Journal of Sustainable Development and Policy

Volume: 16; **Issue:** 2
Pages: 01-40

The Impact of AI-Driven Information Technology on Social Media Communication and Consumer Trust

Md Hasinur Rahman[1]; Tasnim Binte Afsar[2]; Md Shah Alam[3]

1. *PhD in information Technology, University of the Cumberland, USA*
   M.A. in Professional Communication, University of San Francisco, USA
   M.A. in English Language & Literature, Dhaka International University, Bangladesh
   B.A. (Honors) in English Literature, Edward College, National University, Bangladesh
   Email: rahmanhasinur10@gmail.com

2. *Master in Business Analytics Management, Avila University, Kansas City Missouri, USA*
   Bachelor of Business Administration in Finance, United International University, Dhaka, Bangladesh
   Email: tasnim.afsar3@gmail.com

3. *Master in Business Administration (MBA) Mercy University USA*
   Master in Business Analytics, Jagannath University Bangladesh
   Bachelor of Business Studies Jagannath University Bangladesh
   Email: pappurony2020@gmail.com

Doi: 10.63125/2jq2bx46
Received: *22 November 2025;* ***Revised:*** *06 December 2025;* ***Accepted:*** *02 January 2026;* ***Published:*** *03 February 2026;*

Abstract

Artificial intelligence now mediates most brand communication on social media: recommendation models decide what consumers see, generative models write part of it, conversational agents answer them, classifiers moderate them and labels mark what is synthetic. This article asks how these AI-driven information technologies affect consumer trust and what a survey study of the standard design can learn about it. It develops the Algorithmic Communication Trust Framework, in which five exposure constructs (personalised recommendation, AI-generated content, conversational agents, algorithmic moderation and AI labelling) act on three perceptions of AI-mediated communication (communication quality, authenticity, and manipulation and surveillance), which act on trust, which carries their effects to engagement and purchase intention, with AI literacy and privacy concern as moderators. Because no survey observes the true effect, the framework is evaluated on a transparent synthetic testbed: a pre-specified data-generating model with parameters anchored to published findings, ordinal measurement through 48 Likert items and a method factor. The pipeline of a recent template study (reliability, factor analysis, Kruskal-Wallis tests, hierarchical regression, bootstrapped mediation and moderation) is applied to synthetic samples of 130, 500 and 2,000 and repeated in a Monte Carlo evaluation. At n = 500, thirteen of eighteen directional hypotheses are recovered with the designed signs: the three perceptions are the strongest predictors of trust, trust mediates ten of eleven indirect paths, synthetic content lowers trust through two routes and labelling raises it through both. Coefficients are attenuated by 15 to 30% and interactions by more; at n = 130 only ten hypotheses survive, significant estimates are exaggerated up to 2.7-fold and the factor structure replicates in 36% of samples; interactions of 0.10 need 1,000 to 2,000 respondents;

Keywords

Artificial Intelligence; Social Media; AI-Mediated Communication; Consumer Trust; Personalization; Generative Content; Chatbots; Privacy; Survey Methodology; Simulation

Introduction

Social media communication between brands and consumers now passes through artificial intelligence at almost every step. Recommendation and ranking models decide which posts, stories and advertisements a consumer sees; generative models write captions, draft replies and produce images and video; conversational agents answer questions in direct messages and comment threads; classifiers remove, down-rank or flag content; and provenance systems and platform labels mark some of what remains as synthetic (Kaplan & Haenlein, 2019; Huang & Rust, 2021; Hancock et al., 2020; Gorwa et al., 2020). These AI-driven information technologies change the three things that communication theory says matter for trust: who the consumer believes is speaking, whether the message was shaped for them, and what was done with their data to shape it (Sundar, 2020; Hancock et al., 2020; Acquisti et al., 2015). The practitioner evidence is that consumers notice. Public-opinion surveys report that a majority of Americans are more concerned than excited about the role of AI in daily life, that comfort with AI-generated news is low in several countries, and that trust in institutions and media is contested (Tyson & Kikuchi, 2023; Fletcher & Nielsen, 2024; Newman et al., 2024; Edelman, 2025), while experimental studies find that disclosing a chatbot reduces purchases, that content labelled as AI-generated is believed less even when true, and that perceived surveillance erodes the benefit of personalisation (Luo et al., 2019; Altay & Gilardi, 2024; Longoni et al., 2022; Aguirre et al., 2015).

The academic record on these effects is uneven. Experiments on single mechanisms (chatbot disclosure, AI labels, algorithmic transparency) are rigorous but narrow; survey studies of consumers on social media are broad but typically small, cross-sectional and built on composite Likert scores analysed with non-parametric tests and regression, so that the sizes they report are hard to compare and their power to detect the small effects that field experiments find is rarely examined (Srivastava & Gurme, 2026; DellaVigna & Linos, 2022; Kelly et al., 2023). A recent study in this journal's field, which this article takes as its structural template, surveyed 130 digitally active consumers, confirmed that personalisation and chatbot interaction raise trust and that privacy concern and perceived surveillance lower it, and argued that ethical and privacy concerns are direct determinants of trust rather than moderators (Srivastava & Gurme, 2026). Its design is representative of the literature: a theory-driven framework (the technology acceptance model, the elaboration likelihood model and the privacy calculus), a validated questionnaire, reliability and factor analysis, Kruskal-Wallis tests and multiple regression, and a hypothesis summary. What such a design cannot show is how far its estimates are from the truth, because the truth is never observed.

This article asks what AI-driven information technology does to social media communication and consumer trust, and what a survey study of the standard design can and cannot recover about it. Because the true effect of an AI feature on a population of consumers is never observed, and because no survey of a few hundred respondents can be rerun thousands of times, the article builds a transparent synthetic testbed: a pre-specified data-generating model in which seven AI exposure and disposition constructs act on three perceptions of AI-mediated communication (communication quality, authenticity, and manipulation and surveillance), which act on consumer trust, which acts on engagement and purchase intention, with two moderators, demographic shifts, ordinal measurement through 48 Likert items and a common-method factor. The model's parameters are anchored to published findings and documented as assumptions; synthetic respondents are generated from it; the full analysis pipeline of the template study is applied to a primary sample of 500, a replication-size sample of 130 and a large sample of 2,000; and a Monte Carlo layer repeats the pipeline hundreds of times at each sample size and under alternative assumptions to score what it recovers. All respondents are synthetic. No person was surveyed, and every result is a property of stated assumptions. The research questions are: RQ1, to what extent do AI-driven information technologies (personalised recommendation, AI-generated content, conversational agents, algorithmic moderation and AI labelling) affect consumer trust through perceived communication quality, authenticity and manipulation? RQ2, do privacy concern and perceived manipulation act as direct determinants of trust or as moderators, and does AI literacy change how consumers respond to AI-generated content? RQ3, does trust mediate the effects of AI-mediated communication on engagement and purchase intention? RQ4, what do reliability analysis, exploratory factor analysis, Kruskal-Wallis tests, hierarchical regression, bootstrapped mediation and moderation analysis recover at the sample sizes this literature

uses, and under non-normal ordinal data, measurement error and common-method bias?

The contributions are an integrative framework for trust in AI-mediated brand communication that extends trust theory, the technology acceptance model, the elaboration likelihood model and the privacy calculus with the machine heuristic and the AI-mediated communication perspective (Figure 1); a reference map of the AI-driven information technology stack in social media communication and its governance (Figure 2); a reproducible synthetic-respondent testbed and a 48-item instrument with a pre-specified analysis plan (Section 4); a complete empirical demonstration of the template study's pipeline on the synthetic samples, including seven hypotheses tested with a hypothesis summary; a Monte Carlo evaluation of the pipeline's power, attenuation, false positives and factor recovery across sample sizes and assumptions; and a regulatory mapping and roadmap for trustworthy AI-driven communication. The remainder of the article is organised as follows. Section 2 reviews the literature and develops the framework. Section 3 states the hypotheses. Section 4 describes the methodology: the research design, the synthetic sample, the instrument and the analysis plan. Section 5 presents the results, including the Monte Carlo evaluation. Section 6 discusses theoretical, managerial and policy implications and limitations, and Section 7 concludes.

Literature review

AI mechanisms in social media communication

Social media communication was defined, before the current generation of AI, by its functional building blocks: identity, conversations, sharing, presence, relationships, reputation and groups (Kietzmann et al., 2011; Kaplan & Haenlein, 2010). AI-driven information technology acts on each block through five mechanisms. Personalised recommendation ranks content, creators and advertisements by predicted engagement and value, which raises relevance and, at the same time, makes the consumer aware that behaviour is being modelled (Wedel & Kannan, 2016; Huang & Rust, 2021; Davenport et al., 2020; Dwivedi et al., 2021a; Dwivedi et al., 2021b; Boerman et al., 2017). Generative models produce or assist text, image, video and voice content for brands, creators and platforms, so that an increasing share of what consumers see is synthetic or AI-assisted, including virtual influencers and AI-written replies (Sands et al., 2022; Hancock et al., 2020; Hohenstein et al., 2023). Conversational agents handle service and sales conversations at scale, with effects on compliance, satisfaction and anger that depend on disclosure, anthropomorphism and the state of the consumer (Adam et al., 2021; Sheehan et al., 2020; Crolic et al., 2022; Luo et al., 2019; Mozafari et al., 2022). Algorithmic moderation removes, down-ranks and flags content on a scale that no human workforce could reach, with decisions that are opaque to the people affected (Gillespie, 2018; Gorwa et al., 2020). Labelling and provenance systems, including platform AI labels and content credentials, tell the consumer which content is synthetic, and are becoming a legal requirement in several jurisdictions (Coalition for Content Provenance and Authenticity, 2026; European Parliament and Council, 2024; California State Legislature, 2024). These mechanisms are the layer of the stack through which the consumer experiences AI (Figure 2), and they correspond to the exposure constructs of this study.

Communication theory supplies the lens for how these mechanisms are perceived. Social presence and media richness theory predict that richer, more responsive channels carry more trust-relevant cues (Short et al., 1976; Daft & Lengel, 1986); the hyperpersonal model predicts that computer-mediated messages can be perceived as more intimate than face-to-face ones because they are selectively composed (Walther, 1996); the computers-are-social-actors paradigm predicts that people apply social rules to machines that display social cues (Nass & Moon, 2000); and the AI-mediated communication perspective argues that when an intelligent agent modifies, augments or generates a message on a person's behalf, the receiver's attributions of authorship, intent and trustworthiness change (Hancock et al., 2020; Jakesch et al., 2019). Sundar's work on machine agency adds the machine heuristic, the belief that machines are objective and unbiased, which can raise credibility when AI is the perceived source and lower it when the AI is seen as manipulative (Sundar, 2020). Experiments confirm both directions: profile text believed to be AI-written is trusted less (the replicant effect), people's heuristics for detecting AI-generated language are flawed, AI-generated news text is credible enough to serve misinformation, news from generative AI is believed less, and headlines labelled as AI-generated are believed less even when true or human-made (Jakesch et al., 2019; Jakesch et al., 2023; Kreps et al., 2022; Longoni et al., 2022; Altay & Gilardi, 2024; Pennycook & Rand, 2021). Disclosure is therefore double-edged: a label

protects the consumer from deception but lowers the credibility of what it labels, a dilemma that audiences themselves recognise (Toff & Simon, 2025).

Consumer trust and psychology in AI-mediated environments

Trust is a willingness to be vulnerable to another party on the basis of positive expectations about its ability, integrity and benevolence (Mayer et al., 1995; Rousseau et al., 1998). Transferred to electronic commerce, trust has beliefs, intentions and disposition components, is validated in measurement typologies, and predicts purchase and continued use in integrated models with the technology acceptance model and perceived risk (McKnight et al., 2002; Gefen et al., 2003; Pavlou, 2003; Kim et al., 2008). Trust in technology differs from trust in people in the attributes it rests on (functionality, reliability and helpfulness rather than ability, integrity and benevolence), and the two can be measured side by side (Lankton et al., 2015). The review of human trust in AI finds that trust depends on tangibility, transparency, reliability, task characteristics and immediacy behaviours, that embodied and disembodied AI are trusted differently, and that trust develops over time (Glikson & Woolley, 2020; Jacovi et al., 2021). Recommendation agents gain trust through personalisation and familiarity and lose it when their reasoning is hidden; transparency and explanation raise trust up to a point and lower it when they overload the user (Komiak & Benbasat, 2006; Xiao & Benbasat, 2007; Kizilcec, 2016; Shin, 2021; Shin & Park, 2019; Rader et al., 2018). Consumers show algorithm aversion in some tasks and algorithm appreciation in others (Dietvorst et al., 2015; Logg et al., 2019; Castelo et al., 2019; Longoni et al., 2019), and their acceptance of automated decisions depends on fairness, usefulness and risk (Araujo et al., 2020; Kelly et al., 2023; Bedué & Fritzsche, 2022).

Privacy frames the cost side. The privacy calculus holds that people weigh the benefits of disclosure against perceived risk (Dinev & Hart, 2006; Culnan & Armstrong, 1999); information privacy concern is multidimensional (collection, control, awareness) and measurable (Malhotra et al., 2004; Smith et al., 2011); the privacy paradox, the gap between stated concern and behaviour, is well documented and partly explained by bounded rationality and context (Barth & de Jong, 2017; Acquisti et al., 2015). In advertising, the personalisation paradox is the finding that the same data collection that makes a message relevant also makes the consumer feel watched, so that personalised advertising works only when trust-building cues (transparency, control) accompany it (Aguirre et al., 2015; Bleier & Eisenbeiss, 2015; Awad & Krishnan, 2006; Boerman et al., 2017). The template study found privacy concern and perceived surveillance to be direct negative predictors of trust rather than moderators of the personalisation effect (Srivastava & Gurme, 2026), consistent with the view that surveillance is a cost in the consumer's calculus rather than a condition on the benefit (Zuboff, 2019). Persuasion theory completes the picture: the elaboration likelihood model predicts that relevant, personalised content is processed centrally and is more persuasive, that peripheral cues such as an AI label or a chatbot avatar change persuasion when elaboration is low, and that construal of an AI source changes how a message is evaluated (Petty & Cacioppo, 1986; Kim & Duhachek, 2020; Puntoni et al., 2021). Behavioural intentions, finally, follow attitudes and perceived control in the theory of planned behaviour and the unified theory of acceptance (Ajzen, 1991; Venkatesh et al., 2003; Venkatesh et al., 2012).

Re-evaluating the theories: the Algorithmic Communication Trust Framework

The theories above were built for human senders, static interfaces and one-time disclosures. AI-driven communication is adaptive and continuous: the recommendation that a consumer sees today is shaped by yesterday's behaviour, the reply a brand sends may be written by a model, and the label that discloses it changes how the next message is read. This article integrates the theories into an Algorithmic Communication Trust Framework (ACTF) with four propositions. First, the consumer's perceptions of AI-mediated communication form on three dimensions that the mechanisms act on differently: communication quality (relevance, responsiveness, richness), which personalisation and conversational agents raise; authenticity (whether the message is what it appears to be), which synthetic content lowers and labelling partly restores; and manipulation and surveillance (whether the consumer is being modelled and steered), which personalisation, synthetic content and moderation raise and transparency lowers. Second, trust in AI-mediated brand communication is a judgement of competence, integrity and benevolence made about the brand-and-algorithm assemblage, and it is formed by the three perceptions rather than by the mechanisms directly; the machine heuristic and the replicant effect operate through authenticity, the elaboration likelihood model through communication

quality, and the privacy calculus through manipulation and surveillance. Third, privacy concern and AI literacy are dispositions that enter both as antecedents of perception and as moderators of the perception-to-trust links: privacy concern sharpens the penalty for perceived manipulation, and AI literacy sharpens the penalty for undisclosed synthetic content because literate consumers detect and resent it. Fourth, trust mediates the effect of AI-mediated communication on behaviour (engagement, purchase and data sharing), so that the route from an AI feature to a business outcome runs through trust and is therefore reversible by harms to trust. Figure 1 draws the framework with the hypotheses of Section 3, and Table 1 summarises the literature strands and the gaps the framework addresses.

Figure 1: The Algorithmic Communication Trust Framework (ACTF): AI-driven information technology, perceptions of AI-mediated communication, consumer trust, behavioural intentions, moderators and hypotheses

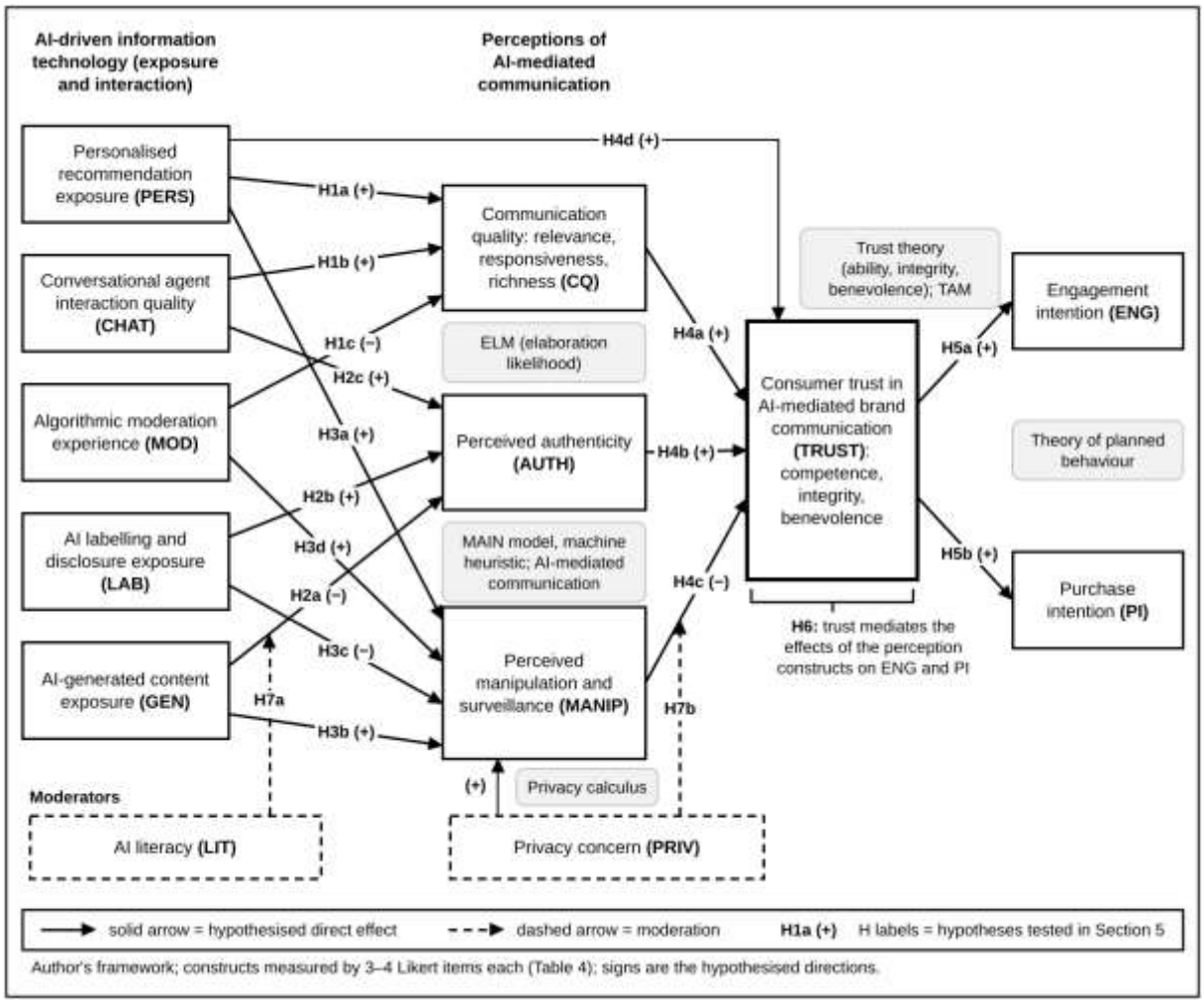


Table 1: Literature strands, what they establish, representative sources and the gap addressed in this article

