended product content; respondents who could not identify a platform were excluded from Wave 1 and are not included in the 472.

Duplicate submissions were prevented through Qualtrics duplicate-entry detection, IP address screening, and a check that no two submissions shared the same device fingerprint. Device fingerprinting was implemented through Qualtrics' built-in duplicate detection, which records browser and device characteristics (e.g., user agent, screen resolution) without collecting personally identifying information. Panel members were provided with standard panel compensation while social media respondents were put into a draw for five shopping vouchers worth PKR 5,000, which took place after the close of Wave 2.

Responses were linked across waves employing a pseudonymous matching code. The code was the first two letters of the respondent's mother first name and the end two digits of their cell phone number. The code was kept separate from survey responses and was only used to link Wave 1 to Wave 2 records and was deleted once the linking was done, it was password protected and only accessible by the lead researcher. The procedure does not include anonymous collection of data but pseudonymous data processing. The analytic data collection did not contain any names, full mobile numbers or other direct identifiers.

The number of Wave 1 respondents who completed Wave 2 and were able to be matched was 387 out of 472, which represents an 82% retention rate. Independent-samples t tests and chi-square tests were used to compare the 85 respondents who were lost to follow-up with the 387 respondents who were retained by age, gender, education, experience with e-commerce, and experience with AI literacy. No significant differences emerged on any of these variables (age: $t(470) = 0.84$, $p = .40$; gender: $\chi^2(1) = 0.52$, $p = .47$; education: $\chi^2(4) = 2.31$, $p = .68$; e-commerce experience: $t(470) = 1.12$, $p = .26$; AI literacy: $t(470) = 0.73$, $p = .47$). The final analytical sample was thus 387.

The sample comprised 54.3% male, with a mean age of 27.8 years (SD = 5.42), with 61.2% having a bachelor's degree or higher. Mean e-commerce experience was 4.6 years (SD = 2.31); and prior scam experience was reported by 51.7%. As with the Visa (2026) study, a significant share of the sample said that they were using AI tools for shopping-related tasks, but this was not measured separately.

Participation was voluntary and informed consent was obtained at the beginning of Wave 1 with a consent screen which asked for a yes/no before proceeding with the survey. Respondents were informed of the study's academic purpose, that participation could be withdrawn at any point without consequence, that responses would be reported only in aggregate, and that data would be stored on a password-protected institutional server accessible only to the research team. No personally identifying information was retained after matching was completed. The matching procedure and privacy safeguards were reviewed and approved by the ethics committee.

With N = 387 and the number of predictors in the interaction model, the achieved sample provides power of approximately 0.56 to detect an interaction effect of $f^2 = 0.01$ at $\alpha = 0.05$ (two-tailed), based on post hoc sensitivity analysis. The study was therefore underpowered for a small interaction. The sample size was set at the upper bound feasible within the eight-week collection window and was not determined by a prospective power calculation for the interaction. This limitation is discussed in Section 5.4.

Measures

All constructs were measured with multi-item, five-point Likert scales (1 = strongly disagree, 5 = strongly agree) unless otherwise noted. Full item wording appears in Appendix A.

Perceived AI marketing transparency (five items) was measured using author-developed items informed by signaling theory (Erdem & Swait, 1998) and prior brand transparency research. The items measure how open the respondents think the focal brand is to using AI in marketing content, recommendations, and personalization ($\alpha = 0.84$; CR = 0.86; AVE = 0.55). The scores are higher, the more open the perception of using AI.

Perceived manipulative intent (four items) was assessed based on items developed by the authors based on the Persuasion Knowledge Model (Friestad & Wright, 1994) and previous research on inferences of manipulative intent (Campbell, 1995). The items measure the belief that the brand's marketing communication using AI is designed to persuade the respondents in a strategic manner ($\alpha = 0.82$; CR = 0.84; AVE = 0.57). Items refer to the brand's marketing communication about AI, rather than the AI itself. The higher scores the more one thinks that the strategy has a higher purpose.

Brand trust (five items) was adapted from the integrative trust framework (Mayer et al., 1995), which includes the three facets of ability, benevolence, and integrity, with the wording of the items made suitable for the online consumer–brand context ($\alpha = 0.88$; CR = 0.89; AVE = 0.60). Scores range from higher to greater trust.

The purchase intention items (three retained) were modified from the standard purchase intention items used in AI marketing research in Pakistan (Ahmed et al., 2026) which included purchase likelihood, purchase intention, and comparative choice intention. A fourth item measuring willingness to recommend was removed for the reasons described in Section 4.1. No retained item measures recommendation ($\alpha = 0.78$; CR = 0.80; AVE = 0.47).

AI literacy was assessed with the 34-item MAILS instrument (Carolus et al., 2023), measured on an 11-point scale (0 = not at all true of me, 10 = completely true of me). The MAILS instrument comprises nine dimensions: (a) use and apply AI, (b) know and understand AI, (c) detect AI, (d) AI ethics, (e) Create AI, (f) AI self-efficacy problem-solving, (g) AI self-efficacy learning, (h) AI self-competency (persuasion literacy), and (i) AI emotion regulation. Following the original MAILS factor structure, the primary AI literacy moderator was operationalized using the four core AI literacy dimensions use and apply AI, know and understand AI, detect AI, and AI ethics which comprise 18 items. These four dimensions constitute the AI literacy construct as originally conceptualized and are distinguished from Create AI, AI self-efficacy, and AI self-competency. The 18-item core AI literacy scale had adequate reliability ($\alpha = 0.87$; CR = 0.88; AVE = 0.50). The full 34-item composite was computed for sensitivity analysis ($\alpha = 0.91$; CR =

0.92; AVE = 0.52) and is reported in Section 4.5. The correlation between the 18-item core measure and the 34-item composite was r = 0.96 (p < .001), indicating that the two scoring approaches share substantial variance. The original MAILS response format is an 11-point scale from 0 to 10; the present study used this original format.

A three-item persuasion knowledge activation measure was developed by the authors with reference to the Persuasion Knowledge Model (Friestad & Wright, 1994) and the accessibility and capacity framework of Campbell and Kirmani (2000), and included for supplementary analysis (α = 0.71); its validity evidence is limited and it is treated as exploratory.

Controls were age (years), gender (0 = female, 1 = male), education (1 = secondary or below to 5 = postgraduate), e-commerce experience (years), prior scam experience (0 = no, 1 = yes), and product category involvement (three items developed by the authors following the product involvement conceptualization of Mittal, 1995; α = 0.81). All six are included in the models reported in Tables 2 and 3.