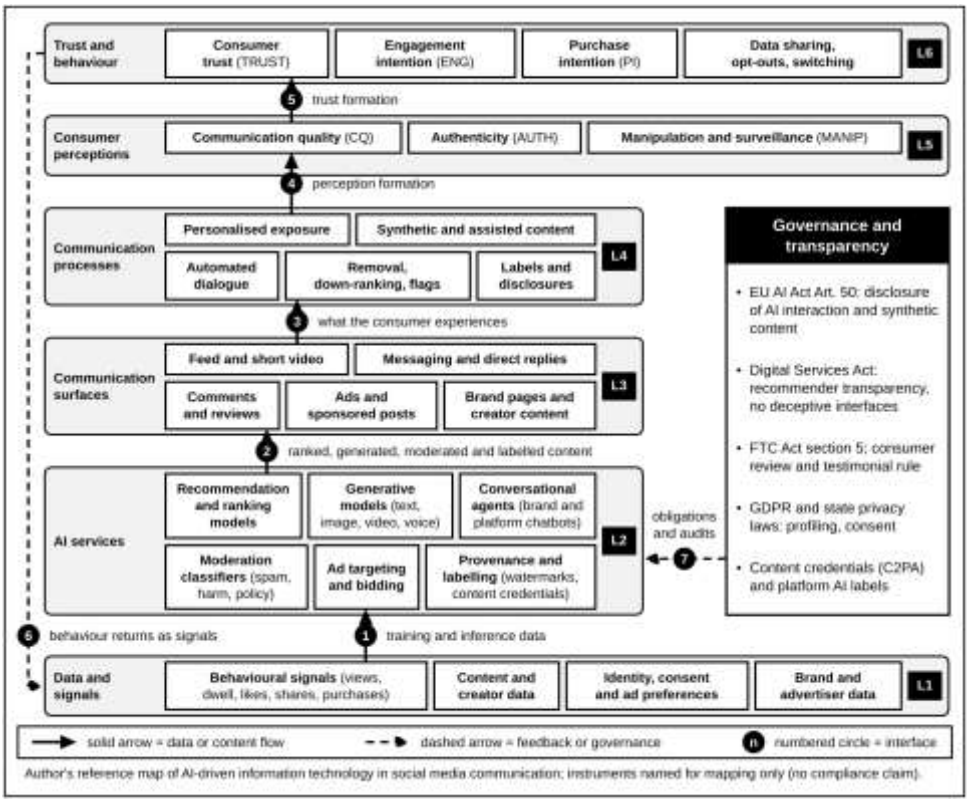
Strand	What the literature establishes	Representative sources	Gap addressed here
AI in marketing and social media	AI personalises, generates and automates brand communication at scale; strategic frameworks describe mechanical, thinking and feeling AI	Kaplan & Haenlein (2010, 2019); Huang & Rust (2021); Davenport et al. (2020); Dwivedi et al. (2021a, 2021b); Kietzmann et al. (2011)	Mechanisms separated (personalisation, generation, agents, moderation, labels) and their perceptual routes to trust compared on one population
AI-mediated communication and the machine heuristic	AI that generates or modifies messages changes attributions of authorship, intent and trustworthiness; machines are heuristically seen as objective yet AI-written text is trusted less	Hancock et al. (2020); Jakesch et al. (2019, 2023); Sundar (2020); Nass & Moon (2000); Hohenstein et al. (2023)	Authenticity modelled as a distinct perception with its own antecedents and a measured path to trust
AI-generated content, labels and misinformation	Generative content is believed less; labels lower credibility of what they mark but protect from deception; synthetic media raises uncertainty and lowers trust in news	Longoni et al. (2022); Altay & Gilardi (2024); Toff & Simon (2025); Kreps et al. (2022); Vaccari & Chadwick (2020); Pennycook & Rand (2021)	Labelling as an exposure construct with opposite effects on authenticity and manipulation; literacy as a moderator
Trust in technology and in AI	Trust rests on ability, integrity and benevolence; trust in AI depends on transparency, reliability, tangibility and task; explanations help up to a point	Mayer et al. (1995); McKnight et al. (2002); Lankton et al. (2015); Glikson & Woolley (2020); Shin (2021); Kizilcec (2016)	Trust in the brand-and-algorithm assemblage measured on the three attributes and placed as the mediator to behaviour
Chatbots, recommendation agents and algorithm aversion	Disclosure, anthropomorphism and competence decide chatbot effects; recommendation agents gain trust through personalisation; aversion and appreciation are task-dependent	Luo et al. (2019); Crolic et al. (2022); Adam et al. (2021); Komiak & Benbasat (2006); Dietvorst et al. (2015); Logg et al. (2019); Castelo et al. (2019)	Agent interaction quality as an antecedent of both communication quality and authenticity
Privacy calculus and the personalisation paradox	Disclosure is a benefit-risk trade-off; concern is multidimensional; personalisation works only with trust-building cues; surveillance is a direct cost	Dinev & Hart (2006); Malhotra et al. (2004); Acquisti et al. (2015); Barth & de Jong (2017); Aguirre et al. (2015); Bleier & Eisenbeiss (2015); Awad & Krishnan (2006)	Manipulation and surveillance as a perception with direct and moderated effects on trust, tested against the moderator reading
Survey methodology and credibility	Reliability, validity, common-method bias, ordinal data, mediation and moderation have established procedures; small samples exaggerate and miss effects	Podsakoff et al. (2003); Fornell & Larcker (1981); Hayes (2018); Preacher & Hayes (2008); Liddell & Kruschke (2018); Gelman & Carlin (2014); Morris et al. (2019)	The full pipeline scored against known truth at the sample sizes this literature uses
Regulation and provenance	Transparency for AI interaction and synthetic content, recommender transparency, consumer review rules and content credentials are now law or standard in several jurisdictions	European Parliament and Council (2016, 2022, 2024); Federal Trade Commission (2022, 2024); California State Legislature (2024); Coalition for Content Provenance and Authenticity (2026)	Instruments mapped to the constructs and to design and governance choices

Theoretical framework and hypotheses

Framework and justification

Figure 2 places the framework in the AI-driven information technology stack of a social media platform. Behavioural signals, content, identity and consent data (L1) feed AI services (L2): recommendation and ranking models, generative models, conversational agents, moderation classifiers, ad targeting and provenance and labelling. These services shape the communication surfaces (L3) the consumer uses and the communication processes (L4) the consumer experiences: personalised exposure, synthetic and assisted content, automated dialogue, removal and flags, and labels and disclosures. The processes form perceptions (L5) and the perceptions form trust and behaviour (L6), which return to the platform as signals. A governance panel lists the instruments that now constrain the stack: the EU AI Act's transparency obligations for AI interaction and synthetic content, the Digital Services Act's rules on recommender transparency and deceptive interfaces, the FTC Act and the consumer review rule, data protection law, and content credentials (European Parliament and Council, 2024; European Parliament and Council, 2022; U.S. Congress, 2023; Federal Trade Commission, 2024; European Parliament and Council, 2016; Coalition for Content Provenance and Authenticity, 2026). The framework's constructs are the consumer-side variables of L4 to L6 and the dispositions that condition them, measured as the consumer perceives them, which is the level at which survey studies of this kind operate and at which trust is formed.

Figure 2: AI-driven information technology stack in social media communication: data and signals, AI services, communication surfaces and processes, consumer perceptions, trust and behaviour, with numbered interfaces and a governance panel



The study models trust with three antecedent perceptions and five exposure constructs rather than with the two or three constructs of the template study, for three reasons. Communication quality, authenticity and manipulation are the dimensions along which the mechanisms differ: personalisation raises quality and manipulation together, synthetic content lowers authenticity and raises manipulation, labelling raises authenticity and lowers manipulation, and conversational agents raise quality and (when disclosed and competent) authenticity. Collapsing them into "AI personalisation" would hide the trade-offs that the design implications depend on. Second, the three perceptions map

one-to-one onto the theories: the elaboration likelihood model and media richness onto communication quality, the machine heuristic, the replicant effect and AI-mediated communication onto authenticity, and the privacy calculus and the personalisation paradox onto manipulation and surveillance. Third, the moderators of interest (privacy concern, AI literacy) act on specific perception-to-trust links, which can be tested only if the perceptions are measured separately.

Hypothesis development

Table 2 lists the hypotheses with their theoretical basis and expected sign. The first group concerns communication quality. Personalised recommendation raises the relevance of what a consumer sees and the responsiveness of the brand's communication, which the elaboration likelihood model and the personalisation literature link to engagement and persuasion (Petty & Cacioppo, 1986; Aguirre et al., 2015; Bleier & Eisenbeiss, 2015); conversational agents raise responsiveness and richness when they resolve requests (Adam et al., 2021; Sheehan et al., 2020); algorithmic moderation, by removing and down-ranking content and replies without explanation, lowers perceived quality for those who experience it (Gorwa et al., 2020; Gillespie, 2018). H1a: personalised recommendation exposure raises perceived communication quality. H1b: conversational agent interaction quality raises perceived communication quality. H1c: algorithmic moderation experience lowers perceived communication quality.

The second group concerns authenticity. Exposure to AI-generated content lowers perceived authenticity because consumers apply the replicant effect and believe such content less, whether or not they can detect it (Jakesch et al., 2019; Jakesch et al., 2023; Longoni et al., 2022); AI labels and content credentials raise perceived authenticity of the communication environment because they restore the consumer's ability to tell synthetic from human content, even though they lower the credibility of the labelled item (Altay & Gilardi, 2024; Toff & Simon, 2025; Coalition for Content Provenance and Authenticity, 2026); competent, disclosed conversational agents raise authenticity relative to undisclosed ones (Luo et al., 2019; Mozafari et al., 2022). H2a: AI-generated content exposure lowers perceived authenticity. H2b: AI labelling exposure raises perceived authenticity. H2c: conversational agent interaction quality raises perceived authenticity.

The third group concerns manipulation and surveillance. Personalisation makes the consumer aware of being modelled (Aguirre et al., 2015; Boerman et al., 2017); AI-generated content, social bots and synthetic media raise the sense of being steered (Ferrara et al., 2016; Vaccari & Chadwick, 2020); labelling and transparency lower it (Kizilcec, 2016; Rader et al., 2018; Zarouali et al., 2021); moderation experienced as opaque raises it (Eslami et al., 2015; Gorwa et al., 2020); and dispositional privacy concern raises it directly (Malhotra et al., 2004; Dinev & Hart, 2006). H3a: personalised recommendation exposure raises perceived manipulation and surveillance. H3b: AI-generated content exposure raises perceived manipulation and surveillance. H3c: AI labelling exposure lowers perceived manipulation and surveillance. H3d: algorithmic moderation experience raises perceived manipulation and surveillance.

The fourth group concerns trust. Trust in AI-mediated brand communication is formed by the three perceptions, with communication quality and authenticity as positive antecedents (competence and integrity) and manipulation and surveillance as a negative antecedent (benevolence) (Mayer et al., 1995; McKnight et al., 2002; Glikson & Woolley, 2020); personalisation retains a direct effect on trust beyond communication quality because relevance signals competence (Komiak & Benbasat, 2006; Srivastava & Gurme, 2026). Following the template study, perceived manipulation and privacy concern are modelled as direct determinants and tested as moderators in H7 rather than assumed to be moderators (Srivastava & Gurme, 2026). H4a: perceived communication quality raises trust. H4b: perceived authenticity raises trust. H4c: perceived manipulation and surveillance lowers trust. H4d: personalised recommendation exposure has a direct positive effect on trust beyond its effect through communication quality.

The fifth and sixth groups concern behaviour. Trust predicts engagement and purchase intention in trust-based models of electronic commerce and in the theory of planned behaviour (Gefen et al., 2003; Kim et al., 2008; Pavlou, 2003; Ajzen, 1991). H5a: trust raises engagement intention. H5b: trust raises purchase intention. H6: trust mediates the effects of communication quality, authenticity and manipulation on engagement and purchase intention, and the perceptions mediate the effects of the

exposure constructs on trust.

Table 2: Hypotheses, theoretical basis, expected sign and the test applied

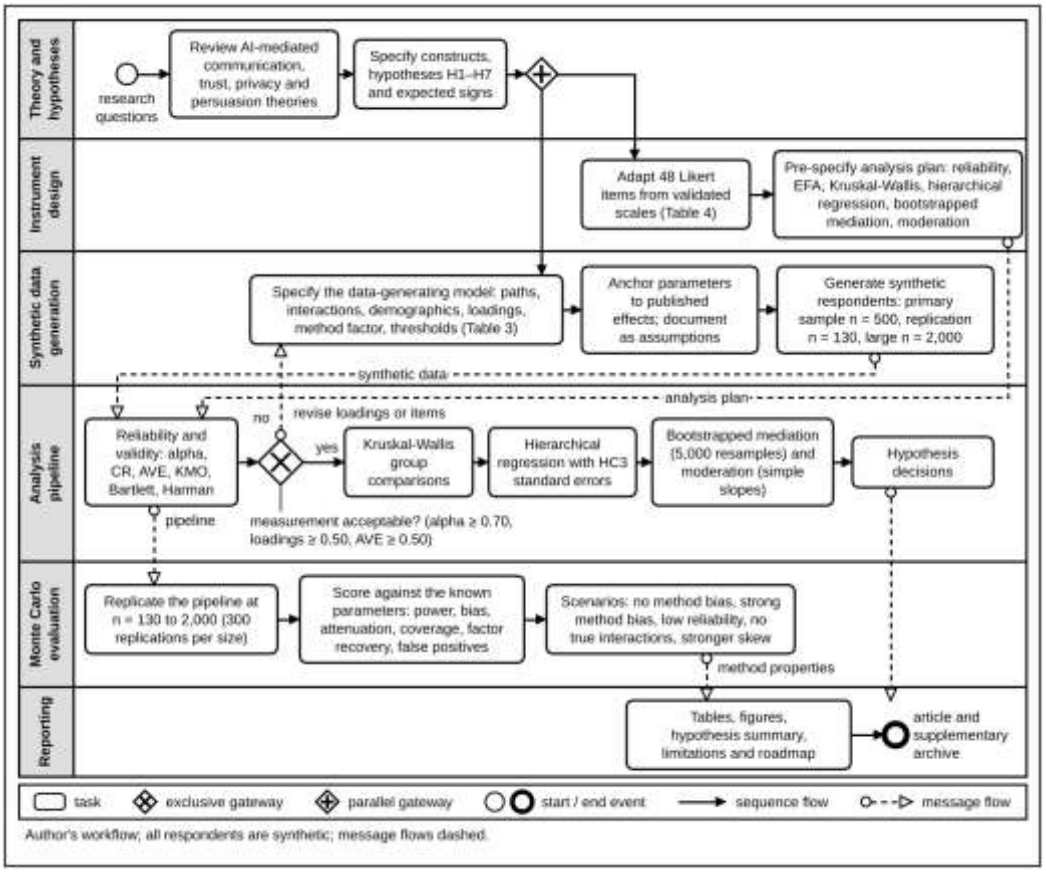
Hypothesis	Statement	Theoretical basis	Sign	Test
H1a	Personalised recommendation exposure raises perceived communication quality	Elaboration likelihood; personalisation and relevance	+	Regression of CQ on exposures; hierarchical step 3 via CQ
H1b	Conversational agent interaction quality raises perceived communication quality	Social presence; chatbot service quality	+	Regression of CQ on exposures
H1c	Algorithmic moderation experience lowers perceived communication quality	Opaque moderation; algorithmic affordance	-	Regression of CQ on exposures
H2a	AI-generated content exposure lowers perceived authenticity	AI-mediated communication; replicant effect	-	Regression of AUTH on exposures; Kruskal-Wallis by exposure tertile
H2b	AI labelling exposure raises perceived authenticity	Disclosure and provenance	+	Regression of AUTH on exposures
H2c	Conversational agent interaction quality raises perceived authenticity	Disclosed, competent agents	+	Regression of AUTH on exposures
H3a	Personalised recommendation exposure raises perceived manipulation and surveillance	Personalisation paradox	+	Regression of MANIP on exposures; Kruskal-Wallis by exposure tertile
H3b	AI-generated content exposure raises perceived manipulation and surveillance	Synthetic media and social bots	+	Regression of MANIP on exposures
H3c	AI labelling exposure lowers perceived manipulation and surveillance	Transparency and control	-	Regression of MANIP on exposures
H3d	Algorithmic moderation experience raises perceived manipulation and surveillance	Invisible algorithms	+	Regression of MANIP on exposures
H4a	Perceived communication quality raises trust	Competence; elaboration likelihood	+	Hierarchical regression step 3 (HC3)
H4b	Perceived authenticity raises trust	Integrity; machine heuristic	+	Hierarchical regression step 3 (HC3)
H4c	Perceived manipulation and surveillance lowers trust	Benevolence; privacy calculus (direct determinant)	-	Hierarchical regression step 3 (HC3)
H4d	Personalised recommendation exposure has a direct positive effect on trust beyond communication quality	Recommendation agents and competence signals	+	Hierarchical regression step 3 (HC3)
H5a	Trust raises engagement intention	Trust-based commerce; planned behaviour	+	Regression of ENG on TRUST and perceptions (HC3)
H5b	Trust raises purchase intention	Trust-based commerce; planned behaviour	+	Regression of PI on TRUST and perceptions (HC3)
H6	Trust mediates the effects of the perceptions on intentions; the perceptions mediate the effects of exposures on trust	Mediation through trust	indirect	Bootstrapped indirect effects (5,000 resamples, percentile CI)
H7a	AI literacy strengthens the negative effect of AI-generated content exposure on trust	AI literacy; detection and resentment	- (interaction)	Mean-centred interaction in step 4; simple slopes
H7b	Privacy concern strengthens the negative effect of perceived manipulation on trust	Privacy calculus weighting	- (interaction)	Mean-centred interaction in step 4; simple slopes

The seventh group concerns moderation. AI literacy, the competencies that let a person evaluate and critically use AI (Long & Magerko, 2020; Zarouali et al., 2021), is expected to sharpen the negative effect of AI-generated content on trust, because literate consumers are more likely to recognise synthetic content and to resent undisclosed use of it; privacy concern is expected to sharpen the negative effect of perceived manipulation on trust, because concerned consumers weigh surveillance more heavily in their calculus (Dinev & Hart, 2006; Smith et al., 2011; Acquisti et al., 2015). H7a: AI literacy strengthens the negative effect of AI-generated content exposure on trust. H7b: privacy concern strengthens the negative effect of perceived manipulation and surveillance on trust.

Methodology

Research design

Figure 3: Study workflow: theory and hypotheses, instrument design, synthetic data generation, analysis pipeline, Monte Carlo evaluation and reporting, as BPMN-style swimlanes



The study adopts a quantitative, cross-sectional survey design of the kind used in the template study and in most of the consumer literature reviewed above, with one departure that governs everything else: the respondents are synthetic. Figure 3 shows the workflow. The constructs and hypotheses are specified from theory, the instrument is adapted from validated scales, and the analysis plan is written before any data exist. A data-generating model is then specified (Figure 4 and Table 3): latent constructs with the structural paths of the framework, two interaction terms, correlations among the exposure constructs, demographic shifts, ordinal measurement through 48 Likert items with loadings drawn from a stated range, a common-method factor that represents response style, and thresholds that produce the agreement-skewed distributions that such surveys show. Its parameters are chosen to be in the range of published effects and are documented as assumptions. Synthetic respondents are drawn from the model at three sample sizes: a primary sample of 500, a replication-size sample of 130 (the size of the template study) and a large sample of 2,000. The analysis pipeline of the template study (reliability, factor analysis, non-parametric group comparisons, hierarchical regression, mediation and moderation) is applied to each, and a Monte Carlo layer repeats the pipeline 300 times at each of five

sample sizes and 150 times under five alternative assumptions, scoring the estimates against the known parameters (Morris et al., 2019; Paxton et al., 2001; Muthén & Muthén, 2002).