Analytical Strategy

Three analytical components were estimated, relating to different hypotheses.

Simple mediation (H1 and H2). Association between perceived AI marketing transparency and purchase intention in terms of brand trust was estimated by PROCESS Model 4 (Hayes, 2022) with 5000 bootstrap resamples. H1 was also tested using a hierarchical regression (Table 2).

Direct association in conditional process model (H3). Perceived manipulative intent was modeled as a mediator on the path to brand trust in PROCESS Model 7. First-stage moderated mediation (H4 and H5). The conditional indirect association of perceived transparency with brand trust through perceived manipulative intent, moderated by the 18-item core AI literacy measure at the first stage, was estimated using PROCESS Model 7, with 5,000 bootstrap resamples.

Coefficient reporting. All PROCESS coefficients entering the conditional process calculations (interaction, conditional slopes, the mediator-to-outcome path, conditional indirect effects, and the index of moderated mediation) are reported as unstandardized coefficients, because the arithmetic of the index requires consistent units. Standardized coefficients are reported separately for the hierarchical regressions (Tables 2 and 3) and labelled as such. Tables 2 and 3 contain standardized coefficients only for reference purposes, and should not be used for arithmetic. Tables 5 contains unstandardized coefficients. The means of perceived AI marketing transparency and AI literacy were centered prior to the creation of the interaction term. The perceived manipulative intent and brand trust were not standardized, and thus the conditional effects are in the original metric of the predictors.

Common method variance. Procedural remedies consisted of confidentiality assurances, voluntary participation and temporal separation of waves. Harman's single factor test was carried out as the four perceptual constructs measured in one wave, the first unrotated factor explained 41.8% of the total variance. This outcome is only descriptive of the data structure and doesn't confirm the absence of common-method variance. Common-method concerns are thus still present and form a limitation in Section 5.4. The VIF values were all less than 2.2 suggesting that there was no multicollinearity among the predictors.

Results

Measurement Model

A hypothesized 5 factor structure was tested with confirmatory factor analysis (CFA). Perceived transparency, perceived manipulative intent, brand trust, and purchase intention were specified with their individual items as indicators (five, four, five, and three items, respectively). Following the original MAILS structure, the four core dimensions constitute the construct of AI literacy, with AI literacy specified as a first order factor with the 18 core AI literacy items as indicators. This gave 35 variables that were observed. The estimator used was maximum likelihood with robust standard errors (MLR) and the responses across the 11-point MAILS were assumed continuous. Fit was acceptable: $\chi^2(544) = 892.47$, $p < 0.001$; CFI = 0.912; TLI = 0.903; RMSEA = 0.041 (90% CI [0.036, 0.046]); SRMR = 0.052. The measurement model exhibited good fit.

Four items had standardized loadings of less than 0.70 (between 0.61 and 0.69). All constructs had composite reliability scores greater than 0.78, which is acceptable. The perceived transparency, perceived manipulative intent, brand trust and core AI literacy had adequate convergent validity (AVE>0.5). The AVE for purchase intention (0.47) was below the traditional 0.50 value and the composite reliability (0.80) was greater than 0.70. This limitation was noted and the construct was retained. Discriminant validity was confirmed with square roots of the AVEs being higher than the inter-construct correlations. The trust – purchase intention

correlation (r = 0.43) was the highest correlation among the various constructs. Furthermore, discriminant validity was explored with regard to heterotrait–monotrait (HTMT) ratios, which were all below the maximum recommended cutoff value of 0.85, with the lowest value being 0.72 for trust–purchase intention. The conceptual difference between perceived transparency and perceived manipulative intent was also explored at the item level: transparency items focus on perceived openness, while the perceived manipulative intent items focus on inferred strategic motive. These are conceptually different, but need to be explored further empirically in future research.

The standardized loading for the recommendation item was 0.47, and it was removed from the purchase intention scale because its conceptual meaning is not related to purchase intention. The three items that remain were retained during the measurement evaluation process and thus the retained scale does not have as broad a content domain as the original scale, an acknowledgment that is included in Section 5.4.

Table 1. Measurement properties and correlations

Construct	Items	M	SD	α	CR	AVE	1	2	3	4	5
1. Perceived AI marketing transparency	5	3.36	0.82	0.84	0.86	0.55	**0.74**				
2. Perceived manipulative intent	4	2.94	0.88	0.82	0.84	0.57	0.09	**0.76**			
3. Brand trust	5	3.17	0.77	0.88	0.89	0.60	0.27**	−0.31**	**0.77**		
4. Purchase intention	3	3.02	0.85	0.78	0.80	0.47	0.19**	−0.22**	0.43**	**0.69**	
5. Core AI literacy (18 items)	18	5.82	1.52	0.87	0.88	0.50	0.15**	−0.03	−0.02	−0.01	**0.71**

*Notes. N = 387. Diagonal values in bold are √AVE. AI literacy was measured on an 11-point scale (0–10); all other constructs on 5-point scales. **p < 0.01 (two-tailed).*

Two features of Table 1 shape later interpretation. The perceived transparency–perceived manipulative intent correlation was small and non-significant (r = 0.09, ns), consistent with the argument that this association is conditional rather than a main effect. Core AI literacy was uncorrelated with brand trust (r = −0.02, ns) and purchase intention (r = −0.01, ns) despite correlating with exposure to transparency (r = 0.15, p < 0.01). This pattern is consistent with the moderation hypothesis that AI literacy does not have a main effect on trust or purchase intention but may moderate the association between transparency and suspicion-related evaluations.

Direct and Mediated Associations (H1, H2)

H1. After controlling for age, gender, education, e-commerce experience, prior scam experience, and product category involvement, perceived AI marketing transparency was positively associated with brand trust ($\beta = 0.24$, $SE = 0.07$, $t = 3.42$, $p < 0.001$, 95% CI [0.10, 0.38]). H1 was confirmed. Trust was negatively related to prior experience with scams (β = −0.13, p = 0.031); other controls were non-significant.