Three consequences follow. The results in Section 5 are not evidence about any population of consumers; they are a demonstration of the pipeline on data with known structure, and the direction and rough size of the effects are inherited from the assumptions in Table 3. What the study can show is methodological: whether the pipeline recovers the structure it is given, how much it attenuates effects, what it misses at the sample sizes that are typical, and what it finds that is not there. The reporting follows the template study's structure so that a reader can see each step of a standard survey analysis performed and then evaluated. All data are synthetic; no student, consumer, platform or brand data were used.

Data-generating model

Figure 4: Data-generating model as a path diagram: latent constructs, structural coefficients, interaction terms, residuals, and the measurement layer with items, loadings, a method factor and ordinal thresholds

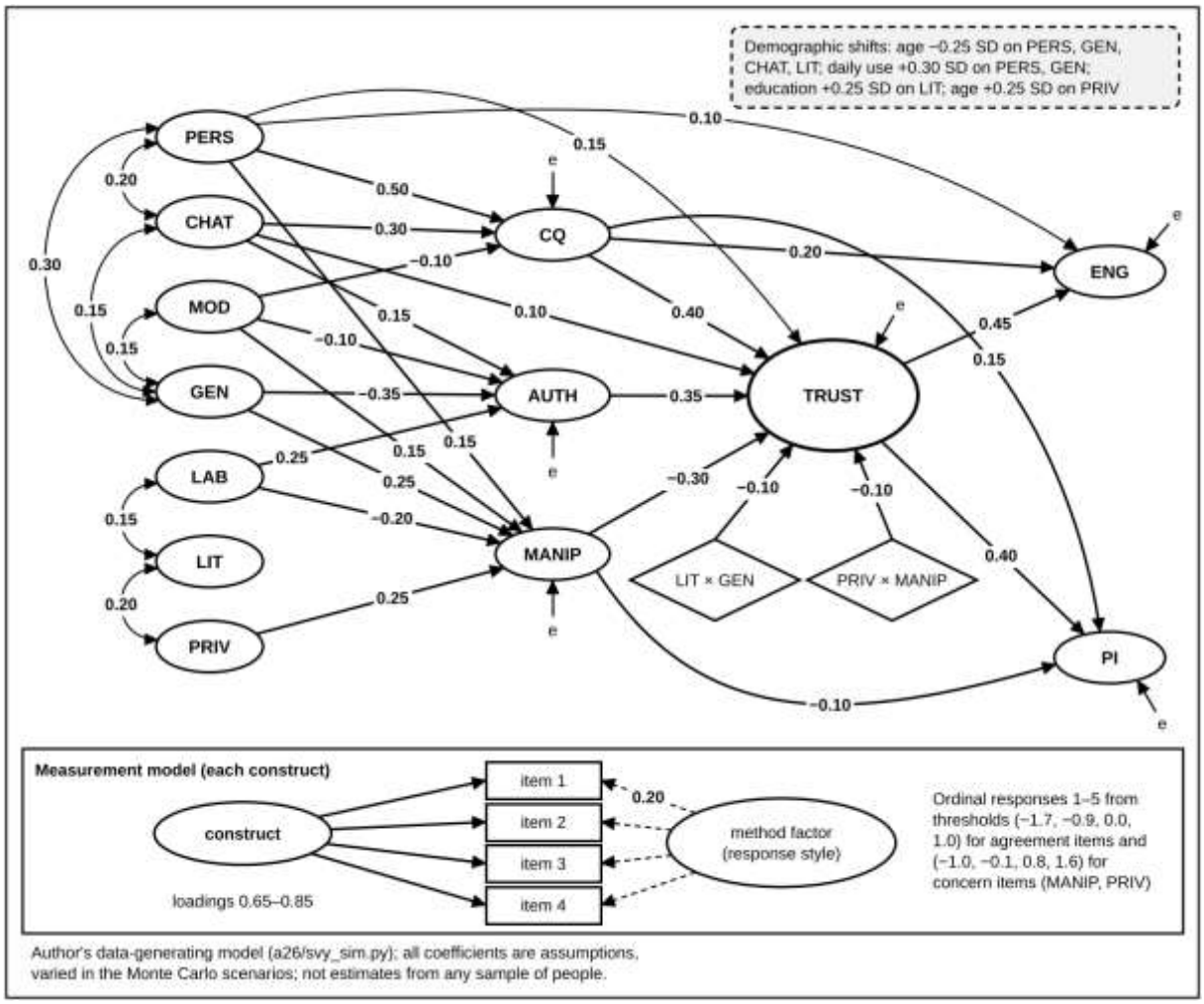


Table 3 gives the model and Figure 4 draws it. The seven exogenous constructs are personalised recommendation exposure (PERS), AI-generated content exposure (GEN), conversational agent interaction quality (CHAT), algorithmic moderation experience (MOD), AI labelling exposure (LAB), AI literacy (LIT) and privacy concern (PRIV), standardised and correlated (PERS with GEN 0.30, PERS with CHAT 0.20, GEN with CHAT 0.15, LIT with PRIV 0.20, LAB with LIT 0.15, MOD with GEN 0.15). Demographics shift them: younger and heavier users see more personalisation and AI-generated content and interact more with agents; literacy falls with age and rises with education; privacy concern rises with age. The structural equations use standardised coefficients: communication quality is 0.50

PERS + 0.30 CHAT − 0.10 MOD; authenticity is −0.35 GEN + 0.25 LAB + 0.15 CHAT − 0.10 MOD; manipulation is 0.15 PERS + 0.25 GEN − 0.20 LAB + 0.25 PRIV + 0.15 MOD; trust is 0.40 CQ + 0.35 AUTH − 0.30 MANIP + 0.15 PERS + 0.10 CHAT − 0.10 (LIT × GEN) − 0.10 (PRIV × MANIP); engagement intention is 0.45 TRUST + 0.20 CQ + 0.10 PERS; purchase intention is 0.40 TRUST + 0.15 CQ − 0.10 MANIP; residuals are scaled so that every construct has unit variance. The sizes are anchored as follows: the trust regression of the template study (standardised coefficients of 0.31 for personalisation, 0.25 for chatbot interaction, −0.23 for privacy concern and −0.18 for perceived monitoring, R-squared 0.46) (Srivastava & Gurme, 2026); the trust-to-intention paths of trust-based commerce models (0.3 to 0.5) (Gefen et al., 2003; Kim et al., 2008); the effects of AI labels and AI-generated content on belief, which are moderate in experiments (Altay & Gilardi, 2024; Longoni et al., 2022); and the small interaction effects that moderation studies in this field typically report (Aguirre et al., 2015; Bleier & Eisenbeiss, 2015). Each construct is measured by four items (three for MOD, LAB, LIT and PRIV) generated as a latent response with a loading drawn uniformly from 0.65 to 0.85, a common-method loading of 0.20 and independent error, cut at thresholds that give a mode of 4 for agreement items and 2 to 3 for the concern items (Podsakoff et al., 2003; Liddell & Kruschke, 2018).

Table 3: Data-generating model: constructs, structural coefficients, correlations, demographic shifts, measurement parameters and their anchors in the literature

Element	Specification in the data-generating model	Anchor or rationale
Exogenous constructs	PERS, GEN, CHAT, MOD, LAB, LIT, PRIV: standard normal, correlated (PERS-GEN 0.30, PERS-CHAT 0.20, GEN-CHAT 0.15, LIT-PRIV 0.20, LAB-LIT 0.15, MOD-GEN 0.15)	Heavier users see more of everything; concern and literacy co-vary (Zarouali et al., 2021; Barth & de Jong, 2017)
Demographic shifts (SD units)	Age: -0.25 on PERS, GEN, LIT, -0.20 on CHAT, +0.25 on PRIV; daily use: +0.30 on PERS and GEN, +0.15 on CHAT; education: +0.25 on LIT, +0.10 on LAB; constructs re-standardised	Age and use gradients in platform surveys (Newman et al., 2024; Tyson & Kikuchi, 2023)
Communication quality (CQ)	0.50 PERS + 0.30 CHAT - 0.10 MOD + residual	Relevance and responsiveness (Aguirre et al., 2015; Adam et al., 2021)
Perceived authenticity (AUTH)	-0.35 GEN + 0.25 LAB + 0.15 CHAT - 0.10 MOD + residual	Replicant effect and labels (Jakesch et al., 2019; Altay & Gilardi, 2024)
Perceived manipulation and surveillance (MANIP)	0.15 PERS + 0.25 GEN - 0.20 LAB + 0.25 PRIV + 0.15 MOD + residual	Personalisation paradox; privacy concern (Aguirre et al., 2015; Malhotra et al., 2004)
Consumer trust (TRUST)	0.40 CQ + 0.35 AUTH - 0.30 MANIP + 0.15 PERS + 0.10 CHAT - 0.10 (LIT x GEN) - 0.10 (PRIV x MANIP) + residual	Trust regression of the template study (0.31, 0.25, -0.23, -0.18; R-squared 0.46) (Srivastava & Gurme, 2026); small interactions
Engagement intention (ENG)	0.45 TRUST + 0.20 CQ + 0.10 PERS + residual	Trust-to-intention paths 0.3 to 0.5 (Gefen et al., 2003; Kim et al., 2008)
Purchase intention (PI)	0.40 TRUST + 0.15 CQ - 0.10 MANIP + residual	Trust-based decision model (Kim et al., 2008; Pavlou, 2003)
Residuals	Scaled so that every endogenous construct has unit variance; latent R-squared for trust 0.51	Standardised model
Measurement	4 items per construct (3 for MOD, LAB, LIT, PRIV): y* = lambda x construct + 0.20 x method factor + error; lambda uniform on 0.65 to 0.85; error variance 1 - lambda-squared - 0.04	Loadings typical of validated scales; method factor as response style (Podsakoff et al., 2003)
Ordinal responses	y* cut at thresholds (-1.7, -0.9, 0.0, 1.0) for agreement items and (-1.0, -0.1, 0.8, 1.6) for MANIP and PRIV; responses 1 to 5	Agreement skew observed in consumer surveys (Liddell & Kruschke, 2018)
Samples and seeds	Primary n = 500, replication n = 130, large n = 2,000 (seed 0); Monte Carlo n = 130, 250, 500, 1,000, 2,000 with 300 replications; five scenarios with 150 replications at n = 130, 500, 2,000	Sample sizes of the literature (Srivastava & Gurme, 2026; Westland, 2010)
Scenarios	No method factor (0); strong method factor (0.35); low reliability (lambda 0.50 to 0.70); no true interactions; stronger skew (thresholds x 0.7)	Common-method bias, reliability, false positives, ceiling effects (Podsakoff et al., 2003; Flake & Fried, 2020)

Sample design

The synthetic population represents digitally active social media users aged 18 and over who have been exposed to AI-driven features, the population the template study targeted purposively. Age groups are 18 to 24 (22%), 25 to 34 (30%), 35 to 44 (22%), 45 to 54 (14%) and 55 and over (12%); gender is 52% female; education is secondary (30%), undergraduate (50%) and postgraduate (20%); daily use is under one hour (25%), one to three hours (45%) and over three hours (30%); and the main platform is Instagram (26%), TikTok (22%), Facebook (20%), YouTube (14%), X (10%) or LinkedIn (8%). These shares are design choices, not estimates of any market. Three samples are drawn with the same seed: 500 (primary), 130 (replication size) and 2,000 (large). For the primary sample the a priori power for a medium standardised regression coefficient (0.25) at alpha 0.05 exceeds 0.99, for a small one (0.10) it is about 0.60, and for an interaction term of 0.10 on a product of two measured variables it is lower, which the Monte Carlo analysis quantifies (Cohen, 1988; Westland, 2010; MacCallum et al., 1999).

Questionnaire design

Table 4: Instrument: constructs, items (five-point Likert scale), sources adapted and the role of each construct in the model

Construct	Items (1 = strongly disagree to 5 = strongly agree)	Adapted from	Role
Personalised recommendation exposure (PERS)	PERS1: The posts and videos I see on this platform are selected for me based on what I have viewed before; PERS2: I often see advertisements that match products I have recently looked at or searched for; PERS3: The platform's feed adapts quickly when my interests change; PERS4: Recommendations on this platform are tailored to me personally	Zarouali et al. (2021); Boerman et al. (2017)	Exposure (exogenous)
AI-generated content exposure (GEN)	GEN1: I regularly see posts, images or videos on this platform that appear to be created by AI; GEN2: Brands I follow use AI-generated text or images in their posts; GEN3: I have interacted with virtual (AI) influencers or AI-generated personas on this platform; GEN4: Many of the replies and comments from brands that I see appear to be written automatically	Hancock et al. (2020); Sands et al. (2022); new items	Exposure (exogenous)
Conversational agent interaction quality (CHAT)	CHAT1: Brand chatbots on this platform understand my questions; CHAT2: Chatbots respond quickly and resolve what I ask; CHAT3: When a chatbot cannot help, it passes me to a person without difficulty; CHAT4: My interactions with brand chatbots are a satisfactory experience	Chung et al. (2020); Sheehan et al. (2020)	Exposure (exogenous)
Algorithmic moderation experience (MOD)	MOD1: Content I posted or shared has been removed, hidden or flagged by the platform's automated systems; MOD2: Some of my posts reach far fewer people than others for reasons I was not told; MOD3: The platform does not explain why content is removed or down-ranked	New items (Gorwa et al., 2020; Eslami et al., 2015)	Exposure (exogenous)
AI labelling exposure (LAB)	LAB1: I see labels on this platform that mark content as AI-generated or AI-assisted; LAB2: When a brand's chatbot talks to me, it is disclosed that I am talking to an AI; LAB3: The platform tells me when a post or advertisement was created with AI tools	New items (Altay & Gilardi, 2024; Coalition for Content Provenance and Authenticity, 2026)	Exposure (exogenous)

AI literacy (LIT)	**LIT1: I understand how AI systems decide what to show me on social media; LIT2: I can usually tell when content has been produced by AI; LIT3: I know what data AI systems use to personalise my experience**	**Long & Magerko (2020)**	**Disposition; moderator (H7a)**
Privacy concern (PRIV)	PRIV1: It bothers me when social media platforms collect information about my behaviour; PRIV2: I am concerned that information I share could be used in ways I did not agree to; PRIV3: I want control over how my personal data are used for AI-driven features	Malhotra et al. (2004)	Disposition; antecedent of MANIP; moderator (H7b)
Communication quality (CQ)	CQ1: The brand communication I receive on this platform is relevant to my needs; CQ2: Brands on this platform respond to me promptly and helpfully; CQ3: The communication feels personal rather than generic; CQ4: The content and replies I receive from brands are rich and informative	Aguirre et al. (2015); Short et al. (1976)	Perception (mediator)
Perceived authenticity (AUTH)	AUTH1: The brand content I see on this platform is genuine; AUTH2: I can tell when a brand's message was written by a person; AUTH3: What brands show me on this platform matches what they really are; AUTH4: The reviews and recommendations I see from other users are real	Lou & Yuan (2019); Jakesch et al. (2019)	Perception (mediator)
Perceived manipulation and surveillance (MANIP)	MANIP1: AI-driven features on this platform make me feel I am being watched; MANIP2: I feel the platform uses what it knows about me to steer my decisions; MANIP3: Content on this platform is designed to pressure me into acting quickly; MANIP4: AI-driven advertising on this platform exploits my personal data	Srivastava & Gurme (2026); Mathur et al. (2019)	Perception (mediator)
Consumer trust (TRUST)	TRUST1: Brands that communicate with me through AI-driven features on this platform are competent; TRUST2: I believe the AI-driven communication I receive from brands is honest; TRUST3: Brands that use AI on this platform have my interests in mind; TRUST4: Overall, I trust the brand communication I receive through this platform	McKnight et al. (2002); Lankton et al. (2015)	Outcome and mediator
Engagement intention (ENG)	ENG1: I intend to keep following and interacting with brands on this platform; ENG2: I am likely to like, comment on or share brand content on this platform; ENG3: I would reply to a brand's message or chatbot on this platform; ENG4: I intend to use this platform to communicate with brands in the next three months	Venkatesh et al. (2012); Hennig-Thurau et al. (2004)	Behavioural intention
Purchase intention (PI)	PI1: I am likely to buy a product I discovered through this platform; PI2: I would purchase from a brand after interacting with it on this platform; PI3: Recommendations on this platform influence what I buy; PI4: I intend to make a purchase through this platform's links or shops in the next three months	Kim et al. (2008); Pavlou (2003)	Behavioural intention

Table 4 gives the instrument: 48 items on five-point Likert scales (1 = strongly disagree, 5 = strongly agree) in thirteen blocks, adapted from validated scales. Exposure to personalised recommendation and to AI-generated content adapts the algorithmic media content awareness scale and behavioural advertising items (Zarouali et al., 2021; Boerman et al., 2017); conversational agent interaction quality adapts chatbot service items (Chung et al., 2020; Sheehan et al., 2020); moderation experience and labelling exposure are new items written for this study; AI literacy adapts the competency framework (Long & Magerko, 2020); privacy concern adapts the IUIPC collection and control dimensions (Malhotra et al., 2004); communication quality adapts relevance and responsiveness items from personalisation and social presence research (Aguirre et al., 2015; Short et al., 1976); authenticity adapts credibility and authenticity items from influencer and AI-mediated communication research (Lou & Yuan, 2019; Jakesch et al., 2019); manipulation and surveillance adapts the perceived monitoring and dark-pattern items (Srivastava & Gurme, 2026; Mathur et al., 2019); trust adapts the competence, integrity and benevolence items of the trust typology (McKnight et al., 2002; Lankton et al., 2015); and engagement and purchase intention adapt behavioural intention items (Venkatesh et al., 2012; Hennig-Thurau et al., 2004). In a field administration the instrument would be pre-tested with expert review and a pilot, as the template study did; here its properties are a consequence of the measurement model,

and the reliability and validity analyses in Section 5 report what a researcher would see.