Table 2. Hierarchical regression predicting brand trust (standardized coefficients)

Predictor	Model 1 (controls)	Model 2a (transparency only)	Model 2b (transparency + PMI)
Age	−0.05	−0.04	−0.04
Gender	0.03	0.02	0.02
Education	0.06	0.05	0.05
E-commerce experience	0.05	0.04	0.04
Prior scam experience	−0.14*	−0.13*	−0.13*
Product category involvement	0.04	0.03	0.03
Perceived AI marketing transparency (AIT)		0.24***	0.21**

Predictor	Model 1 (controls)	Model 2a (transparency only)	Model 2b (transparency + PMI)
Perceived manipulative intent (PMI)			−0.29***
R^2	0.04	0.10	0.19
ΔR^2		0.06***	0.09***
F	2.19*	6.04***	11.46***

*Notes. N = 387. Standardised coefficients. Model 2a provides the H1 estimate (β = 0.24). Model 2b is a descriptive hierarchical specification including both predictors simultaneously; it is not the formal mediation test. The coefficient for transparency differs between Models 2a and 2b (0.24 vs. 0.21) because the specifications differ, and this reduction does not by itself establish mediation. These standardised coefficients are not used in the conditional process arithmetic. F values are overall model Fs. *p < 0.05. **p < 0.01. ***p < 0.001.*

H2. The formal mediation test used PROCESS Model 4. Perceived transparency was positively associated with brand trust (a = 0.23, SE = 0.07, t = 3.42, p < 0.001), and brand trust was positively associated with purchase intention (b = 0.26, SE = 0.06, t = 4.33, p < 0.001). The total association between perceived transparency and purchase intention was positive and significant (c = 0.16, SE = 0.06, p = 0.008). The indirect relationship through brand trust was statistically significant (indirect effect = 0.06, SE = 0.02, 95% bootstrap CI [0.02, 0.11]), while the direct relationship between perceived AI marketing transparency and purchase intention was not statistically significant after controlling for brand trust (c′ = 0.10, SE = 0.06, p = 0.081). The results suggest significant indirect association with non-significant direct association. This pattern is consistent with full mediation, but a nonsignificant direct effect does not demonstrate that the direct association is exactly zero. The indirect effect accounted for approximately 37.5% of the estimated total association (0.06/0.16) but should be interpreted with caution as the proportion is descriptive, and the total association is relatively small while the direct association is non-significant, thus making the proportion unstable. H2 was supported.

AI Literacy as Moderator (H4)

There was a positive and significant interaction of transparency × core AI literacy predicting perceived manipulative intent (unstandardised B = 0.13, SE = 0.06, t = 2.17, p = 0.031, 95% CI [0.01, 0.25]). The interaction accounted for an additional 1.2% of the variance in perceived manipulative intent (ΔR2 = 0.012; F (1, 377) = 4.71, p = 0.031). The degrees of freedom were calculated as N - k - 1 = 387 - 9 - 1 = 377, with k being the number of predictors (six controls, transparency, AI literacy, and interaction term). H4 was supported but the incremental contribution was minimal.

Table 3. Hierarchical regression predicting perceived manipulative intent (standardised betas)

Predictor	Model 1 (controls)	Model 2 (main effects)	Model 3 (interaction)
Age	0.02	0.02	0.02
Gender	−0.03	−0.03	−0.03
Education	0.04	0.04	0.04
E-commerce experience	0.03	0.03	0.03
Prior scam experience	0.09	0.09	0.09
Product category involvement	0.02	0.02	0.02
Perceived AI marketing transparency (AIT)		0.06	0.06
AI literacy (AIL)		−0.02	−0.01
AIT × AIL			0.13*

Predictor	Model 1 (controls)	Model 2 (main effects)	Model 3 (interaction)
R^2	0.03	0.040	0.052
ΔR^2		0.010	0.012*
F	1.79	1.97	2.30*

*Notes. N = 387. Standardised coefficients reported for the hierarchical regression. F values are overall model Fs. Model 2 R^2 = 0.040; Model 3 R^2 = 0.052; ΔR^2 for the interaction = 0.012. The F-change for the interaction is F(1, 377) = 4.71, p = 0.031. The standardised interaction coefficient in this table and the unstandardised PROCESS interaction coefficient in Table 5 are both reported as 0.13 after rounding; they are distinct quantities derived from different model specifications. *p < 0.05. Predictors mean-centred before forming the interaction term.*

Table 4. Conditional association of perceived AI marketing transparency with perceived manipulative intent at levels of AI literacy (unstandardised)

AI literacy level	Unstandardized B	SE	T	p	95% CI
Low (−1 SD)	−0.13	0.07	−1.86	0.064	[−0.26, 0.00]
Mean	0.06	0.05	1.20	0.231	[−0.04, 0.16]
High (+1 SD)	0.25	0.07	3.57	< 0.001	[0.11, 0.39]

Notes. N=387. PROCESS Model 7. Unstandardised Coefficients. These are the coefficients for computing the conditional indirect associations in Table 6. The transparency → PMI path in Table 5 (a1 = 0.06) and the conditional slope at the mean (0.06) are the same estimate and reported similarly in both the tables. The exact interaction coefficient for these conditional slopes is about 0.125 (rounding to the reported 0.13) and the standard deviation of AI literacy is 1.52 (Table 1). Thus, the difference between the high and low slopes (0.25 − (−0.13) = 0.38) is exactly 2 × 0.125 × 1.52.

The conditional association was negative and non-significant when AI literacy was low (B = −0.13, p = 0.064), positive and non-significant when AI literacy was at the mean (B = 0.06, p = 0.231), and positive and significant when AI literacy was high (B = 0.25, p < 0.001). The association between perceived transparency and perceived manipulative intent therefore became more positive as AI literacy increased, and it was statistically distinguishable from zero only at the high level.

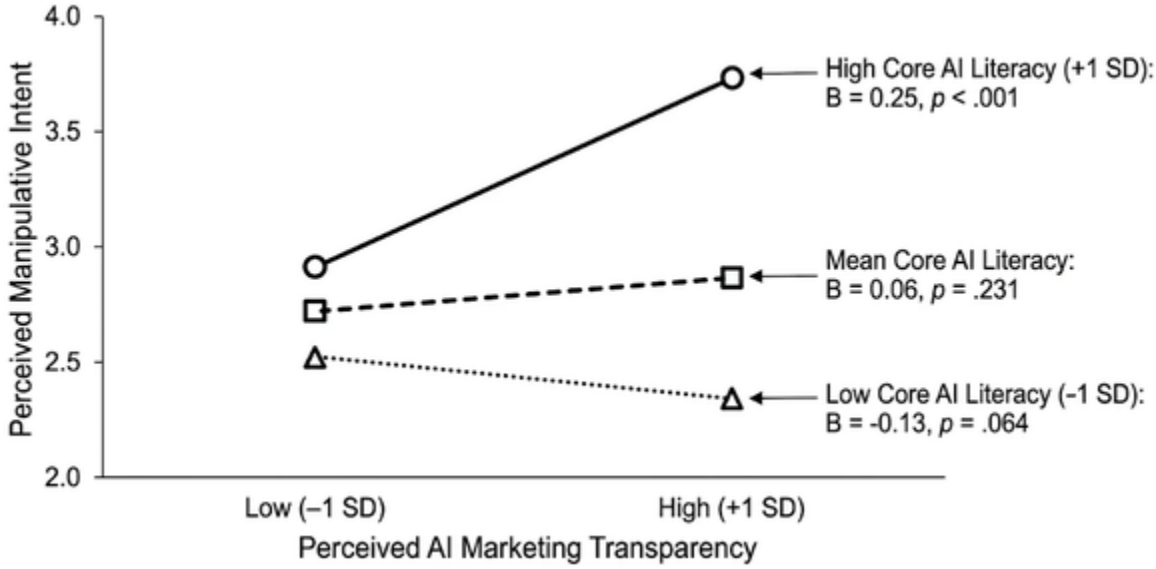

Figure 2. Conditional effect of perceived AI marketing transparency on perceived manipulative intent across levels of core AI literacy

Note. N = 387. Unstandardized conditional slopes derived from PROCESS Model 7. Low and high levels of core AI literacy represent -1 SD and 1 SD from the mean, respectively. Perceived AI marketing transparency is positively associated with perceived manipulative intent only at high levels of core AI literacy (β = 0.25, p < .001$)

Conditional Indirect Association (H5)

The conditional indirect association was positive and non-significant at low AI literacy (effect = 0.03, Boot SE = 0.03, 95% CI [−0.01, 0.09]), negative and non-significant at mean AI literacy (effect = −0.02, Boot SE = 0.02, 95% CI [−0.05, 0.01]), and negative and significant at high AI literacy (effect = −0.06, Boot SE = 0.03, 95% CI [−0.13, −0.01]).

The moderated mediation index was -0.03, 95% bootstrap CI [-0.07, -0.001]. The index of moderated mediation was statistically significant at the 5% level, but the confidence interval barely excluded zero. The result should therefore be interpreted with caution, especially in view of the small interaction effect and the limited statistical power. There was statistical support for H5, but the evidence is best characterised as fragile. The fact that the indirect effect is significant when conditional on high literacy does not in and of itself represent moderated mediation.

Table 5. Unstandardized PROCESS coefficients used in the conditional process calculations

Model and parameter	Unstandardized B	SE	p / 95% CI
Model 4 (H2): transparency → trust → purchase intention			
Transparency → brand trust (a)	0.23	0.07	p < 0.001
Brand trust → purchase intention (b)	0.26	0.06	p < 0.001
Indirect effect (a × b)	0.06	0.02	[0.02, 0.11]
Direct effect (c′)	0.10	0.06	p = 0.081
Model 7 (H3, H4, H5): transparency → PMI → brand trust, moderated at first stage			
Transparency → PMI at mean AIL (a_1)	0.06	0.05	p = 0.231
AIL → PMI (a_2)	−0.01	0.05	p = 0.842
Transparency × AIL (a_3)	0.13	0.06	p = 0.031, [0.01, 0.25]
PMI → brand trust (b)	−0.25	0.05	p < 0.001
Transparency → brand trust (c′)	0.21	0.06	p = 0.001
Index of moderated mediation (a_3 × b)	−0.03	0.02	[−0.07, −0.001]

Notes. N = 387, 5,000 bootstrapped re-samples. The two models answer different questions. Their routes are not comparable. The trust-mediated relationship with purchase intention is estimated in Model 4; the conditional suspicion-related relationship with brand trust is estimated in Model 7. All coefficients are in unstandardised form. Here the PMI → brand trust path is shown unstandardised. The standardised coefficient in Table 2 is of similar numerical value, but represents a different quantity, and can't be used in the arithmetic. The arithmetic is 3 × b = 0.13 × − 0.25 = − 0.0325 which rounds nicely to the reported index of − 0.03. Confidence intervals from bootstrapping are not reproducible from point estimates alone.

Table 6. Conditional indirect effect of perceived AI marketing transparency on brand trust through perceived manipulative intent (unstandardised)

AI literacy level	Indirect effect	Boot SE	95% bootstrapped CI
Low (−1 SD)	0.03	0.03	[−0.01, 0.09]
Mean	−0.02	0.02	[−0.05, 0.01]
High (+1 SD)	−0.06	0.03	[−0.13, −0.01]

Notes. N = 387; 5,000 bootstrapped resamples. Unstandardized coefficients. A negative indirect effect indicates that perceived transparency is associated with higher perceived manipulative intent, which in turn is associated with lower brand trust. These are conditional associations, not causal effects. The arithmetic follows from Tables 4 and 5: at high literacy, 0.25 × −0.25 = −0.0625 ≈

−0.06; at low literacy, −0.13 × −0.25 = 0.0325 ≈ 0.03; at the mean, 0.06 × −0.25 = −0.015 ≈ −0.02. The bootstrap confidence intervals were obtained from the resampling procedure and cannot be derived from these products.

The conditional indirect association was statistically distinguishable from zero only at high AI literacy. The corresponding estimates at low and mean AI literacy were not statistically significant; these results should not be interpreted as proof that the indirect association is absent at those levels. The effect size at high literacy is small and the practical implication is that there is a suspicion-related pathway among the most AI-literate respondents, not that the trust-building association disappears or reverses.

Supplementary and Exploratory Analyses

To check that the results were consistent regardless of the alternative scoring of the MAILS instrument, we repeated the moderated mediation analysis with the total score on the MAILS (comprising the 34 items) as the moderator. The correlation between the 18-item core measure and the 34-item combined measure was r = 0.96 ($p < .001$), suggesting that the two scoring methods share a substantial amount of variance, and thus the sensitivity analysis is a limited test of the operationalization issue. The interaction remained positive and significant (B = 0.12, SE = 0.06, t = 2.00, p = 0.046, 95% CI [0.002, 0.24]), and the index of moderated mediation remained statistically significant at the 5% level, although its confidence interval narrowly excluded zero (index = −0.03, 95% CI [−0.07, −0.002]). The pattern of conditional indirect associations was also similar, with positive and non-significant conditional indirect associations at low literacy (effect = 0.03, 95% CI [−0.01, 0.09]) and negative and non-significant conditional indirect associations at mean literacy (effect = −0.02, 95% CI [−0.05, 0.01]), and negative and significant conditional indirect associations at high literacy (effect = −0.06, 95% CI [−0.14, −0.01]). The robustness of the results to the additional MAILS dimensions is limited, given that the two dimensions are highly correlated, but nevertheless these results are robust. The 18-core AI literacy measure is the key measure for this test, as it adheres more closely to the original MAILS factor structure.