Analysis plan

Table 5 gives the analysis plan, written to match the template study and the recommendations of the survey-methods literature. Reliability is assessed by Cronbach's alpha (threshold 0.70), composite reliability and average variance extracted (0.50), with discriminant validity by the Fornell-Larcker criterion (Cronbach, 1951; Nunnally & Bernstein, 1994; Fornell & Larcker, 1981; Henseler et al., 2015). Dimensionality is assessed by principal component analysis with varimax rotation, the Kaiser-Meyer-Olkin measure (0.60), Bartlett's test of sphericity, the eigenvalue-greater-than-one rule and loadings of at least 0.50 with cross-loadings below 0.40 (Kaiser, 1974; Bartlett, 1950; Hair et al., 2019). Common-method bias is screened by Harman's single-factor test (Podsakoff et al., 2003; Harman, 1976). Normality is tested by the Shapiro-Wilk test on composite scores (Shapiro & Wilk, 1965), and because Likert composites are ordinal and non-normal, between-group differences are tested by Kruskal-Wallis H with the epsilon-squared effect size (Kruskal & Wallis, 1952; Jamieson, 2004; Norman, 2010). Hypotheses are tested by hierarchical multiple regression on standardised composites in four steps (demographics; exposure constructs; perceptions; interactions) with heteroskedasticity-consistent (HC3) standard errors (White, 1980; Aiken & West, 1991); indirect effects are estimated by bootstrapping with 5,000 resamples and percentile intervals (Preacher & Hayes, 2008; Hayes, 2018; Baron & Kenny, 1986); moderation is tested by mean-centred interaction terms and simple slopes at one standard deviation above and below the moderator (Aiken & West, 1991; Hayes, 2018). Structural equation modelling is not used for the primary analysis for the reasons the template study gave (sample size, non-normality, exploratory purpose) (Kline, 2016; Westland, 2010), and the regression system on composites is instead compared with the same regressions on the latent constructs, which the synthetic design makes possible. The Monte Carlo layer records, for each replication, reliability, the KMO measure, whether the factor structure is recovered, the Harman statistic, the regression coefficients and p-values of every step, the indirect effects and their significance, and the Kruskal-Wallis p-values, and summarises them as power, bias, attenuation, Type S error, factor recovery and false-positive rates (Morris et al., 2019; Gelman & Carlin, 2014; Flake & Fried, 2020).

Verification and reproducibility

Three checks were applied before the hypotheses were tested. On a population of 200,000 synthetic respondents, the regressions on the latent constructs reproduce the structural coefficients of Table 3 to within 0.01 (for example 0.40, 0.35 and −0.30 for the three perceptions on trust, −0.10 for each interaction), which confirms the generator. The composite scores on the same population have Cronbach's alphas of 0.73 to 0.86 and the same regressions attenuate the three perception coefficients to 0.33, 0.30 and −0.22 and the interactions to −0.07, which is the attenuation that measurement error of this reliability implies; the Monte Carlo estimates are scored against both the latent and the composite-level values. On the primary sample, the thirteen-factor structure is recovered exactly by the eigenvalue rule, with every item loading on its intended factor and no cross-loading above 0.40. The generator, the runner, the aggregated results and the figure and table scripts are in the supplementary archive; fixed seeds reproduce every number (Law, 2015; Sargent, 2013).

Table 5: Analysis plan: procedures, statistics, decision rules and the quantity each one is scored on in the Monte Carlo evaluation

Procedure	Statistic and decision rule	Purpose	Monte Carlo score
Reliability	Cronbach's alpha >= 0.70; composite reliability >= 0.70; AVE >= 0.50	Internal consistency of each block	Minimum and mean alpha by sample size
Discriminant validity	Square root of AVE above every inter-construct correlation (Fornell-Larcker)	Constructs distinct	Share of replications passing

Sampling adequacy and sphericity	KMO >= 0.60; Bartlett's test p < 0.05	Factorability of the item correlation matrix	Mean KMO
Exploratory factor analysis	Principal components, varimax rotation; eigenvalue > 1; loading >= 0.50 on the intended factor; cross-loading < 0.40	Dimensionality and item retention	Share of items correctly assigned; share of clean solutions; number of eigenvalues > 1
Common-method bias	Harman's single factor: first unrotated component < 50% of variance	Screen for response-style variance	Harman statistic under method-factor loadings 0, 0.20, 0.35
Normality	Shapiro-Wilk on composites; skewness and kurtosis	Choice of non-parametric tests	Reported
Group comparisons	Kruskal-Wallis H on composites across exposure tertiles, age, platform and use; epsilon-squared	Ordinal between-group differences (as in the template study)	Rejection rate by sample size
Hierarchical regression	Four steps (demographics; exposures; perceptions; interactions) on standardised composites; HC3 standard errors; R-squared change	H1 to H5, H7; incremental explanatory power	Power, mean estimate, attenuation ratio against latent and composite references, Type S error
Mediation	Product of coefficients with 5,000 bootstrap resamples; percentile 95% interval excluding zero	H6	Power to detect each indirect effect; mean estimate against the latent indirect effect
Moderation	Mean-centred interactions in step 4; simple slopes at +/- 1 SD of the moderator	H7a, H7b	Power at each sample size; false-positive rate when the true interaction is zero
Latent comparison	The same regressions on the latent constructs of the generator	Attenuation due to measurement error	Ratio of composite to latent coefficients

Data analysis and results

Sample profile and descriptive statistics

Table 6 and Figure 5 describe the primary synthetic sample of 500 respondents. The age distribution is 19.6% aged 18 to 24, 26.4% aged 25 to 34, 25.0% aged 35 to 44, 15.2% aged 45 to 54 and 13.8% aged 55 and over; 52.0% are female; 31.2% have secondary education, 47.4% an undergraduate degree and 21.4% a postgraduate degree; 29.0% use social media for under an hour a day, 44.4% for one to three hours and 26.6% for more; and the main platform is Instagram or TikTok for 25.6% each, Facebook for 18.4%, YouTube for 14.0%, X for 8.8% and LinkedIn for 7.6%. The replication-size sample of 130 and the large sample of 2,000 are drawn from the same population with the same seed. Item responses show the agreement skew that consumer surveys of this kind report: for the agreement-worded items the modal response is 4, between 12% and 19% of responses are 5, and between 3% and 6% are 1; for the concern items (privacy concern, manipulation and surveillance) the mode is 2 or 3 and the distribution is skewed the other way (Figure 5c). The composite scores (Table 7) have means of 3.37 to 3.44 for the agreement constructs and 2.62 (manipulation and surveillance) and 2.67 (privacy concern) for the concern constructs, standard deviations of 0.83 to 0.92, skewness between −0.38 and 0.27, and Shapiro-Wilk statistics of 0.972 to 0.980 with p-values below 0.001 for every construct, so that none is normally distributed, which motivates the non-parametric group comparisons of Section 5.3 and the robust standard errors of Section 5.4.

Figure 5: Profile of the primary synthetic sample and response distributions of the 48 items

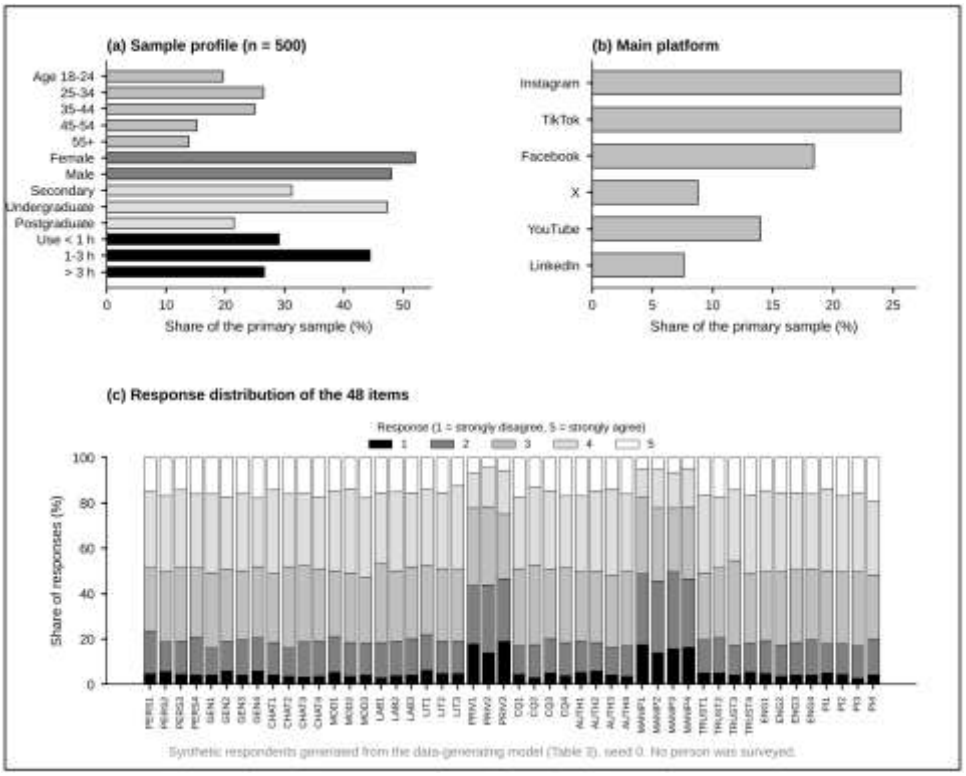


Table 6: Profile of the primary synthetic sample (n = 500): age, gender, education, daily use and main platform

Variable	Category	n	%
Age group	18-24	98	19.6%
	25-34	132	26.4%
	35-44	125	25.0%
	45-54	76	15.2%
	55 and over	69	13.8%
Gender	Female	260	52.0%
	Male	240	48.0%
Education	Secondary	156	31.2%
	Undergraduate	237	47.4%
	Postgraduate	107	21.4%
Daily social media use	Under 1 h/day	145	29.0%
	1-3 h/day	222	44.4%
	Over 3 h/day	133	26.6%
Main platform	Instagram	128	25.6%
	TikTok	128	25.6%
	Facebook	92	18.4%
	X	44	8.8%
	YouTube	70	14.0%
	LinkedIn	38	7.6%
Total		500	100%

Measurement reliability and validity

Table 7 reports reliability and validity and Figure 6 the factor analysis. Cronbach's alpha ranges from 0.74 (privacy concern) and 0.75 (moderation experience), the two three-item blocks with the lowest loadings, to 0.87 (AI-generated content exposure), above the 0.70 threshold for all thirteen constructs;

composite reliability ranges from 0.85 to 0.91 and average variance extracted from 0.63 to 0.74, above 0.50 for all. The Fornell-Larcker criterion holds for every construct: the square root of AVE (0.79 to 0.86) exceeds the largest correlation of the construct with any other (0.47 for trust with engagement intention and for personalisation with communication quality). The 48-item correlation matrix is factorable (KMO = 0.846; Bartlett's chi-squared(1128) = 9,942, $p < 0.001$). Principal component analysis extracts exactly thirteen components with eigenvalues above 1 (7.19 to 1.03; the fourteenth is 0.74), which together explain 67.9% of the variance, and after varimax rotation every item loads on its intended component (loadings 0.70 to 0.86, Table 8) with no cross-loading above 0.40 and the highest cross-loading 0.27. Harman's single-factor test gives 15.0% of variance on the first unrotated component, far below the 50% criterion, even though the data were generated with a method factor loading 0.20 on every item; Section 5.7 shows that the test does not detect a method factor loading 0.35 either, which is the known insensitivity of the test (Podsakoff et al., 2003). On the replication-size sample of 130 the same procedures give alphas of 0.75 to 0.87, KMO = 0.709, thirteen eigenvalues above 1 and every item on its intended component, so that a researcher with 130 respondents would reach the same measurement conclusions as one with 500, which the Monte Carlo analysis shows is true in only 36% of replications at that size.

Figure 6: Measurement model of the primary sample: rotated loadings of the 48 items, scree plot and reliability and convergent validity by construct

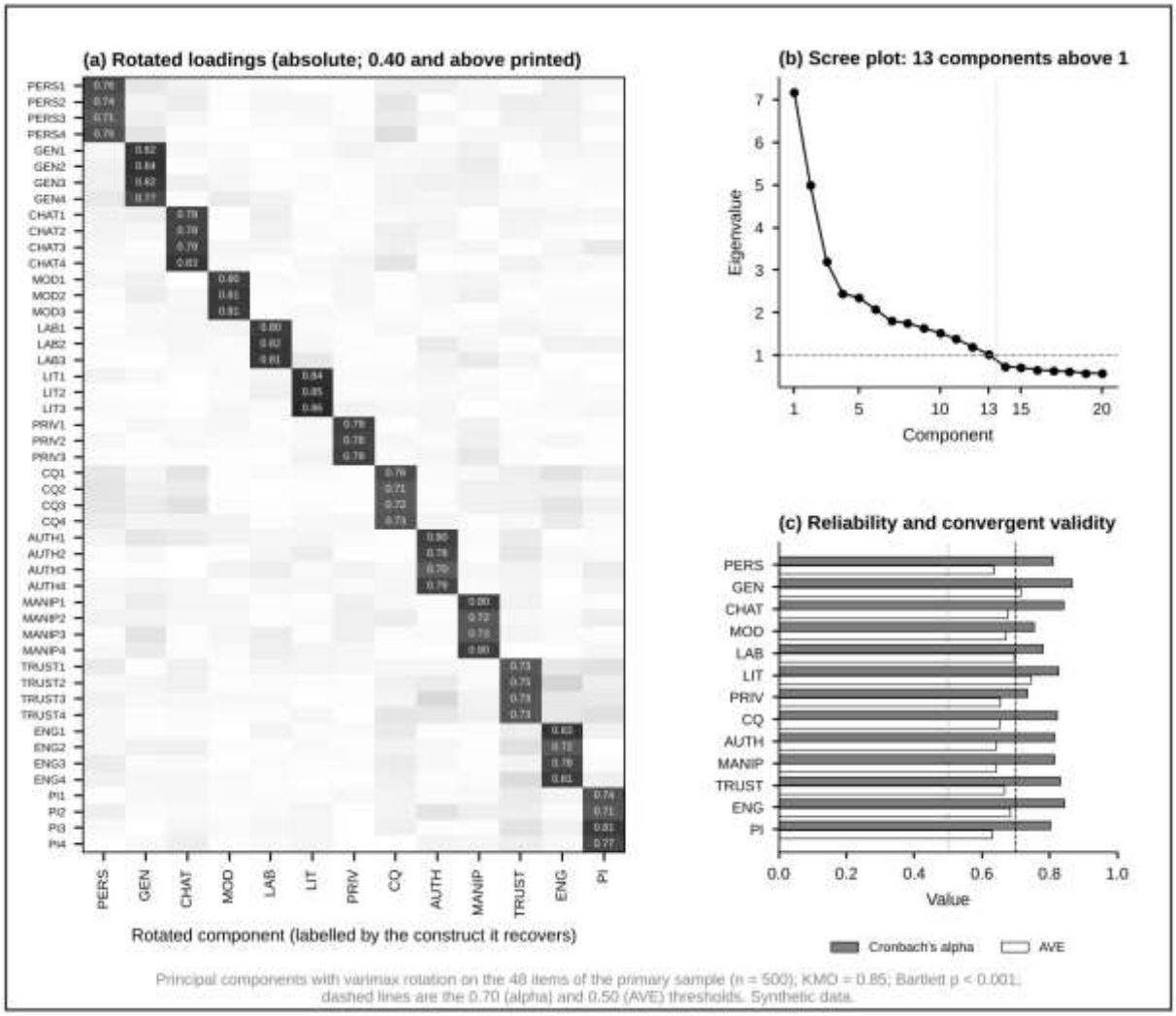


Table 7: Reliability, convergent and discriminant validity and distribution of the thirteen constructs in the primary sample (n = 500)

Construct	Items	Alpha	CR	AVE	Root AVE	Largest correlation with another construct	Mean (SD)	Skew	Shapiro-Wilk W
Personalised recommendation exposure (PERS)	4	0.81	0.87	0.63	0.80	0.47	3.39 (0.86)	-0.23	0.980
AI-generated content exposure (GEN)	4	0.87	0.91	0.71	0.85	0.35	3.42 (0.91)	-0.30	0.976
Conversational agent interaction quality (CHAT)	4	0.84	0.89	0.68	0.82	0.37	3.43 (0.85)	-0.25	0.980
Algorithmic moderation experience (MOD)	3	0.75	0.86	0.67	0.82	0.22	3.42 (0.86)	-0.28	0.973
AI labelling exposure (LAB)	3	0.78	0.87	0.69	0.83	0.23	3.40 (0.87)	-0.26	0.974
AI literacy (LIT)	3	0.83	0.90	0.74	0.86	0.19	3.37 (0.92)	-0.29	0.972
Privacy concern (PRIV)	3	0.74	0.85	0.65	0.81	0.29	2.67 (0.90)	0.15	0.975
Communication quality (CQ)	4	0.82	0.88	0.65	0.81	0.47	3.41 (0.84)	-0.25	0.980
Perceived authenticity (AUTH)	4	0.81	0.88	0.64	0.80	0.34	3.43 (0.84)	-0.38	0.978
Perceived manipulation and surveillance (MANIP)	4	0.81	0.88	0.64	0.80	0.32	2.62 (0.87)	0.27	0.978
Consumer trust (TRUST)	4	0.83	0.89	0.67	0.82	0.47	3.41 (0.87)	-0.22	0.978
Engagement intention (ENG)	4	0.84	0.90	0.68	0.83	0.47	3.42 (0.86)	-0.31	0.977
Purchase intention (PI)	4	0.80	0.87	0.63	0.79	0.44	3.44 (0.83)	-0.31	0.979

Descriptive statistics and correlations

Table 9 gives the correlations among the composites. Trust correlates most with engagement intention (0.47), purchase intention (0.44), perceived authenticity (0.34) and communication quality (0.33), negatively with manipulation and surveillance (−0.15), and weakly with the exposure constructs (0.27 with personalisation, 0.24 with conversational agent quality, 0.15 with labelling, 0.09 with AI-generated content, −0.01 with moderation). The near-zero bivariate correlation between AI-generated content exposure and trust is worth noting now because it recurs below: in the data-generating model the total effect of such exposure on trust is −0.20 at the latent level, but the exposure is correlated with personalisation (0.35) and conversational agent interaction, both of which raise trust, so the raw association is washed out. Among the exposure constructs the correlations are as designed (personalisation with AI-generated content 0.35, with conversational agent quality 0.25); the perceptions correlate with their antecedents (communication quality with personalisation 0.47 and with conversational agent quality 0.37; authenticity with AI-generated content −0.19 and with labelling 0.23; manipulation with AI-generated content 0.32, with privacy concern 0.29 and with labelling −0.20); and the two intentions correlate 0.22 with each other.