There was a significant positive correlation between persuasion knowledge activation and core AI literacy (r = 0.28, $p < 0.01$), but it was not empirically identical. The interaction between transparency and persuasion knowledge was significant (B = 0.18, t = 2.52, p = 0.012, $\Delta R2 = 0.017$) and had a larger $\Delta R2$ than the core AI literacy interaction ($\Delta R2 = 0.012$). The interaction between transparency and persuasion knowledge activation led to a larger incremental R^2 in the supplementary model than the interaction between transparency and AI literacy.The descriptive difference indicates that persuasion knowledge activation could be a more proximal construct, although the descriptive difference was not formally compared in a joint model. The discovery is preliminary. Due to the limited validity evidence and only three items in the measure ($\alpha = 0.71$), these results cannot be used in place of the main model.

The interaction was significant when respondents reported having been a victim of a scam in their prior lives (B = 0.16, p = 0.048, n = 200) and non-significant when they did not report prior scam experience (B = 0.09, p = 0.26, n = 187). If there is a significant group difference in one subgroup, but not in the other, this does not imply a difference in the group differences, it requires a formal cross-group interaction test. The results presented here are exploratory in nature and should not be viewed as providing evidence of moderation of the relationship.

We analyzed whether there was a difference between Qualtrics panel respondents (n = 219) and social media respondents (n = 168). The interaction was significant in the panel subsample, but not in the social media subsample. Because of limited power, a formal cross-group comparison was not performed, and these findings were descriptive, and should be interpreted with caution.

The achieved sample led to about 0.56 power to detect an interaction of $f^2 = 0.01$, as mentioned in Section 3.1. The value of H4 and H5 could be due to inadequate power rather than a small true effect; the null low-literacy slope should not be interpreted as absence of an effect. However, this power constraint makes it impossible to rule out a chance finding from the results of this study, so the result should be regarded as preliminary.

Summary of Hypothesis Tests

Table 7. Summary of hypothesis tests

Hypothesis	Path	Estimate (type)	SE p / 95% CI	Decision
H1	AIT → Brand trust	β = 0.24 (std.)	0.07 < 0.001	Supported

Hypothesis	Path	Estimate (type)	SE	p / 95% CI	Decision
H2	AIT → Trust → PI (Model 4)	0.06 (unstd.)	0.02	[0.02, 0.11]	Supported (significant indirect association; direct effect non-significant, p = 0.081)
H3	PMI → Brand trust	b = −0.25 (unstd.)	0.05	< 0.001	Supported
H4	AIT × AIL → PMI (Model 7)	B = 0.13 (unstd.)	0.06	0.031	Supported (small effect)
H5	Index of moderated mediation	−0.03 (unstd.)	0.02	[−0.07, −0.001]	Supported (statistically significant but borderline)

Notes. AIT = perceived AI marketing transparency; AIL = AI literacy; PMI = perceived manipulative intent; PI = purchase intention. For the hierarchical regressions, standardised (std.) coefficients are reported, and for the conditional process analyses, unstandardised (unstd.) coefficients are reported. H2 and H5 use bootstrapped confidence intervals (5,000 resamples). H5 is described as statistically significant but borderline because the confidence interval narrowly excludes zero.

Discussion

Overview of Findings

This study explored the relationship between perceived AI marketing transparency and brand trust and purchase intention, proposed a mechanism as perceived manipulative intent, and proposed a boundary condition as core AI literacy. The results are consistent with two theoretically competing pathways, examined in different empirical models. The trust building pathway was analyzed using PROCESS Model 4, while the conditional suspicion related pathway was analyzed using PROCESS Model 7. Note that these pathways were estimated independently, and so do not represent an independent estimation of the decomposition of the overall transparency-trust relationship.

The trust-building pathway revealed that perceived AI marketing transparency had a positive association with brand trust (β = 0.24, $p < 0.001$), which directly showed an indirect association with purchase intention (indirect effect = 0.06, 95% CI [0.02, 0.11]). The direct association was not statistically significant after accounting for the mediation role of brand trust ($p = 0.081$), confirming full mediation. In the suspicion-related pathway, brand trust was negatively related to perceived manipulative intent ($b = -0.25$, $p < 0.001$); at the higher levels of core AI literacy, the indirect association between perceived manipulative intent and transparency and trust became more negative and the index of moderated mediation was statistically significant at the 5% level but just narrowly missed zero (index = −0.03, 95% CI [−0.07, −0.001]). The same pattern was obtained from the sensitivity analysis with the full 34 item MAILS composite, but again because of the high correlation between the two scoring procedures ($r = 0.96$), this sensitivity analysis is limited.

Theoretical Contributions

The study helps to alert researchers and practitioners to the perspectives on AI for marketing by distinguishing between three stages that are often mixed up (Connelly et al., 2011; Spence, 1973). Perceived transparency is a receiver-centered construct that refers to how the receiver interprets the actual disclosure and does not indicate whether the disclosure is costly, verifiable or complete. The results reinforce the notion that perceived transparency can be linked to positive and negative evaluations based on receiver characteristics and suggest that perceived transparency need not be the same as objectively credible transparency. This is a contribution that is specific to marketing because the boundary condition resides in the consumers' marketplace knowledge (Erdem & Swait, 1998).

The study identifies core AI literacy operationalized using the 18-item core dimensions of the MAILS instrument as a potential boundary condition in the association between perceived AI marketing transparency and perceived manipulative intent (B = 0.13, $p = 0.031$; $\Delta R^2 = 0.012$). The observed interaction suggests that consumers' AI-related knowledge may be relevant to how transparency is interpreted. The contribution is preliminary: the interaction is small, the index of moderated mediation is statistically significant but borderline (index = −0.03, 95% CI [−0.07, −0.001]), and the mechanism through which AI-related knowledge would translate into suspicion of strategic intent is not directly tested. Importantly, the core AI literacy dimensions

capture technological knowledge and competency rather than persuasion-related awareness, which makes them a more appropriate boundary condition for testing whether general AI literacy conditions the transparency–suspicion association. The sensitivity analysis using the full 34-item composite produced a similar interaction (B = 0.12, p = 0.046) and a similar index of moderated mediation (index = −0.03, 95% CI [−0.07, −0.002]), although the high correlation between the two measures (r = 0.96) limits the in formativeness of this robustness check. The contribution should therefore be understood as evidence that core AI literacy competencies condition the transparency–suspicion association, with the full composite producing a similar pattern.

The supplementary analysis suggests that persuasion knowledge activation may exhibit a larger interaction with perceived transparency than core AI literacy (B = 0.18, p = 0.012, ΔR^2 = 0.017, versus B = 0.13, p = 0.031, ΔR^2 = 0.012). The two measures, while related, correlated at r = 0.28, and the constructs produced different empirical patterns, consistent with the argument that a focused measure of awareness of persuasive intentions captures a somewhat different consumer response than a broad AI literacy composite. However, the difference between the two interaction effects was not formally tested in a joint model and the supplementary measure requires further validation. These preliminary results suggest that context-specific persuasion knowledge may be more salient to consumers' interpretation of AI-related disclosure than general AI literacy. This suggestion, however, should be seen as exploratory.

The main theoretical contribution of the paper is the coexistence of a positive transparency–trust association (β = 0.24) and a conditional negative indirect association through perceived manipulative intent (index = −0.03). Transparency should not be thought of in terms of good or bad perceptions. Consumers may associate transparency with openness and evaluate the communicated communication against the persuasive goals of the brand. Two qualifications tied this contribution. First, the pathways were estimated in separate models, rather than in a single integrated structural model. This means that the relative magnitude of the pathways should not be interpreted as a jointly estimated decomposition of the overall association of transparency with trust. The two pathways are best described as theoretically opposing pathways, examined in separate empirical models. Second, the suspicion pathway is about how disclosure is read (not how consumers believe the brand's underlying AI practices), and the measure captures only the former.

Thus, the interpretation of AI-related brand communication is particularly relevant in the Pakistani market, which is a setting where AI-assisted shopping and concerns about transactional trust co-exist. The study did not assess institutional trust and verification mechanisms and hence cannot claim that these contextual conditions account for the observed moderation. Future research could directly investigate institutional trust, perceived regulatory credibility, and scam exposure as contextual or individual level boundary conditions.

The results are not against Qadri et al. (2025). This study is focused on online literacy and the relationship between perceived manipulation and unethical behaviour. The current study is focused on core AI literacy and the relationship between perceived transparency and perceived manipulative intent. Online literacy and AI literacy are different constructs, and the two moderation mechanisms are not directly comparable. The results are in line with the research on AI-based marketing and young consumers (Waqas & Qadri, 2026) which emphasises the role of consumer autonomy and ethical assessment in AI-mediated marketplace encounters. The contradictory finding from Ali and Sunela (2026) that trust did not mediate the transparency–purchase intention relationship highlights the importance of examining conditional pathways rather than assuming a uniform transparency–trust association.

Practical Implications

First, transparency should communicate meaningful information about AI use rather than generic disclosure statements. Clear explanations of how AI contributes to customer-facing activities may help consumers understand the role of technology in brand interactions.

Second, the results indicate that the marketer needs to take into account whether the AI disclosures are easily comprehensible and meaningful to AI-savvy and other consumer groups. While overall, perceived transparency was positively correlated with brand trust, the conditional suspicion pathway indicated that higher core AI literacy could potentially be a factor in consumers perceiving AI-related disclosure as a strategic, persuasive communication. This does not mean that organizations should decrease transparency to the AI-literate consumers, nor does this current study suggest a segmentation approach to AI-literacy. Instead, disclosures should include meaningful explanations of how AI is used, why data-driven personalization is used, and how much automated systems have an impact on any decisions made for customers. Before the segmentation approach can be recommended generally for management, it should be experimented.

Third, transparency is not sufficient to foster trust. Disclosure should come with information on the purpose of the use of AI, their respective consumer protection protocols, and how much AI is used in the decisions made for their customers. The present results should not be used to justify lack of transparency for highly AI-literate consumers; they call for a more general focus on credible and context-sensitive communication with AI.

Limitations

This study did not employ an experimental design that had standard conditions for disclosure. These findings are for association between perceived constructs and do not imply the causal effects of AI marketing transparency, nor do they prove the effects of disclosure manipulation on consumers.

At Wave 2, we measured perceived transparency, perceived manipulative intent, brand trust, and purchase intention. While separating waves decreased immediate consistency stress between AI literacy and the focal perceptions, it did not create a temporal precedence between the main explanatory and outcome constructs. Common-method issues are still possible, and the Harman's test result (41.8%) does not exclude these issues.

A small interaction ($f^2 = 0.01$) was under powered and the study had a power of about 0.56. Statistically significant results (H4, $p = 0.031$; H5, index 95% CI [−0.07, −0.001]) may be due to lack of statistical power, rather than to a small true effect. The study is unable to rule out the possibility of a chance finding. It must be understood that the non-significant low-literacy slope does not mean the absence of the low literacy.

The main AI literacy moderator was measured with the 18-item core AI literacy dimensions in the original structure of the MAILS instrument. The operationalization is used to differentiate the core AI literacy dimensions from Create AI, AI self-efficacy, and AI self-competency. A sensitivity analysis was conducted using the full 34-item composite, and was similar, although the measures are highly correlated ($r = 0.96$). Future studies need to explore the results when AI literacy is measured through other types of scoring, and should provide the complete specification of the measure, including the item-level indicators included.

The item for recommendation was deleted during item evaluation of the purchase intention construct, and the AVE value was 0.47, which is below the conventional value. The three item scale retained is less comprehensive than the original instrument. The persuasion knowledge measure was poorly validated and only had three items ($\alpha = 0.71$), and the results from this measure are presented in exploratory manner.

The study did not rely on a probability-based national sampling frame, but instead recruited respondents with the help of online panels and social media. The two sources of recruitment can vary on the study variables and thus the results can be taken as evidences from the recruited urban consumers sample only and cannot be interpreted as national sample estimates of consumers in Pakistan.

Respondents who had experienced AI-recommended content without the information from the focal brand on AI content were not included. The analytical sample is thus representative of consumers who had seen AI-recommended content, and some AI-related brand communication, but is not representative of all consumers who encountered AI recommendations. This screening is recognized as a weakness.