Table 8: Rotated component loadings of the 48 items (principal components, varimax) with the highest cross-loading of each item

Item	Loading	Highest cross-loading (factor)	Item	Loading	Highest cross-loading (factor)	Item	Loading	Highest cross-loading (factor)
PERS1	0.76	0.16 (GEN)	LAB2	0.82	0.15 (AUTH)	MANIP1	0.80	0.14 (GEN)
PERS2	0.74	0.18 (CQ)	LAB3	0.81	0.16 (LIT)	MANIP2	0.72	0.14 (CQ)
PERS3	0.71	0.15 (CQ)	LIT1	0.84	0.09 (PERS)	MANIP3	0.73	0.18 (GEN)
PERS4	0.75	0.21 (CQ)	LIT2	0.85	0.11 (LAB)	MANIP4	0.80	0.14 (GEN)
GEN1	0.82	0.12 (MANIP)	LIT3	0.86	0.09 (PRIV)	TRUST1	0.73	0.19 (PI)
GEN2	0.84	0.14 (MANIP)	PRIV1	0.79	0.10 (MANIP)	TRUST2	0.75	0.27 (ENG)
GEN3	0.82	0.12 (CHAT)	PRIV2	0.78	0.12 (MANIP)	TRUST3	0.73	0.24 (AUTH)
GEN4	0.77	0.16 (PERS)	PRIV3	0.79	0.16 (MANIP)	TRUST4	0.73	0.21 (PI)
CHAT1	0.78	0.13 (GEN)	CQ1	0.76	0.19 (ENG)	ENG1	0.83	0.15 (CQ)
CHAT2	0.78	0.12 (CQ)	CQ2	0.71	0.19 (PERS)	ENG2	0.72	0.20 (TRUST)
CHAT3	0.79	0.15 (PI)	CQ3	0.72	0.20 (CHAT)	ENG3	0.78	0.14 (CQ)
CHAT4	0.83	0.18 (CQ)	CQ4	0.73	0.15 (ENG)	ENG4	0.81	0.23 (TRUST)
MOD1	0.80	0.10 (GEN)	AUTH1	0.80	0.17 (GEN)	PI1	0.74	0.15 (TRUST)
MOD2	0.81	0.13 (GEN)	AUTH2	0.78	0.17 (TRUST)	PI2	0.71	0.18 (AUTH)
MOD3	0.81	0.07 (CHAT)	AUTH3	0.70	0.14 (PI)	PI3	0.81	0.19 (TRUST)
LAB1	0.80	0.09 (MOD)	AUTH4	0.79	0.10 (GEN)	PI4	0.77	0.13 (TRUST)

Table 9: Means, standard deviations and correlations of the thirteen composite scores (n = 500)

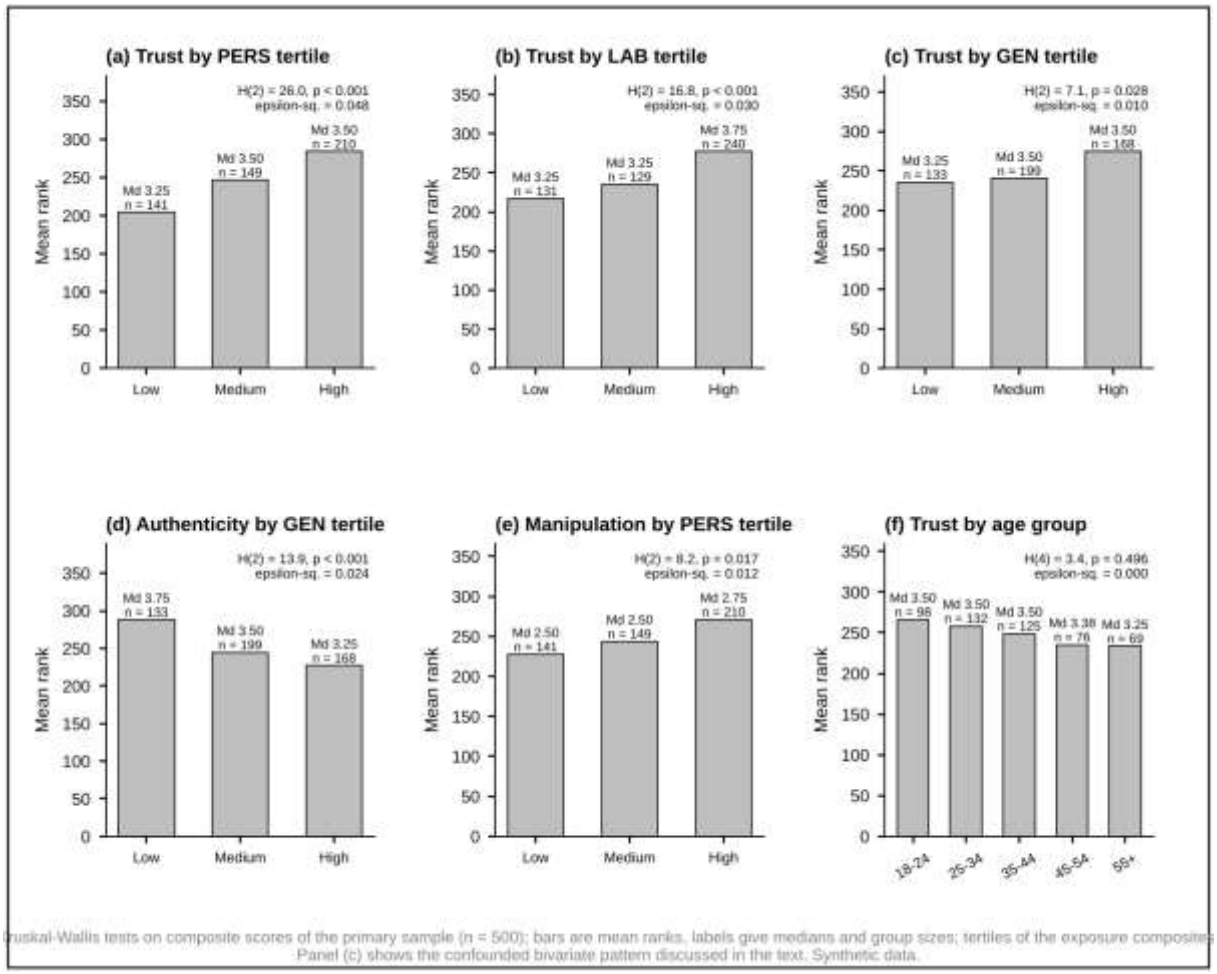
Construct	Mean	SD	1	2	3	4	5	6	7	8	9	10	11	12	13
1 PERS	3.39	0.86	–												
2 GEN	3.42	0.91	0.35	–											
3 CHAT	3.43	0.85	0.25	0.17	–										
4 MOD	3.42	0.86	0.07	0.22	0.10	–									
5 LAB	3.40	0.87	-0.01	0.03	0.04	0.00	–								
6 LIT	3.37	0.92	-0.03	0.05	0.01	0.05	0.19	–							
7 PRIV	2.67	0.90	-0.01	0.06	-0.04	0.11	0.01	0.14	–						
8 CQ	3.41	0.84	0.47	0.27	0.37	0.06	-0.03	-0.04	0.05	–					
9 AUTH	3.43	0.84	-0.11	-0.19	0.12	-0.06	0.23	0.15	0.11	-0.07	–				
10 MANIP	2.62	0.87	0.17	0.32	0.06	0.17	-0.20	0.02	0.29	0.16	-0.10	–			
11 TRUST	3.41	0.87	0.27	0.09	0.24	-0.01	0.15	0.01	0.02	0.33	0.34	-0.15	–		
12 ENG	3.42	0.86	0.31	0.15	0.20	0.00	0.12	0.02	-0.01	0.40	0.09	0.00	0.47	–	
13 PI	3.44	0.83	0.15	0.00	0.21	0.01	0.14	-0.01	-0.03	0.23	0.22	-0.15	0.44	0.22	–

Group comparisons: Kruskal-Wallis tests

Table 10 and Figure 7 report the Kruskal-Wallis tests. Trust differs across tertiles of personalisation exposure ($H(2) = 25.95$, $p < 0.001$, epsilon-squared = 0.048; mean ranks 204, 247, 284 from low to high), across tertiles of labelling exposure ($H(2) = 16.84$, $p < 0.001$, epsilon-squared = 0.030; mean ranks 217, 235, 277) and, more weakly, across tertiles of AI-generated content exposure ($H(2) = 7.13$, $p = 0.028$, epsilon-squared = 0.010), where the highest-exposure tertile has the highest trust (mean ranks 236, 240, 275). Perceived authenticity falls across tertiles of AI-generated content exposure ($H(2) = 13.87$, $p = 0.001$; mean ranks 288, 245, 227), and perceived manipulation rises across tertiles of personalisation

exposure (H(2) = 8.18, p = 0.017; mean ranks 228, 244, 271). Trust does not differ by age group (H(4) = 3.38, p = 0.496), by main platform (H(5) = 3.34, p = 0.648) or by daily use (H(2) = 0.39, p = 0.825), which reflects the design, in which demographics act on trust only through exposure. Purchase intention differs sharply by trust tertile (H(2) = 79.06, p < 0.001, epsilon-squared = 0.155; mean ranks 170, 236, 310). The direction of the AI-generated content comparison is the point to carry forward: a researcher who used this bivariate test as the test of H2a or of the total effect of synthetic content on trust would conclude that exposure to AI-generated content raises trust, the opposite of the data-generating model, because the heavy users who see the most synthetic content also see the most personalised, responsive communication. The same reversal appears in the sample of 130 (H(2) = 7.39, p = 0.025) and in the sample of 2,000 (H(2) = 7.56, p = 0.023). The multivariate tests of the next section, which hold the other exposures constant, recover the designed sign.

Figure 7: Kruskal-Wallis comparisons of trust, authenticity and manipulation across exposure tertiles and age groups in the primary sample



Kruskal-Wallis tests on composite scores of the primary sample (n = 500); bars are mean ranks, labels give medians and group sizes; tertiles of the exposure composites. Panel (c) shows the confounded bivariate pattern discussed in the text. Synthetic data.

Table 10: Kruskal-Wallis tests on the composite scores of the primary sample: groups, medians, mean ranks, H, p and epsilon-squared

Comparison (n = 500)	Groups	Medians	Mean ranks	H	df	p	Epsilon-squared
Trust by AI-generated content exposure (tertiles)	Low (n = 133); Medium (n = 199); High (n = 168)	3.25 / 3.50 / 3.50	235 / 240 / 275	7.13	2	0.028	0.010
Trust by exposure to AI labels (tertiles)	Low (n = 131); Medium (n = 129); High (n = 240)	3.25 / 3.25 / 3.75	217 / 235 / 277	16.84	2	< 0.001	0.030
Trust by personalisation exposure (tertiles)	Low (n = 141); Medium (n = 149); High (n = 210)	3.25 / 3.50 / 3.50	204 / 247 / 284	25.95	2	< 0.001	0.048
Authenticity by AI-generated content exposure (tertiles)	Low (n = 133); Medium (n = 199); High (n = 168)	3.75 / 3.50 / 3.25	288 / 245 / 227	13.87	2	< 0.001	0.024
Perceived manipulation by personalisation exposure (tertiles)	Low (n = 141); Medium (n = 149); High (n = 210)	2.50 / 2.50 / 2.75	227 / 244 / 271	8.18	2	0.017	0.012
Trust by age group	18-24 (n = 98); 25-34 (n = 132); 35-44 (n = 125); 45-54 (n = 76); 55+ (n = 69)	3.50 / 3.50 / 3.50 / 3.38 / 3.25	266 / 258 / 249 / 235 / 233	3.38	4	0.496	0.000
Trust by main platform	Instagram (n = 128); TikTok (n = 128); Facebook (n = 92); X (n = 44); YouTube (n = 70); LinkedIn (n = 38)	3.50 / 3.50 / 3.25 / 3.38 / 3.75 / 3.50	250 / 249 / 243 / 241 / 278 / 237	3.34	5	0.648	0.000
Trust by daily use	Under 1 h/day (n = 145); 1-3 h/day (n = 222); Over 3 h/day (n = 133)	3.50 / 3.50 / 3.50	244 / 252 / 254	0.39	2	0.825	0.000
Purchase intention by trust (tertiles)	Low (n = 132); Medium (n = 151); High (n = 217)	3.00 / 3.25 / 4.00	170 / 236 / 309	79.06	2	< 0.001	0.155

Hierarchical regression

Table 11 and Figure 8 give the hierarchical regression of standardised trust with HC3 standard errors. Demographics alone (step 1) explain 0.9% of the variance ($F = 1.20$, $p = 0.309$) and no demographic coefficient is significant at any step. Adding the five exposure constructs and the two dispositions (step 2) raises R-squared to 0.134 (change 0.125, $F = 6.75$, $p < 0.001$): personalisation ($\beta = 0.235$, $p < 0.001$), conversational agent quality ($\beta = 0.184$, $p < 0.001$) and labelling exposure ($\beta = 0.149$, $p < 0.001$) raise trust, while AI-generated content exposure ($\beta = -0.019$, $p = 0.676$), moderation experience ($\beta = -0.042$, $p = 0.345$), AI literacy ($\beta = -0.031$) and privacy concern ($\beta = 0.041$) have no significant total effect at this sample size. The AI-generated content coefficient deserves comment: the model's latent total effect is −0.198 and the expected composite-level coefficient is −0.126, so the estimate of −0.019 is within two standard errors of the expectation (0.045) but reads as a null; the Monte Carlo analysis shows that this test has 77% power at n = 500 and 19% at n = 130. Adding the three perceptions (step 3) raises R-squared to 0.322 (change 0.188, $F = 18.48$, $p < 0.001$): communication quality ($\beta = 0.272$, $p < 0.001$), authenticity ($\beta = 0.357$, $p < 0.001$) and manipulation and surveillance ($\beta = -0.219$, $p < 0.001$) are the three strongest predictors, personalisation retains a direct effect ($\beta = 0.174$, $p < 0.001$), and the exposure coefficients of conversational agent quality and labelling fall to non-significance, consistent with full mediation through the perceptions. Adding the two interactions (step 4) changes R-squared by 0.004 (F for the step not significant): the coefficients of AI literacy by AI-generated content ($\beta = -0.026$, $p = 0.483$) and privacy concern by manipulation ($\beta = -0.055$, $p = 0.144$) have the designed sign but neither is significant, although both are at n = 2,000 ($\beta = -0.080$, $p < 0.001$ and $\beta = -0.051$, $p = 0.003$). The same regression on the latent constructs of the generator gives coefficients of 0.347 for communication quality, 0.408 for authenticity, −0.315 for manipulation and −0.081 and −0.087 for the interactions, with R-squared 0.478, against 0.326 for the composites: the composite regression attenuates the perception

coefficients by 13 to 30% and the interaction coefficients by 35 to 68%, and understates the explained variance by a third, because the composites carry measurement error and the products of two error-laden composites carry more. The regressions of the two intentions give trust coefficients of 0.363 on engagement intention (R-squared 0.295; communication quality 0.229, personalisation 0.108) and 0.333 on purchase intention (R-squared 0.226; communication quality 0.135, authenticity 0.111, manipulation −0.119), all with $p < 0.02$.

Figure 8: Hierarchical regression of trust: final-step coefficients with 95% intervals against the latent and composite-level reference values, and explained variance by step and sample size

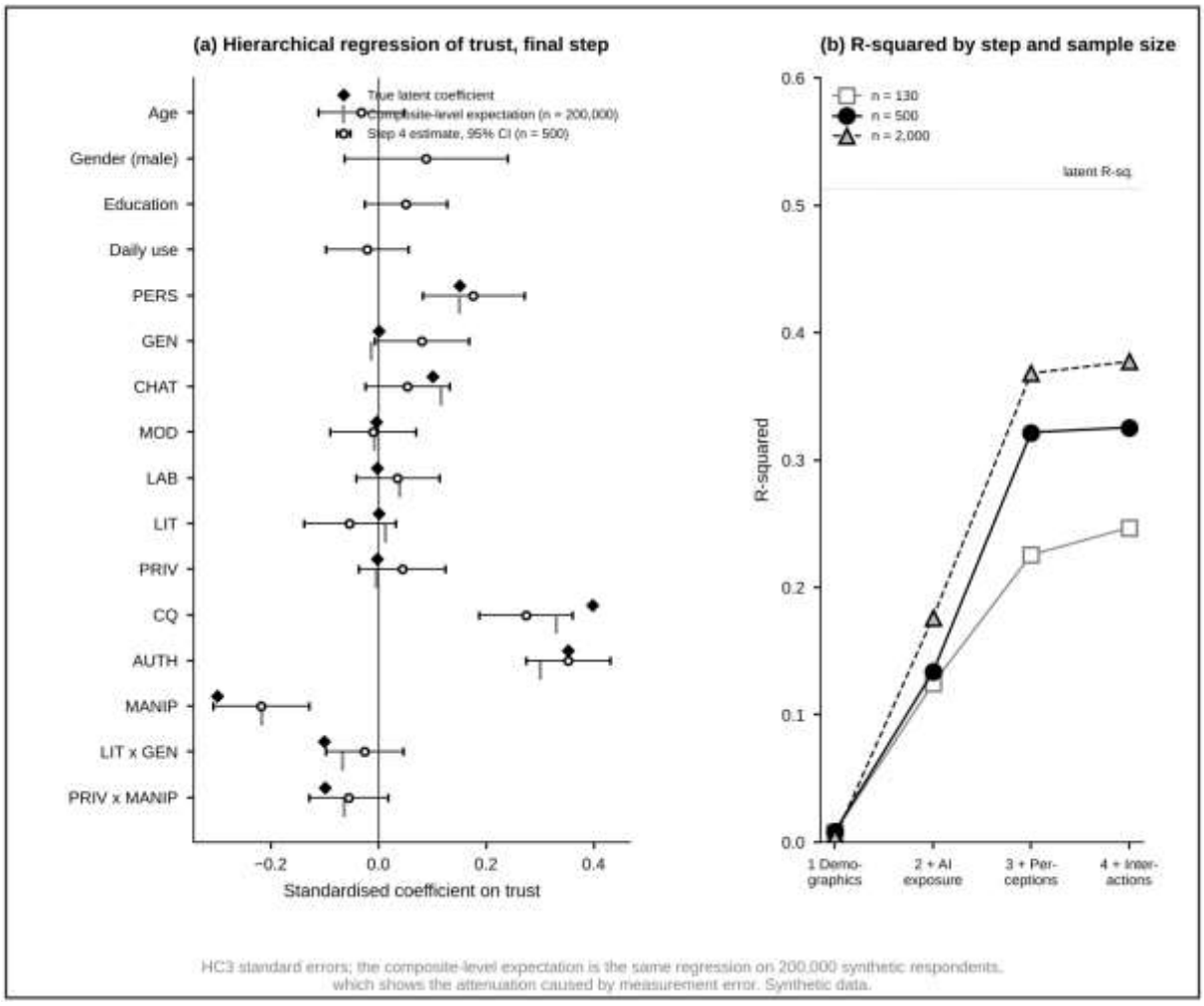


Table 11: Hierarchical regression of consumer trust on standardised composites in four steps (n = 500): standardised coefficients with HC3 standard errors, R-squared and change in R-squared