The number of degrees of freedom reported in the measurement model and how the MAILS instrument is treated should be understood based on the model being analysed. For the purpose of the present analysis, core AI literacy was chosen as the first-order factor in the indata analysis, which had 18 items as indicators, resulting in 35 observed variables. Future work should include the full specification – including any item parceling or error covariance specification. There has been some other evidence of discriminant validity, including the maximum heterotrait–monotrait ratio (HTMT = 0.72).

The study was carried out among the Pakistani consumers aged 18-40 from 5 urban centers. Results might not be applicable to older consumers, rural consumers, other national contexts, or consumers who have different exposure to AI, digital infrastructure, and regulatory experience.

Estimates were made separately for the trust-building pathway (Model 4) and the conditional suspicion pathway (Model 7). Although their results are theoretically consistent, no single integrated structural model was estimated that allows a joint evaluation of the two pathways.

The study participants were prompted to reflect on a brand that they had come across with AI recommended information, and a screening question ensured that they had seen information about the brand's use of AI. This study also did not include a standardized disclosure stimulus, and the perceived transparency may be due to pre-exposure to the brand, as well as actual brand communication. Experimental designs with uniform disclosure stimuli could be utilized in future studies.

The matching code was a mix of a family member's name and digits from the respondent's cellphone. Though there was no complete anonymisation of personal data, the access was limited and after matching, the code was deleted. The procedure was reviewed and accepted by the ethics committee and the difference between confidentiality, pseudonymisation and anonymisation is recognized.

Future Research

Future research could extend the present study in four ways. Experimental designs could manipulate content, specificity, and credibility of AI marketing disclosures to establish causal effects and test whether explainable disclosure changes the pattern at high AI literacy. The integrated structural models could simultaneously estimate the accountability-related and suspicion-related pathways. Core AI literacy and persuasion knowledge should be analysed as independent moderators, with larger samples, validated multi-item measures, and pre-registered hypotheses. Researchers should also compare alternative scoring approaches for AI literacy including the 18-item core measure and the full 34-item composite to identify whether specific competencies are more closely linked to consumers' understanding of AI-related brand disclosure. Due to the high correlation between the 18-item and 34-item measures in the present study, future work should utilise measures that more clearly distinguish technological AI literacy from persuasion-related awareness. Generalisability would be established by replication across countries, industries and age groups, ideally with prospective power analysis calibrated for small interaction effects.

Conclusion

The current study investigated the relationships between perceived AI marketing transparency, perceived manipulative intent, brand trust, and purchase intention in Pakistan, focusing on core AI literacy as a possible boundary condition. AI marketing transparency was positively associated with brand trust and brand trust significantly mediated the relationship between transparency and purchase intention. Perceived manipulative intent was negatively related to brand trust. At high levels of core AI literacy, the indirect relationship between transparency and trust through perceived manipulative intent was more negative. The index of moderated mediation was significant at the 5% level but just barely missed zero indicating statistically significant but borderline support for the proposed conditional indirect association.

The findings suggest that perceived transparency of AI marketing may be associated with both positive trust evaluation and conditional suspicion-based evaluation. However, the weak interaction (ΔR 2 = 0.012) and the statistically fragile mediated moderation evidence suggest caution in interpretation. The study does not suggest that transparency decreases trust among consumers with high AI literacy. Instead, it emphasises a conditional indirect path that can weaken the positive association between transparency and trust. The study disentangles perceived transparency from its interpretation and explores key competencies of AI literacy and a specific measure of persuasion knowledge activation, providing a basis for future research on consumer evaluation of AI-related brand communication. Future experimental and longitudinal studies are needed to test the temporal and causal mechanisms underlying these associations, to validate the measures of AI literacy and perceived manipulative intent, and to assess whether the conditional indirect association is robust across consumer groups and national contexts.

References

1. Ahmed, A., Khan, I. U., Farooq, U., & Haque, A. N. (2026). The impact of content quality, authenticity, and AI transparency on purchase intention: The mediating role of brand perception and trust. *Journal of Arts & Social Sciences*, 13(1), 125–133. https://doi.org/10.46662/jass.v13i1.590

2. Ali, M. B., & Sunela, S. F. (2026). AI-driven purchase intention in Pakistan's e-commerce: The roles of personalization, anthropomorphism, usefulness, transparency, trust, and AI anxiety. *Journal of Social and Organizational Matters*, 5(2), 44–71. https://doi.org/10.56976/jsom.v5i2.431

3. Campbell, M. C. (1995). When attention-getting advertising tactics elicit consumer inferences of manipulative intent: The importance of balancing benefits and investments. *Journal of Consumer Psychology*, 4(3), 225–254. https://doi.org/10.1207/s15327663jcp0403_02

4. Campbell, M. C., & Kirmani, A. (2000). Consumers' use of persuasion knowledge: The effects of accessibility and cognitive capacity on perceptions of an influence agent. *Journal of Consumer Research*, 27(1), 69–83. https://doi.org/10.1086/314309

5. Carolus, A., Koch, M. J., Straka, S., Latoschik, M. E., & Wienrich, C. (2023). MAILS Meta AI literacy scale: Development and testing of an AI literacy questionnaire based on well-founded competency models and psychological change- and meta-competencies. *Computers in Human Behavior: Artificial Humans*, 1(2), 100014. https://doi.org/10.1016/j.chbah.2023.100014

6. Chaudhuri, A., & Holbrook, M. B. (2001). The chain of effects from brand trust and brand affect to brand performance: The role of brand loyalty. *Journal of Marketing*, 65(2), 81–93. https://doi.org/10.1509/jmkg.65.2.81.18255

7. Connelly, B. L., Certo, S. T., Ireland, R. D., & Reutzel, C. R. (2011). Signaling theory: A review and assessment. *Journal of Management*, 37(1), 39–67. https://doi.org/10.1177/0149206310388419

8. Erdem, T., & Swait, J. (1998). Brand equity as a signaling phenomenon. *Journal of Consumer Psychology*, 7(2), 131–157. https://doi.org/10.1207/s15327663jcp0702_02

9. Friestad, M., & Wright, P. (1994). The persuasion knowledge model: How people cope with persuasion attempts. *Journal of Consumer Research*, 21(1), 1–31. https://doi.org/10.1086/209380