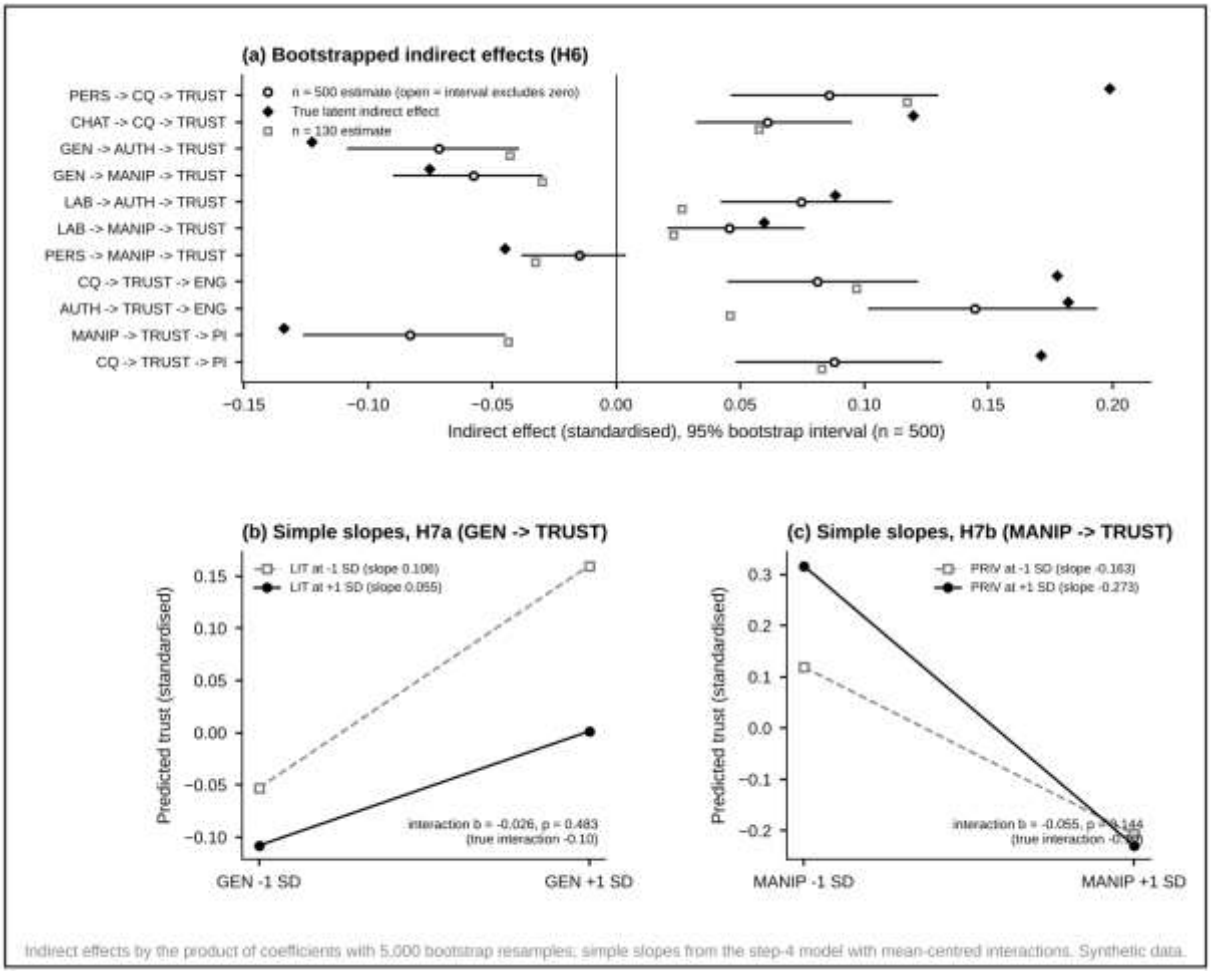
Predictor	Step 1: demographics	Step 2: AI exposure and dispositions	Step 3: perceptions	Step 4: interactions
Age (standardised)	-0.080 (0.042)	-0.044 (0.044)	-0.035 (0.041)	-0.032 (0.041)
Gender (male = 1)	0.038 (0.090)	0.076 (0.086)	0.088 (0.077)	0.088 (0.077)
Education	0.032 (0.047)	0.033 (0.045)	0.046 (0.039)	0.051 (0.039)
Daily use	0.021 (0.045)	-0.030 (0.044)	-0.022 (0.039)	-0.022 (0.039)
PERS		0.235 (0.049)***	0.174 (0.048)***	0.176 (0.048)***
GEN		-0.019 (0.045)	0.085 (0.044)	0.081 (0.045)
CHAT		0.184 (0.045)***	0.054 (0.040)	0.054 (0.040)
MOD		-0.042 (0.044)	-0.004 (0.041)	-0.011 (0.041)
LAB		0.149 (0.042)***	0.036 (0.040)	0.036 (0.040)
LIT		-0.031 (0.046)	-0.055 (0.043)	-0.053 (0.043)
PRIV		0.041 (0.044)	0.038 (0.041)	0.044 (0.041)
CQ			0.272 (0.044)***	0.274 (0.044)***
AUTH			0.357 (0.040)***	0.353 (0.040)***
MANIP			-0.219 (0.045)***	-0.218 (0.045)***
LIT x GEN (H7a)				-0.026 (0.037)
PRIV x MANIP (H7b)				-0.055 (0.038)
R-squared (adjusted)	0.009 (0.001)	0.134 (0.114)	0.322 (0.302)	0.326 (0.303)
Change in R-squared	0.009	0.125	0.188	0.004
F (p)	1.20 (0.309)	6.75 (< 0.001)	18.48 (< 0.001)	16.58 (< 0.001)
HC3 SE in parentheses	*** p < 0.001	** p < 0.01	* p < 0.05	predictors standardised (gender 0/1)

Mediation and moderation

Table 12 and Figure 9 report the bootstrapped indirect effects and the moderation analysis. Ten of the eleven hypothesised indirect effects have 95% percentile intervals that exclude zero in the primary sample. Personalisation raises trust through communication quality (a = 0.372, b = 0.231, indirect = 0.086, interval 0.047 to 0.130) and conversational agent quality does the same (indirect = 0.061, interval 0.032 to 0.095); AI-generated content lowers trust through authenticity (a = −0.203, b = 0.353, indirect = −0.072, interval −0.108 to −0.039) and through manipulation and surveillance (indirect = −0.058, interval −0.090 to −0.030); labelling raises trust through authenticity (indirect = 0.074) and through lower manipulation (indirect = 0.045); and trust carries the effects of communication quality and authenticity to engagement intention (indirect effects 0.081 and 0.145) and of communication quality and manipulation to purchase intention (0.088 and −0.083). The one indirect effect that is not supported, personalisation through manipulation (indirect = −0.015, interval −0.038 to 0.004), has a true latent value of −0.045 and is the smallest in the model. Every estimated indirect effect is smaller in absolute value than its latent counterpart (for example 0.086 against 0.199 for personalisation through communication quality, and −0.072 against −0.123 for AI-generated content through authenticity), because the product of two attenuated coefficients is attenuated twice. In the replication-size sample only four of the eleven indirect effects are significant, and in the large sample all eleven are. For moderation, the simple slopes have the hypothesised pattern: the slope of AI-generated content on trust is 0.106 at low AI literacy and 0.055 at high literacy, and the slope of manipulation on trust is −0.163 at low privacy concern and −0.273 at high concern, but neither interaction is significant at n = 500. The latent-level interactions are −0.101 and −0.099, and their composite-level expectations are −0.067 and

−0.064; the Monte Carlo analysis puts the power to detect them at 36% and 33% for a sample of this size.

Figure 9: Bootstrapped indirect effects against their latent values and the n = 130 estimates, and simple slopes for the two moderation hypotheses



Validated model and hypothesis summary

Figure 10 draws the validated model with the primary-sample estimates, and Table 13 summarises the hypothesis tests at the three sample sizes. In the primary sample, thirteen of the eighteen directional hypotheses are supported with the designed sign: H1a and H1b (personalisation and conversational agent quality raise communication quality: β = 0.372 and 0.265), H2a, H2b and H2c (AI-generated content lowers authenticity, β = −0.203; labelling and conversational agent quality raise it, β = 0.211 and 0.173), H3b and H3c (AI-generated content raises manipulation, β = 0.265; labelling lowers it, β = −0.209), H4a to H4d (the three perceptions and the direct personalisation effect on trust) and H5a and H5b (trust on the two intentions). Five are not supported at n = 500: H1c and H3d, the two moderation-experience effects, whose latent coefficients are −0.10 and 0.15 (estimates −0.017, p = 0.673, and 0.074, p = 0.075); H3a, personalisation on manipulation, whose latent coefficient is 0.15 (estimate 0.069, p = 0.122); and the two interactions H7a and H7b. H6 is largely supported (ten of eleven indirect effects). The pattern across sample sizes is the point of the exercise. At n = 130, ten hypotheses are supported: four of the thirteen supported at n = 500 are lost (authenticity on trust, β = 0.108, p = 0.227, against a latent value of 0.35; manipulation on trust, p = 0.068; the direct personalisation effect; and labelling on manipulation), while personalisation on manipulation (H3a) happens to be found; the mediation analysis finds four of eleven indirect effects. At n = 2,000, every hypothesis including the two interactions and H1c, H3a and H3d is supported, with coefficients within 0.05 of their composite-level expectations. The hypotheses supported at every sample size are those whose latent coefficients are

0.25 or larger; those lost at 130 have latent coefficients of 0.15 to 0.35 on a noisy outcome; those found only at 2,000 have latent coefficients of 0.10 to 0.15 or are interactions.

Table 12: Mediation (bootstrapped indirect effects, 5,000 resamples) and moderation (interaction coefficients and simple slopes) in the primary sample, with the true latent values

Path or interaction	Hypotheses	a (X to M)	b (M to Y)	Direct c'	Indirect a x b	95% bootstrap CI	Latent (true) value	Supported (n = 500)
PERS -> CQ -> TRUST	H1a, H4a, H6	0.372	0.231	0.151	0.086	[0.047, 0.130]	0.199	Yes
CHAT -> CQ -> TRUST	H1b, H4a, H6	0.265	0.231	0.122	0.061	[0.032, 0.095]	0.120	Yes
GEN -> AUTH -> TRUST	H2a, H4b, H6	-0.203	0.353	0.051	-0.072	[-0.108, -0.039]	-0.123	Yes
GEN -> MANIP -> TRUST	H3b, H4c, H6	0.265	-0.217	0.037	-0.058	[-0.090, -0.030]	-0.075	Yes
LAB -> AUTH -> TRUST	H2b, H4b, H6	0.211	0.353	0.073	0.074	[0.042, 0.111]	0.088	Yes
LAB -> MANIP -> TRUST	H3c, H4c, H6	-0.209	-0.217	0.102	0.045	[0.021, 0.076]	0.059	Yes
PERS -> MANIP -> TRUST	H3a, H4c, H6	0.069	-0.217	0.251	-0.015	[-0.038, 0.004]	-0.045	No
CQ -> TRUST -> ENG	H5a, H6	0.231	0.351	0.237	0.081	[0.045, 0.121]	0.178	Yes
AUTH -> TRUST -> ENG	H5a, H6	0.353	0.410	-0.043	0.145	[0.102, 0.194]	0.182	Yes
MANIP -> TRUST -> PI	H5b, H6	-0.217	0.382	-0.085	-0.083	[-0.126, -0.045]	-0.134	Yes
CQ -> TRUST -> PI	H5b, H6	0.231	0.381	0.091	0.088	[0.049, 0.131]	0.171	Yes
Moderation H7a: LIT x GEN	H7a	-0.026	p = 0.483	GEN slope at LIT -1 SD 0.106	at +1 SD 0.055		-0.101	No
Moderation H7b: PRIV x MANIP	H7b	-0.055	p = 0.144	MANIP slope at PRIV -1 SD -0.163	at +1 SD -0.273		-0.099	No

Figure 10: Validated model: standardised coefficients from the primary sample on the paths of the framework, with significance levels

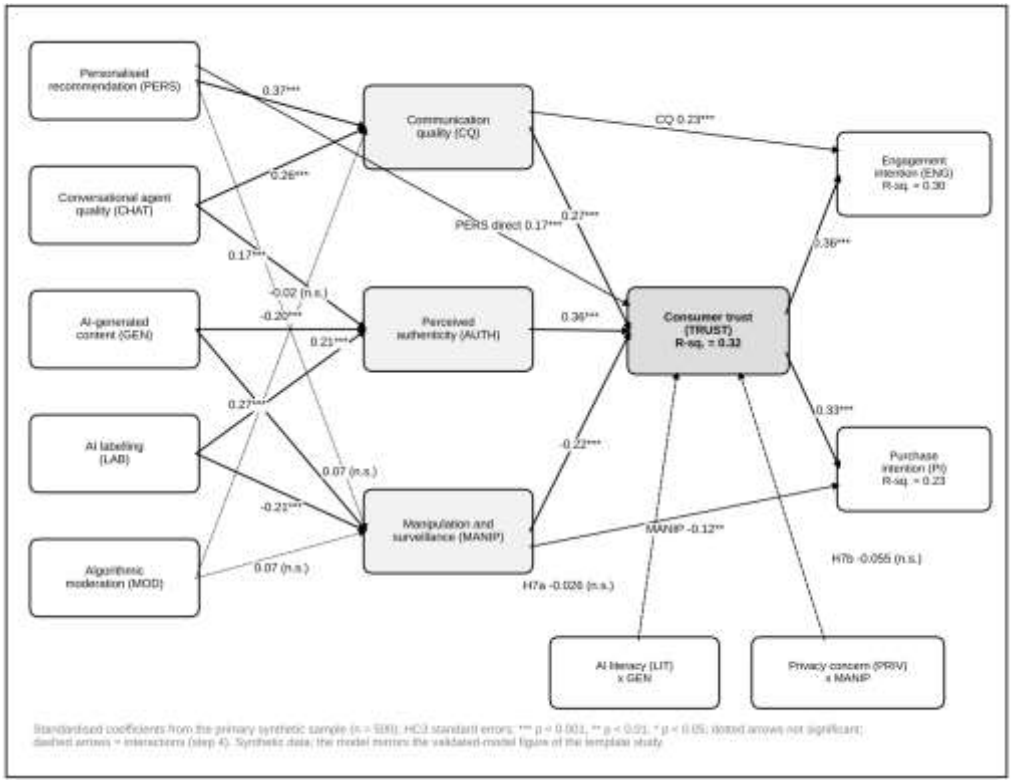


Table 13: Summary of hypothesis tests: estimates, p-values and outcomes in the primary sample, the replication-size sample and the large sample, against the true latent coefficients

Hypothesis	Path (expected sign)	Standardised estimate (n = 500)	p (HC3)	True latent coefficient	Outcome (n = 500)	Outcome at n = 130 (estimate)	Outcome at n = 2,000 (estimate)
H1a	PERS -> CQ (+)	0.372	< 0.001	+0.50	Supported	Supported (0.41)	Supported (0.41)
H1b	CHAT -> CQ (+)	0.265	< 0.001	+0.30	Supported	Supported (0.20)	Supported (0.28)
H1c	MOD -> CQ (-)	-0.017	0.673	-0.10	Not supported	Not supported (0.03)	Supported (-0.04)
H2a	GEN -> AUTH (-)	-0.203	< 0.001	-0.35	Supported	Supported (-0.36)	Supported (-0.24)
H2b	LAB -> AUTH (+)	0.211	< 0.001	+0.25	Supported	Supported (0.22)	Supported (0.24)
H2c	CHAT -> AUTH (+)	0.173	< 0.001	+0.15	Supported	Supported (0.21)	Supported (0.17)
H3a	PERS -> MANIP (+)	0.069	0.122	+0.15	Not supported	Supported (0.22)	Supported (0.12)
H3b	GEN -> MANIP (+)	0.265	< 0.001	+0.25	Supported	Supported (0.20)	Supported (0.22)
H3c	LAB -> MANIP (-)	-0.209	< 0.001	-0.20	Supported	Not supported (-0.16)	Supported (-0.12)

H3d	**MOD -> MANIP (+)**	**0.074**	**0.075**	**+0.15**	**Not supported**	**Not supported (-0.00)**	**Supported (0.12)**
H4a	CQ -> TRUST (+)	0.272	< 0.001	+0.40	Supported	Supported (0.33)	Supported (0.33)
H4b	AUTH -> TRUST (+)	0.357	< 0.001	+0.35	Supported	Not supported (0.11)	Supported (0.28)
H4c	MANIP -> TRUST (-)	-0.219	< 0.001	-0.30	Supported	Not supported (-0.21)	Supported (-0.23)
H4d	PERS -> TRUST direct (+)	0.174	< 0.001	+0.15	Supported	Not supported (0.11)	Supported (0.14)
H5a	TRUST -> ENG (+)	0.363	< 0.001	+0.45	Supported	Supported (0.32)	Supported (0.40)
H5b	TRUST -> PI (+)	0.333	< 0.001	+0.40	Supported	Supported (0.28)	Supported (0.34)
H7a	LIT x GEN -> TRUST (-)	-0.026	0.483	-0.10	Not supported	Not supported (-0.16)	Supported (-0.08)
H7b	PRIV x MANIP -> TRUST (-)	-0.055	0.144	-0.10	Not supported	Not supported (-0.01)	Supported (-0.05)
H6	Indirect effects through CQ, AUTH, MANIP and TRUST (11 paths)	10 of 11 CIs exclude zero	bootstrap	10 of 11 non-zero	Largely supported	4 of 11 significant	11 of 11 significant

Monte Carlo evaluation of the pipeline

Table 14 and Figure 11 report the Monte Carlo evaluation over 300 replications at each of five sample sizes and 150 replications in each of five scenarios. In Table 14, "Alpha / KMO" is the minimum Cronbach's alpha and the mean KMO; "Clean EFA / k" the share of replications in which every item loads on its intended factor with no cross-loading above 0.40 and the mean number of eigenvalues above 1; "Harman" the share of replications in which the first unrotated component exceeds 50% of variance, with its mean share in parentheses; the power columns give the share of replications in which the named coefficient is significant at 0.05 (for the no-interaction scenario, the false-positive rate); "Indirect" the power of the indirect effects of AI-generated content through authenticity and of personalisation through communication quality; "Attenuation" the mean estimate over the latent coefficient for communication quality and manipulation; and "Exaggeration" the mean absolute significant estimate over its composite-level expectation for manipulation and for the literacy interaction. Four results stand out. First, measurement looks adequate at every size: the minimum alpha is 0.74 to 0.75, the mean alpha 0.82, and KMO rises from 0.73 at n = 130 to 0.89 at n = 2,000; but the thirteen-factor structure is recovered cleanly (every item on its factor, no cross-loading above 0.40) in only 36% of replications at n = 130 and 91% at n = 250, and the eigenvalue-greater-than-one rule returns exactly thirteen components in 60% of replications at n = 130 and 76% at n = 2,000, returning twelve or fourteen otherwise. A study of 130 respondents that reports a clean factor solution has, in this testbed, more often than not reported a solution that would not replicate. Second, power is adequate for the large perception effects and poor for everything else. The three perception coefficients on trust are detected in 91%, 94% and 72% of replications at n = 130 and in 95% or more from n = 250; the direct personalisation effect in 34% at 130 and 94% at 500; the total effects of AI-generated content, labelling and moderation in 19%, 36% and 13% at 130, 77%, 92% and 41% at 500, and 100%, 100% and 93% at 2,000; and the two interactions in 10% and 11% at 130, 36% and 33% at 500, 70% at 1,000 and 90 to 95% at 2,000. The indirect effects are detected more often than the total effects they compose (75% for AI-

generated content through authenticity at n = 130, 100% from 500) because the product test uses the stronger a and b paths. Third, estimates are attenuated but unbiased relative to the composite-level expectation: the mean coefficient of communication quality is 0.32 to 0.33 at every size against a latent 0.40 (attenuation 0.82), of manipulation −0.22 against −0.30 (0.72 to 0.75), of the interactions −0.06 against −0.10 (0.63); no replication at any size gives a significant coefficient of the wrong sign for the perception effects (Type S error 0%). Significant estimates are exaggerated when power is low: at n = 130 the mean significant interaction estimate is 2.7 times its expected value and the mean significant manipulation estimate 1.2 times; at n = 500 the figures are 1.5 and 1.0. Fourth, the scenarios show what the standard diagnostics do and do not catch. Removing the method factor lowers alpha slightly (0.72 minimum) and R-squared (0.35) and raises the power of the total-effect tests; a strong method factor (0.35) raises alpha (0.80), KMO (0.90) and R-squared (0.46), so that the data look better while the exposure tests lose power (62% for AI-generated content), and Harman's first component rises only to 22%, never near the 50% criterion. Low reliability (loadings 0.50 to 0.70) lowers the minimum alpha to 0.58, halves the share of clean factor solutions and reduces the attenuation ratio of the manipulation coefficient to 0.55. When the true interactions are zero, the interaction tests reject in 4% and 7% of replications, which is the nominal rate, so the power figures above are not inflated by false positives. Stronger agreement skew changes little. The Kruskal-Wallis test of trust across AI-generated content tertiles rejects in only 3% to 8% of replications at every size, which quantifies the confounding seen in Section 5.3: the bivariate test has no power against a total effect of −0.20 because the exposure is confounded with the exposures that raise trust.

Figure 11: Monte Carlo evaluation: power of the regression tests by sample size, mean estimates against reference values, exaggeration of significant estimates and scenario comparisons at n = 500

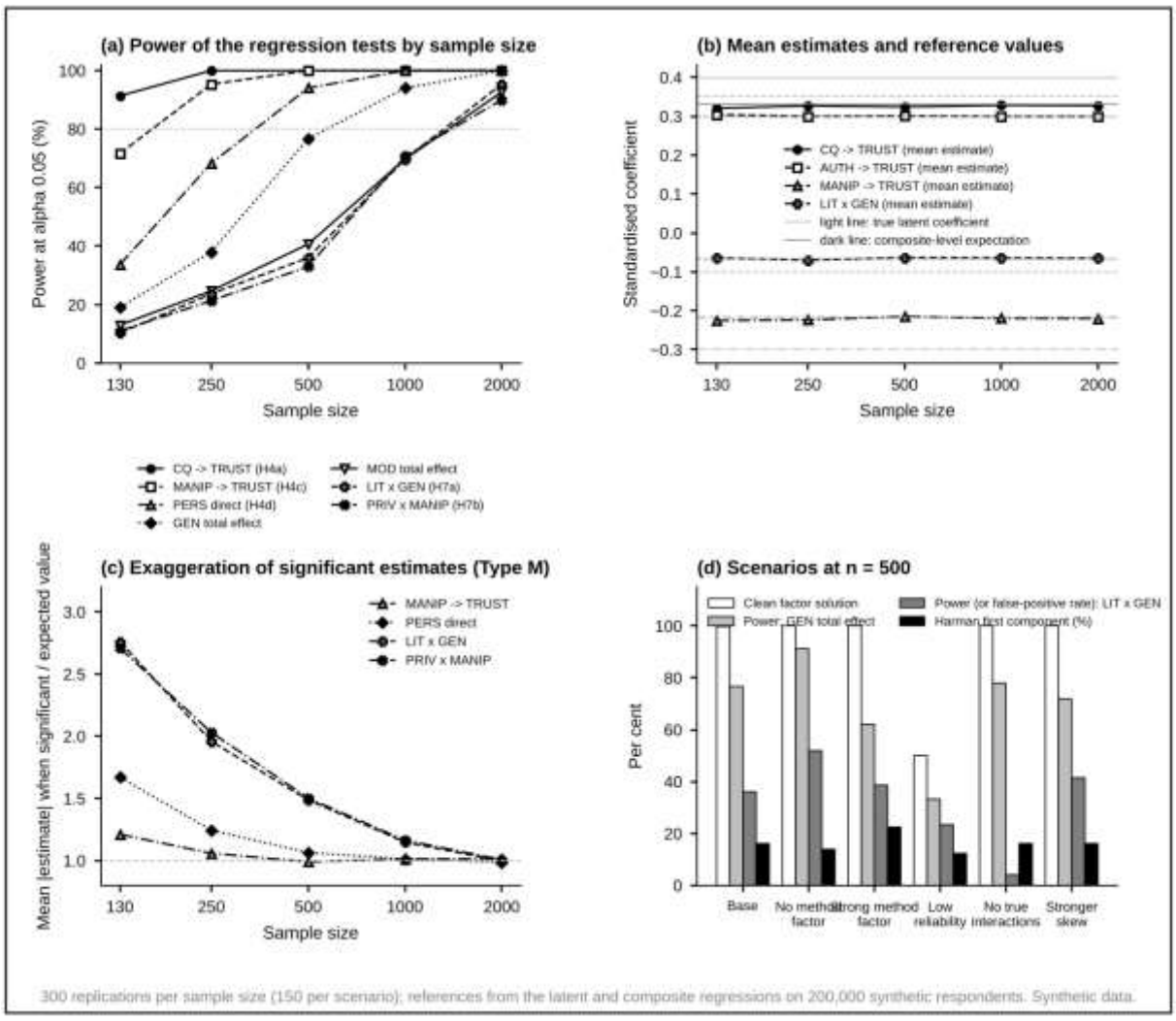


Table 14: Monte Carlo properties of the survey pipeline by sample size and scenario: reliability, factor recovery, Harman's test, power of the hypothesis tests, attenuation and exaggeration

Condition	Alpha / KMO	Clean EFA / k	Harman	H4a-c	H4d / CHAT	GEN / LAB / MOD	H7a / H7b	Indirect	Attenu-ation	Exagger-ation	R-sq.
Base, n = 130	0.74 / 0.73	36% / 12.6	0% (17%)	91% / 94% / 72%	34% / 27%	19% / 36% / 13%	10% / 11%	75% / 91%	0.81 / 0.75	1.21 / 2.76	0.44
Base, n = 250	0.74 / 0.81	91% / 12.5	0% (16%)	100% / 100% / 95%	68% / 53%	38% / 64% / 25%	24% / 21%	98% / 100%	0.82 / 0.75	1.06 / 1.96	0.41
Base, n = 500	0.75 / 0.86	100% / 12.6	0% (16%)	100% / 100% / 100%	94% / 86%	77% / 92% / 41%	36% / 33%	100% / 100%	0.82 / 0.72	0.99 / 1.49	0.39
Base, n = 1,000	0.75 / 0.88	100% / 12.7	0% (16%)	100% / 100% / 100%	100% / 99%	94% / 100% / 70%	70% / 71%	100% / 100%	0.83 / 0.73	1.01 / 1.15	0.38
Base, n = 2,000	0.75 / 0.89	100% / 12.8	0% (16%)	100% / 100% / 100%	100% / 100%	100% / 100% / 93%	95% / 90%	100% / 100%	0.82 / 0.73	1.01 / 1.00	0.37
No method factor, n = 500	0.72 / 0.83	100% / 12.8	0% (14%)	100% / 100% / 100%	87% / 75%	91% / 79% / 69%	52% / 48%	100% / 100%	0.80 / 0.74	1.03 / 1.55	0.35
Strong method factor (0.35), n = 500	0.80 / 0.90	100% / 12.1	0% (22%)	100% / 100% / 100%	93% / 90%	62% / 99% / 23%	39% / 25%	100% / 100%	0.88 / 0.71	0.99 / 1.32	0.46
Low reliability (loadings 0.50 to 0.70), n = 500	0.58 / 0.80	50% / 12.9	0% (12%)	100% / 100% / 98%	87% / 80%	33% / 84% / 25%	23% / 23%	97% / 100%	0.70 / 0.55	0.77 / 1.41	0.30
No true interactions, n = 500	0.75 / 0.86	100% / 12.7	0% (16%)	100% / 100% / 100%	93% / 83%	78% / 93% / 49%	4% / 7%	100% / 100%	0.83 / 0.72	0.99 / –	0.38
Stronger agreement skew, n = 500	0.74 / 0.85	100% / 12.7	0% (16%)	100% / 100% / 100%	87% / 81%	72% / 93% / 44%	41% / 35%	100% / 100%	0.83 / 0.71	0.98 / 1.45	0.37

Interpretation and theoretical contribution

Read as a demonstration on known structure, the results say three things about AI-driven communication and trust that survive the caveat that the structure was assumed. The mechanisms sort by the perception they act on: personalisation and conversational agents work through communication quality and carry a manipulation cost that is small when personalisation is relevant; synthetic content works against trust through two routes, lower authenticity and higher manipulation, that a single "AI exposure" construct would merge; and labelling is the one exposure with a positive effect through both routes, which is why its total effect on trust (0.15 at the latent level) is positive even though labels lower the credibility of what they mark (Altay & Gilardi, 2024; Toff & Simon, 2025). Trust is the channel to behaviour: the direct effects of the perceptions on the intentions are small once trust is in the model, and the indirect effects through trust are as large as the trust coefficients themselves. Manipulation and privacy concern act directly, as the template study argued, and also as moderators of small size that a survey of a few hundred respondents cannot detect, which reconciles the template study's rejection of the moderator reading with the privacy-calculus prediction that concern conditions the trust cost of

surveillance (Srivastava & Gurme, 2026; Dinev & Hart, 2006). The theoretical contribution of the framework is the separation of the three perceptions and the placement of trust as the mediator, which turns the elaboration likelihood model, the machine heuristic and the privacy calculus into three measurable routes and makes the opposing effects of labels and of synthetic content visible in one model (Petty & Cacioppo, 1986; Sundar, 2020; Hancock et al., 2020).

Discussion and implications

Theoretical implications

The study makes four theoretical points. First, it extends the technology acceptance model, the elaboration likelihood model and the privacy calculus, which the template study integrated, with the trust typology, the machine heuristic and the AI-mediated communication perspective into a single framework in which AI-driven communication is perceived on three dimensions and trusted on three attributes (Davis, 1989; Petty & Cacioppo, 1986; Dinev & Hart, 2006; Mayer et al., 1995; Sundar, 2020; Hancock et al., 2020). The empirical demonstration shows that the three perceptions are separable (thirteen clean factors, discriminant validity by the Fornell-Larcker criterion), that they carry different antecedents (personalisation and agents on quality; synthetic content and labels on authenticity; synthetic content, personalisation, labels, moderation and concern on manipulation) and that each has an independent effect on trust of the order of 0.2 to 0.4 in standardised units. A single "AI personalisation" construct cannot represent the finding that the same mechanism raises one perception and lowers another, which is the finding that the personalisation paradox and the labelling dilemma both depend on (Aguirre et al., 2015; Altay & Gilardi, 2024).

Second, the study supports the template study's reading of ethical and privacy concerns as direct determinants of trust rather than as moderators, and qualifies it (Srivastava & Gurme, 2026). Perceived manipulation and surveillance is the third-largest predictor of trust in every sample and the only perception with a direct negative effect on purchase intention beyond trust; privacy concern acts on trust through perceived manipulation and not directly. But the two moderation hypotheses are also true in the data-generating model, and the demonstration shows why a survey of 130 or even 500 respondents would reject them: an interaction of 0.10 between two measured composites has an expected estimate of 0.065 and a power of 10% to 36% at those sizes. The theoretical claim that concern conditions the trust cost of surveillance, and that literacy conditions the trust cost of synthetic content, is therefore not refuted by null interactions in small samples; it is untested by them. This is the Type S and Type M reasoning of the credibility literature applied to a specific design (Gelman & Carlin, 2014; Simmons et al., 2011).

Third, the study separates the bivariate and the multivariate reading of AI-generated content. In every sample the Kruskal-Wallis test says that heavy consumers of synthetic content trust brands more; the regression says that, holding personalisation and agent interaction constant, synthetic content lowers trust through authenticity and manipulation. Both are true of the data: heavy users see more of everything. Survey studies in this literature that compare groups on a single exposure should expect this confounding, because every AI mechanism on a platform is correlated with use (Zarouali et al., 2021; Boerman et al., 2017). Fourth, trust mediates: the perceptions reach the intentions almost entirely through trust, which is the structural claim of trust-based commerce models (Gefen et al., 2003; Kim et al., 2008; Pavlou, 2003) and the reason the harms to trust from undisclosed synthetic content and surveillance are harms to engagement and purchase.

Managerial implications

For brands and platforms the demonstration gives five implications, each conditional on the assumed structure holding in the field. Personalisation and conversational agents are worth investing in because they raise communication quality, the largest route to trust, but the investment pays only if the agent is competent and the relevance is real: the direct effect of personalisation on trust exists alongside its effect on perceived manipulation, and the net effect turns on which dominates for a given consumer, which the privacy calculus literature has long shown (Aguirre et al., 2015; Bleier & Eisenbeiss, 2015; Komiak & Benbasat, 2006). Synthetic content is a trust cost: undisclosed AI-generated posts, replies and reviews lower authenticity and raise manipulation, and the cost compounds through trust into engagement and purchase; the experiments that find AI-written text believed less were not an artefact of novelty (Jakesch et al., 2019; Longoni et al., 2022; Luo et al., 2019). Labelling is the countervailing

investment: in the model it is the only exposure that improves both authenticity and manipulation, so a brand that discloses its AI use and preserves content credentials gains trust on net even though each labelled item is believed a little less (Altay & Gilardi, 2024; Toff & Simon, 2025; Coalition for Content Provenance and Authenticity, 2026). Conversational agents should be disclosed and should escalate to a person, because interaction quality raises authenticity only when the agent is experienced as competent and honest (Mozafari et al., 2022; Crolic et al., 2022; Adam et al., 2021). Moderation should be explained, because unexplained removals and reach reductions raise perceived manipulation and, in the model, lower communication quality; the effects are small per consumer but the exposure is wide (Gorwa et al., 2020; Eslami et al., 2015).

For researchers and analytics teams the implications are methodological. A survey of 130 respondents is adequate to find the two or three strongest perception effects and to validate an instrument with a clean factor structure, but it will miss most exposure effects, all interactions and most indirect effects, exaggerate the ones it finds by a factor of two or more, and report a factor solution that replicates less than half the time; a survey of 500 detects effects of 0.15 or larger with reasonable power; interactions of 0.10 need 1,000 to 2,000 respondents. Composite-score regression attenuates effects by 15 to 30% and interactions by more, so that the comparison of coefficients across studies with different reliabilities is not a comparison of effects, and structural models with latent variables or reliability corrections should be used when the sample allows (Kline, 2016; Westland, 2010). Harman's test does not detect a method factor that inflates reliability and explained variance by a quarter, so marker variables or multi-source designs are the remedy, not the test (Podsakoff et al., 2003). And the direction of a bivariate comparison on a confounded exposure should not be reported as the effect of that exposure.

Policy and regulatory mapping

Table 15 maps the constructs to the instruments that govern them. The transparency obligations of the EU AI Act (Article 50) require disclosure of AI interaction and the marking of synthetic content, which in the framework act on labelling exposure, authenticity and manipulation; the Digital Services Act's rules on recommender transparency and on deceptive interfaces act on personalisation and perceived manipulation; the FTC Act, the FTC's dark-pattern report and the 2024 consumer review rule, which prohibits fake and AI-generated reviews and testimonials, act on authenticity and manipulation; data protection law and the California AI Transparency Act act on the data side of personalisation and on provenance; and content credentials give labelling a technical substrate (European Parliament and Council, 2024; European Parliament and Council, 2022; U.S. Congress, 2023; Federal Trade Commission, 2022; Federal Trade Commission, 2024; European Parliament and Council, 2016; California State Legislature, 2024; Coalition for Content Provenance and Authenticity, 2026). The model's finding that labelling has a positive net effect on trust is the behavioural counterpart of these obligations, and its finding that labels lower the credibility of what they mark is the cost that the obligations impose on honest users of AI, which the public-opinion evidence on comfort with AI-generated news also shows (Fletcher & Nielsen, 2024; Newman et al., 2024). The ethics guidance for AI-mediated communication (transparency, accountability, autonomy, non-manipulation) maps onto the same constructs (Hancock et al., 2020; Jobin et al., 2019; Floridi et al., 2018). The instruments are read as of their cited editions; implementing acts, enforcement priorities and state laws change, and the mapping is a reading for design, not legal advice.

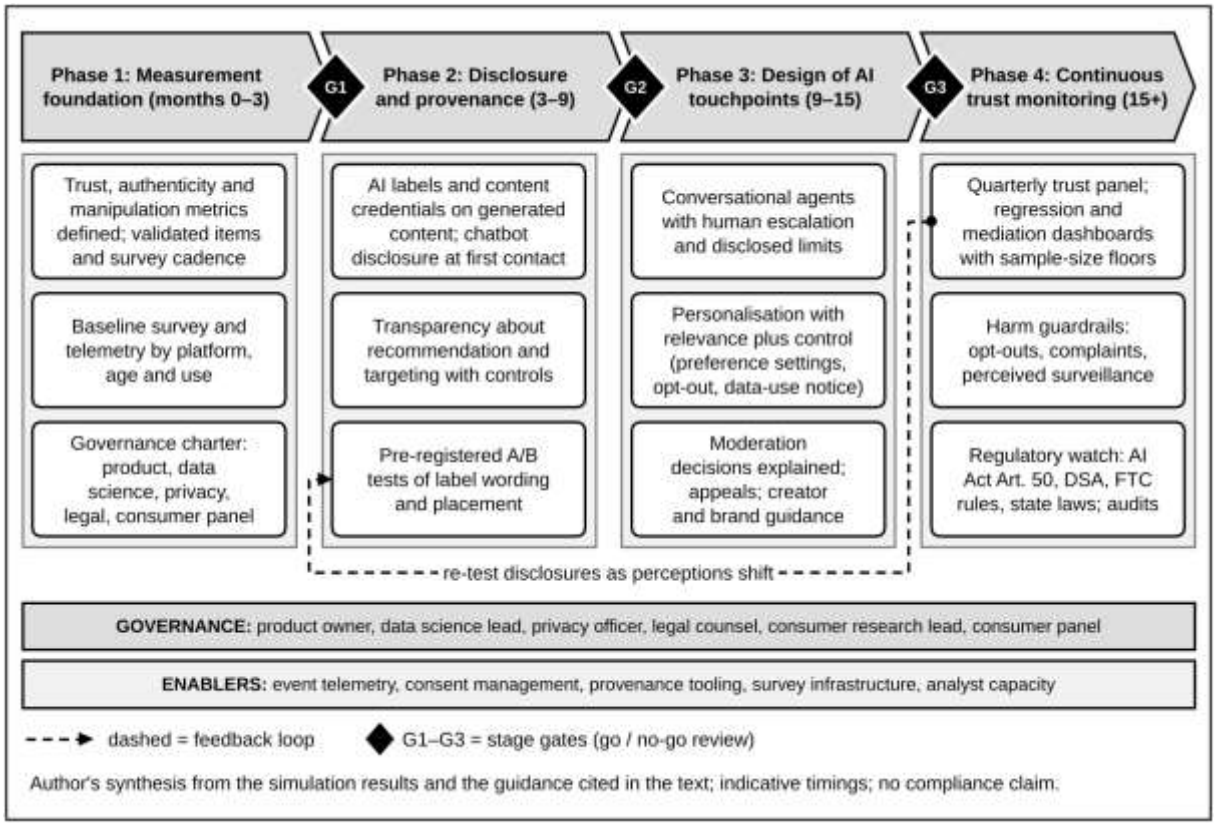
Table 15: Mapping of the framework's constructs and mechanisms to regulatory, standards and ethics instruments and to design and governance implications

Instrument or guidance	What it requires	Constructs and mechanisms affected	Design and governance implication
EU AI Act, Article 50 (European Parliament and Council, 2024)	Providers and deployers must disclose AI interaction (chatbots) and label synthetic audio, image, video and text content in specified cases	LAB, CHAT, GEN; authenticity and manipulation perceptions	Disclosure at first contact with an agent; machine-readable and visible labels on generated content
Digital Services Act, Articles 25 and 27 (European Parliament and Council, 2022)	No deceptive or manipulative interface design; recommender systems must explain main parameters and offer options	PERS, MOD; manipulation and surveillance	Explain and control personalisation; explain moderation decisions and offer appeal
FTC Act section 5 and the consumer review rule (U.S. Congress, 2023; Federal Trade Commission, 2022, 2024)	Unfair or deceptive practices unlawful; fake or AI-generated reviews and testimonials prohibited; dark patterns enforceable	GEN, AUTH (reviews), MANIP (false urgency)	No synthetic reviews; no undisclosed AI personas selling; urgency cues only when true
GDPR and state privacy law (European Parliament and Council, 2016; California State Legislature, 2024)	Lawful basis, purpose limitation and rights over profiling; California requires AI detection tools and latent disclosures from large generative providers	PERS, PRIV; data use for personalisation and generation	Consent and control settings; data-use notice in personalised communication; provenance support
Content credentials (Coalition for Content Provenance and Authenticity, 2026)	Cryptographically signed provenance manifests for media, including AI generation history	LAB, AUTH	Attach and preserve credentials; show provenance in the interface
Ethics of AI-mediated communication (Hancock et al., 2020; Jobin et al., 2019; Floridi et al., 2018)	Transparency, accountability, autonomy and non-manipulation as principles	All exposure constructs; trust	Public AI-use policy for brand communication; human escalation; audit of automated replies
Survey and measurement standards (Podsakoff et al., 2003; Flake & Fried, 2020; Morris et al., 2019)	Validated measures, method-bias controls, pre-specified analysis, power and simulation evidence	The evidence base itself	Pre-registered instruments and sample-size floors for trust monitoring

Implementation roadmap

Figure 12 arranges the implications as a roadmap with stage gates. Phase 1 (months 0 to 3) builds the measurement foundation: trust, authenticity and manipulation metrics with validated items and a survey cadence, a baseline survey and telemetry by platform, age and use, and a governance charter across product, data science, privacy, legal and consumer research with a consumer panel. Phase 2 (months 3 to 9) delivers disclosure and provenance: AI labels and content credentials on generated content, chatbot disclosure at first contact, transparency about recommendation and targeting with controls, and pre-registered tests of label wording and placement. Phase 3 (months 9 to 15) redesigns the AI touchpoints: conversational agents with human escalation and disclosed limits, personalisation with relevance plus control, and explained moderation with appeals. Phase 4 (from month 15) is continuous trust monitoring: a quarterly trust panel with regression and mediation dashboards that respect sample-size floors, harm guardrails (opt-outs, complaints, perceived surveillance) and a regulatory watch. The feedback loop from Phase 4 to Phase 2 re-tests disclosures as perceptions shift, because the credibility cost of a label is a property of consumers' current beliefs about AI, which are moving (Tyson & Kikuchi, 2023; Edelman, 2025).