10. Hayes, A. F. (2022). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach* (3rd ed.). Guilford Press.

11. Igartua, J.-J., & Hayes, A. F. (2021). Mediation, moderation, and conditional process analysis: Concepts, computations, and some common confusions. *The Spanish Journal of Psychology*, 24, e49. https://doi.org/10.1017/SJP.2021.46

12. Kim, M. (2026). When AI says who made it: A meta-analysis of AI disclosure effects on consumer responses. *Computers in Human Behavior Reports*. https://doi.org/10.1016/j.chbr.2026.101345

13. Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *Academy of Management Review*, 20(3), 709–734. https://doi.org/10.5465/amr.1995.9508080335

14. Mittal, B. (1995). A comparative analysis of four scales of consumer involvement. *Psychology & Marketing*, 12(7), 663–682. https://doi.org/10.1002/mar.4220120708

15. Preacher, K. J., Rucker, D. D., & Hayes, A. F. (2007). Addressing moderated mediation hypotheses: Theory, methods, and prescriptions. *Multivariate Behavioral Research*, 42(1), 185–227. https://doi.org/10.1080/00273170701341316

16. Qadri, U. A., Moustafa, A. M. A., & Abd Ghani, M. (2025). They misused me! Digital literacy's dual role in AI marketing manipulation and unethical young consumer behavior. *Young Consumers*, 26(5), 744–764. https://doi.org/10.1108/YC-08-2024-2207

17. Spence, M. (1973). Job market signaling. *Quarterly Journal of Economics*, 87(3), 355–374. https://doi.org/10.2307/1882010

18. Visa. (2026, June 9). *Visa study: 82% of Pakistanis are shopping with AI, but trust is key at checkout*. Visa Middle East. https://pk.visamiddleeast.com/about-visa/newsroom/press-releases/prl-09062026.html

19. Waqas, M., & Qadri, U. A. (2026). When consumer autonomy is compromised: Adopting a dual-type view to understand AI-driven marketing and its impact on young consumers' ethical behavior. *Journal of Business Ethics*, 206(2), 237–255. https://doi.org/10.1007/s10551-025-06155-x

Appendix A. Measurement items

All items were measured on a five-point Likert scale (1 = strongly disagree, 5 = strongly agree) unless otherwise indicated.

A.1 Perceived AI marketing transparency (5 items)

1. This brand clearly explains in its marketing communications that it uses AI.
2. This brand is open about how AI is used to generate the recommendations I see.

3. This brand tells customers when the content they are viewing has been created or optimized using AI.
4. Information about this brand's use of AI is easy to find in its communications.
5. This brand does not hide the role that AI plays in its marketing.

A.2 Perceived manipulative intent (4 items)

1. This brand's statements about its use of AI are intended to persuade me rather than to inform me.
2. This brand makes its AI use visible in order to make me view it more favorably.
3. This brand uses its disclosure of AI use as a marketing tactic.
4. This brand's communication about AI seems designed to shape my impression of the brand.

A.3 Brand trust (5 items)

1. This brand is competent at delivering what it promises. (ability)
2. This brand has the knowledge and resources to serve its customers well. (ability)
3. This brand acts in its customers' best interests. (benevolence)
4. This brand is honest with its customers. (integrity)
5. I can rely on this brand to keep its promises. (integrity)

A.4 Purchase intention (3 retained items)

1. I would be willing to buy from this brand.
2. I intend to purchase from this brand in the future.
3. I would choose this brand over comparable alternatives.

The recommendation item ("I would recommend this brand to friends") was removed during measurement evaluation because its standardized factor loading was 0.47 and because recommendation intention reflects advocacy-related behavior conceptually distinct from purchase intention. The retained scale measures purchase intention only; no retained item measures recommendation.

A.5 AI literacy (MAILS, 34 items)

AI literacy was assessed using the 34-item Meta AI Literacy Scale (MAILS; Carolus et al., 2023), measured on an 11-point scale (0 = not at all true of me, 10 = completely true of me). All 34 items were administered using the published item wording and response format. The instrument includes nine dimensions: AI literacy competencies (use and apply AI, know and understand AI, detect AI, and AI ethics), Create AI, AI self-efficacy (problem-solving and learning), AI self-competency (persuasion literacy), and AI emotion regulation. The component groupings and representative items are as follows.

Core AI literacy (18 items; primary moderator). This domain includes four components: (a) use and apply AI (e.g., "I can use AI-based applications to accomplish tasks"); (b) know and understand AI (e.g., "I understand what AI can and cannot do"); (c) detect AI (e.g., "I can distinguish between AI-supported applications and conventional applications"); and (d) AI ethics (e.g., "I can evaluate the ethical implications of AI-based applications"). These four dimensions constitute the AI literacy construct as originally conceptualized and are used as the primary moderator in the present study. $\alpha = 0.87$; CR = 0.88; AVE = 0.50.

Create AI (4 items; separate construct). Representative item: "I can use AI-based tools to create content or solve problems."

AI self-efficacy (6 items; separate construct). This domain includes problem-solving and learning components (e.g., "I am confident that I can learn to use new AI applications"; "I can solve problems that arise when using AI-based applications").

AI self-competency (6 items; separate construct). This domain includes persuasion literacy (e.g., "I can recognize when AI-based systems are trying to persuade me") and emotion regulation (e.g., "I can regulate negative feelings about developments in artificial intelligence"; "I am not unsettled when I encounter news about AI developments").

Scoring. The 18 core AI literacy items were scored according to the instrument's published procedure and used as the primary moderator in the PROCESS analyses. The full 34-item composite was computed for sensitivity analysis. The correlation between the 18-item core measure and the 34-item composite was $r = 0.96$ ($p < .001$). In the CFA, core AI literacy was specified as a first-order factor with the 18 items as indicators. $\alpha = 0.87$ (18-item core); $\alpha = 0.91$ (34-item composite); CR = 0.88 (18-item core); CR = 0.92 (34-item composite); AVE = 0.50 (18-item core); AVE = 0.52 (34-item composite).

A.6 Persuasion knowledge activation (3 items, supplementary)

1. I recognize when a brand is trying to persuade me.
2. I can tell when a brand's communication is intended to influence my opinion.
3. I understand the tactics that brands use to persuade consumers.

A.7 Product category involvement (3 items, control)

1. The product category I was thinking about is important to me.
2. I care about the kind of product I was thinking about.
3. I pay attention to information about this product category.



2026 by the authors; Journal of Global Social Transformation (JGST). This is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC-BY) license (http://creativecommons.org/licenses/by/4.0/).