Figure 12: Roadmap for trustworthy AI-driven communication in social media: four phases with stage gates, feedback, governance and enablers



Limitations and future research

Table 16 lists the limitations. The decisive one is that the respondents are synthetic: the effects are consequences of the data-generating model, whose signs and sizes were chosen by the author with reference to the literature, so the study shows what a survey of this design can recover from a world with that structure, not what the world is like. The second is the cross-sectional, linear, composite-score design inherited from the template, which the Monte Carlo analysis evaluates but does not escape. The third is the simplicity of the mechanisms: one method factor, two interactions, one platform per respondent, no feedback from behaviour to exposure. Future research should administer the instrument in the field with probability sampling and compare the estimates with the testbed's predictions at the realised sample size; should manipulate labels, disclosure and personalisation controls experimentally, where the causal order is identifiable; should estimate the model with latent variables at adequate sample sizes; and should extend the testbed with dynamics, so that trust formed in one period changes exposure in the next, which is the feedback that AI-driven communication introduces and that no cross-sectional survey can see (Morris et al., 2019; Flake & Fried, 2020; Hancock et al., 2020).

Table 16: Limitations of the study and what would address them

Limitation	Why it matters	What would address it
Synthetic respondents	The effects reported are consequences of the data-generating model, not measurements of any consumer population; signs and sizes are assumptions anchored to the literature	Field administration of the instrument with probability sampling; comparison of field estimates with the testbed's predictions
Cross-sectional structure	Causal order among perceptions, trust and intentions is assumed; a survey of this design cannot establish it	Longitudinal panels; experiments that manipulate labels, disclosure and personalisation controls
Composite-score regression	Measurement error attenuates coefficients by 15 to 35% here; interactions are attenuated most	Structural equation models with latent variables at adequate sample sizes; reliability corrections
Linear, additive model with two interactions	Threshold effects, non-linear trust dynamics and feedback from behaviour to exposure are absent	Dynamic and agent-based extensions; non-linear estimators
Method factor as a single response style	Real common-method variance has several sources (consistency motif, social desirability, item context)	Marker variables; multi-wave or multi-source designs
Single platform context per respondent	Cross-platform and cross-cultural differences are represented only by a platform category	Multi-country samples; platform-specific instruments
Regulatory mapping	Instruments change and differ by jurisdiction; the mapping summarises texts as of their cited editions and is not legal advice	Legal review; monitoring of AI Act implementing acts, DSA enforcement and FTC rules

Conclusion

AI-driven information technology has changed social media communication from a human-to-human exchange mediated by a platform into a continuous, adaptive process in which models decide what a consumer sees, write some of what they see, answer them, remove what they post and label what remains. This article asked what that does to consumer trust and what a survey study of the standard design can learn about it. On a transparent synthetic testbed built from an integrative framework, personalisation and conversational agents raise trust through communication quality, synthetic content lowers it through authenticity and manipulation, labelling raises it through both, manipulation and surveillance is a direct cost, and trust carries the effects to engagement and purchase intention. The same testbed shows that the pipeline of the template study, applied to 500 synthetic respondents, recovers thirteen of eighteen directional hypotheses with the designed signs, attenuates the coefficients by 15 to 30%, misses the small exposure effects and both interactions, and, when applied to 130 respondents, keeps ten hypotheses, loses most indirect effects, exaggerates what it finds and reports a factor structure that replicates in a minority of samples; a bivariate comparison of trust across exposure to AI-generated content points the wrong way at every sample size. None of the effect sizes is a measurement of any population; all of them are reasons to build the instrument, the sample and the governance of Figures 2, 3 and 12 before concluding that an AI feature has, or has not, changed what consumers trust.

REFERENCES

[1]. Acquisti, A., Brandimarte, L., & Loewenstein, G. (2015). Privacy and human behavior in the age of information. *Science*, *347*(6221), 509–514. https://doi.org/10.1126/science.aaa1465

[2]. Adam, M., Wessel, M., & Benlian, A. (2021). AI-based chatbots in customer service and their effects on user compliance. *Electronic Markets*, *31*(2), 427–445. https://doi.org/10.1007/s12525-020-00414-7

[3]. Aguirre, E., Mahr, D., Grewal, D., de Ruyter, K., & Wetzels, M. (2015). Unraveling the personalization paradox: The effect of information collection and trust-building strategies on online advertisement effectiveness. *Journal of Retailing*, *91*(1), 34–49. https://doi.org/10.1016/j.jretai.2014.09.005

[4]. Aiken, L. S., & West, S. G. (1991). *Multiple regression: Testing and interpreting interactions*. Sage. https://www.sagepub.com/shop/buy-a-book/multiple-regression-1-3045

[5]. Ajzen, I. (1991). The theory of planned behavior. *Organizational Behavior and Human Decision Processes*, *50*(2), 179–211. https://doi.org/10.1016/0749-5978(91)90020-T

[6]. Altay, S., & Gilardi, F. (2024). People are skeptical of headlines labeled as AI-generated, even if true or human-made, because they assume full AI automation. *PNAS Nexus*, *3*(10), Article pgae403. https://doi.org/10.1093/pnasnexus/pgae403

[7]. Araujo, T., Helberger, N., Kruikemeier, S., & de Vreese, C. H. (2020). In AI we trust? Perceptions about automated decision-making by artificial intelligence. *AI & Society*, *35*(3), 611–623. https://doi.org/10.1007/s00146-019-00931-w

[8]. Awad, N. F., & Krishnan, M. S. (2006). The personalization privacy paradox: An empirical evaluation of information transparency and the willingness to be profiled online for personalization. *MIS Quarterly*, *30*(1), 13–28. https://doi.org/10.2307/25148715

[9]. Baron, R. M., & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, *51*(6), 1173–1182. https://doi.org/10.1037/0022-3514.51.6.1173

[10]. Barth, S., & de Jong, M. D. T. (2017). The privacy paradox: Investigating discrepancies between expressed privacy concerns and actual online behavior: A systematic literature review. *Telematics and Informatics*, *34*(7), 1038–1058. https://doi.org/10.1016/j.tele.2017.04.013

[11]. Bartlett, M. S. (1950). Tests of significance in factor analysis. *British Journal of Statistical Psychology*, *3*(2), 77–85. https://doi.org/10.1111/j.2044-8317.1950.tb00285.x

[12]. Bedué, P., & Fritzsche, A. (2022). Can we trust AI? An empirical investigation of trust requirements and guide to successful AI adoption. *Journal of Enterprise Information Management*, *35*(2), 530–549. https://doi.org/10.1108/JEIM-06-2020-0233

[13]. Bleier, A., & Eisenbeiss, M. (2015). The importance of trust for personalized online advertising. *Journal of Retailing*, *91*(3), 390–409. https://doi.org/10.1016/j.jretai.2015.04.001

[14]. Boerman, S. C., Kruikemeier, S., & Zuiderveen Borgesius, F. J. (2017). Online behavioral advertising: A literature review and research agenda. *Journal of Advertising*, *46*(3), 363–376. https://doi.org/10.1080/00913367.2017.1339368

[15]. California State Legislature. (2024). *Senate Bill No. 942, California AI Transparency Act*. California Legislative Information. https://leginfo.legislature.ca.gov/faces/billNavClient.xhtml?bill_id=202320240SB942

[16]. Castelo, N., Bos, M. W., & Lehmann, D. R. (2019). Task-dependent algorithm aversion. *Journal of Marketing Research*, *56*(5), 809–825. https://doi.org/10.1177/0022243719851788

[17]. Chung, M., Ko, E., Joung, H., & Kim, S. J. (2020). Chatbot e-service and customer satisfaction regarding luxury brands. *Journal of Business Research*, *117*, 587–595. https://doi.org/10.1016/j.jbusres.2018.10.004

[18]. Coalition for Content Provenance and Authenticity. (2026). *C2PA technical specification (version 2.4)*. C2PA. https://spec.c2pa.org/specifications/specifications/2.4/specs/C2PA_Specification.html

[19]. Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates. https://www.routledge.com/Statistical-Power-Analysis-for-the-Behavioral-Sciences/Cohen/p/book/9780805802832

[20]. Crolic, C., Thomaz, F., Hadi, R., & Stephen, A. T. (2022). Blame the bot: Anthropomorphism and anger in customer-chatbot interactions. *Journal of Marketing*, *86*(1), 132–148. https://doi.org/10.1177/00222429211045687

[21]. Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297–334. https://doi.org/10.1007/BF02310555

[22]. Culnan, M. J., & Armstrong, P. K. (1999). Information privacy concerns, procedural fairness, and impersonal trust: An empirical investigation. *Organization Science*, *10*(1), 104–115. https://doi.org/10.1287/orsc.10.1.104

[23]. Daft, R. L., & Lengel, R. H. (1986). Organizational information requirements, media richness and structural design. *Management Science*, *32*(5), 554–571. https://doi.org/10.1287/mnsc.32.5.554

[24]. Davenport, T., Guha, A., Grewal, D., & Bressgott, T. (2020). How artificial intelligence will change the future of marketing. *Journal of the Academy of Marketing Science*, *48*(1), 24–42. https://doi.org/10.1007/s11747-019-00696-0

[25]. Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, *13*(3), 319–340. https://doi.org/10.2307/249008

[26]. DellaVigna, S., & Linos, E. (2022). RCTs to scale: Comprehensive evidence from two nudge units. *Econometrica*, *90*(1), 81–116. https://doi.org/10.3982/ECTA18709

[27]. Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, *144*(1), 114–126. https://doi.org/10.1037/xge0000033

[28]. Dinev, T., & Hart, P. (2006). An extended privacy calculus model for e-commerce transactions. *Information Systems Research*, *17*(1), 61–80. https://doi.org/10.1287/isre.1060.0080

[29]. Dwivedi, Y. K., Hughes, L., Ismagilova, E., Aarts, G., Coombs, C., Crick, T., Duan, Y., Dwivedi, R., Edwards, J., Eirug, A., Galanos, V., Ilavarasan, P. V., Janssen, M., Jones, P., Kar, A. K., Kizgin, H., Kronemann, B., Lal, B., Lucini, B., ... Williams, M. D. (2021a). Artificial Intelligence (AI): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy. *International Journal of Information Management*, *57*, Article 101994. https://doi.org/10.1016/j.ijinfomgt.2019.08.002

[30]. Dwivedi, Y. K., Ismagilova, E., Hughes, D. L., Carlson, J., Filieri, R., Jacobson, J., Jain, V., Karjaluoto, H., Kefi, H., Krishen, A. S., Kumar, V., Rahman, M. M., Raman, R., Rauschnabel, P. A., Rowley, J., Salo, J., Tran, G. A., & Wang, Y. (2021b). Setting the future of digital and social media marketing research: Perspectives and research propositions. *International Journal of Information Management*, *59*, Article 102168. https://doi.org/10.1016/j.ijinfomgt.2020.102168

[31]. Edelman. (2025). *2025 Edelman Trust Barometer: Trust and the crisis of grievance*. Edelman. https://www.edelman.com/trust/2025/trust-barometer

[32]. Eslami, M., Rickman, A., Vaccaro, K., Aleyasen, A., Vuong, A., Karahalios, K., Hamilton, K., & Sandvig, C. (2015). "I always assumed that I wasn't really that close to [her]": Reasoning about invisible algorithms in news feeds. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)* (pp. 153–162). ACM. https://doi.org/10.1145/2702123.2702556

[33]. European Parliament and Council. (2016). *Regulation (EU) 2016/679 on the protection of natural persons with regard to the processing of personal data (General Data Protection Regulation)*. Official Journal of the European Union, L 119, 1-88. https://eur-lex.europa.eu/eli/reg/2016/679/oj

[34]. European Parliament and Council. (2022). *Regulation (EU) 2022/2065 on a Single Market for Digital Services (Digital Services Act)*. Official Journal of the European Union. https://eur-lex.europa.eu/eli/reg/2022/2065/oj

[35]. European Parliament and Council. (2024). *Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)*. Official Journal of the European Union. https://eur-lex.europa.eu/eli/reg/2024/1689/oj

[36]. Federal Trade Commission. (2022). *Bringing dark patterns to light (FTC staff report)*. Federal Trade Commission. https://www.ftc.gov/reports/bringing-dark-patterns-light

[37]. Federal Trade Commission. (2024). *Trade regulation rule on the use of consumer reviews and testimonials (16 CFR Part 465)* (89 FR 68034). Federal Register. https://www.federalregister.gov/documents/2024/08/22/2024-18519/trade-regulation-rule-on-the-use-of-consumer-reviews-and-testimonials

[38]. Ferrara, E., Varol, O., Davis, C., Menczer, F., & Flammini, A. (2016). The rise of social bots. *Communications of the ACM*, *59*(7), 96–104. https://doi.org/10.1145/2818717

[39]. Flake, J. K., & Fried, E. I. (2020). Measurement schmeasurement: Questionable measurement practices and how to avoid them. *Advances in Methods and Practices in Psychological Science*, *3*(4), 456–465. https://doi.org/10.1177/2515245920952393

[40]. Fletcher, R., & Nielsen, R. K. (2024). *What does the public in six countries think of generative AI in news?*. Reuters Institute for the Study of Journalism. https://doi.org/10.60625/risj-4zb8-cg87

[41]. Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People: An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds and Machines*, *28*(4), 689–707. https://doi.org/10.1007/s11023-018-9482-5

[42]. Fornell, C., & Larcker, D. F. (1981). Evaluating structural equation models with unobservable variables and measurement error. *Journal of Marketing Research*, *18*(1), 39–50. https://doi.org/10.1177/002224378101800104

[43]. Gefen, D., Karahanna, E., & Straub, D. W. (2003). Trust and TAM in online shopping: An integrated model. *MIS Quarterly*, *27*(1), 51–90. https://doi.org/10.2307/30036519

[44]. Gelman, A., & Carlin, J. (2014). Beyond power calculations: Assessing Type S (sign) and Type M (magnitude) errors. *Perspectives on Psychological Science*, *9*(6), 641–651. https://doi.org/10.1177/1745691614551642

[45]. Gillespie, T. (2018). *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press.

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

[47]. Gorwa, R., Binns, R., & Katzenbach, C. (2020). Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society*, *7*(1), Article 205395171989794. https://doi.org/10.1177/2053951719897945

[48]. Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2019). *Multivariate data analysis* (8th ed.). Cengage. https://books.google.com/books?vid=ISBN9781473756540

[49]. Hancock, J. T., Naaman, M., & Levy, K. (2020). AI-mediated communication: Definition, research agenda, and ethical considerations. *Journal of Computer-Mediated Communication*, *25*(1), 89–100. https://doi.org/10.1093/jcmc/zmz022

[50]. Harman, H. H. (1976). *Modern factor analysis* (3rd ed.). University of Chicago Press. https://openlibrary.org/isbn/0226316521

[51]. Hayes, A. F. (2018). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach* (2nd ed.). Guilford Press. https://books.google.com/books?vid=ISBN9781462534654

[52]. Hennig-Thurau, T., Gwinner, K. P., Walsh, G., & Gremler, D. D. (2004). Electronic word-of-mouth via consumer-opinion platforms: What motivates consumers to articulate themselves on the Internet? *Journal of Interactive Marketing*, *18*(1), 38–52. https://doi.org/10.1002/dir.10073

[53]. Henseler, J., Ringle, C. M., & Sarstedt, M. (2015). A new criterion for assessing discriminant validity in variance-based structural equation modeling. *Journal of the Academy of Marketing Science*, *43*(1), 115–135. https://doi.org/10.1007/s11747-014-0403-8

[54]. Hohenstein, J., Kizilcec, R. F., DiFranzo, D., Aghajari, Z., Mieczkowski, H., Levy, K., Naaman, M., Hancock, J., & Jung, M. F. (2023). Artificial intelligence in communication impacts language and social relationships. *Scientific Reports*, *13*(1), Article 5487. https://doi.org/10.1038/s41598-023-30938-9

[55]. Huang, M.-H., & Rust, R. T. (2021). A strategic framework for artificial intelligence in marketing. *Journal of the Academy of Marketing Science*, *49*(1), 30–50. https://doi.org/10.1007/s11747-020-00749-9

[56]. Jacovi, A., Marasović, A., Miller, T., & Goldberg, Y. (2021). Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in AI. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)* (pp. 624–635). ACM. https://doi.org/10.1145/3442188.3445923

[57]. Jakesch, M., French, M., Ma, X., Hancock, J. T., & Naaman, M. (2019). AI-mediated communication: How the perception that profile text was written by AI affects trustworthiness. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)* (pp. 1–13). ACM. https://doi.org/10.1145/3290605.3300469

[58]. Jakesch, M., Hancock, J. T., & Naaman, M. (2023). Human heuristics for AI-generated language are flawed. *Proceedings of the National Academy of Sciences*, *120*(11), Article e2208839120. https://doi.org/10.1073/pnas.2208839120

[59]. Jamieson, S. (2004). Likert scales: How to (ab)use them. *Medical Education*, *38*(12), 1217–1218. https://doi.org/10.1111/j.1365-2929.2004.02012.x

[60]. Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, *1*(9), 389–399. https://doi.org/10.1038/s42256-019-0088-2

[61]. Kaiser, H. F. (1974). An index of factorial simplicity. *Psychometrika*, *39*(1), 31–36. https://doi.org/10.1007/BF02291575

[62]. Kaplan, A. M., & Haenlein, M. (2010). Users of the world, unite! The challenges and opportunities of Social Media. *Business Horizons*, *53*(1), 59–68. https://doi.org/10.1016/j.bushor.2009.09.003

[63]. Kaplan, A., & Haenlein, M. (2019). Siri, Siri, in my hand: Who's the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence. *Business Horizons*, *62*(1), 15–25. https://doi.org/10.1016/j.bushor.2018.08.004

[64]. Kelly, S., Kaye, S.-A., & Oviedo-Trespalacios, O. (2023). What factors contribute to the acceptance of artificial intelligence? A systematic review. *Telematics and Informatics*, *77*, Article 101925. https://doi.org/10.1016/j.tele.2022.101925

[65]. Kietzmann, J. H., Hermkens, K., McCarthy, I. P., & Silvestre, B. S. (2011). Social media? Get serious! Understanding the functional building blocks of social media. *Business Horizons*, *54*(3), 241–251. https://doi.org/10.1016/j.bushor.2011.01.005

[66]. Kim, D. J., Ferrin, D. L., & Rao, H. R. (2008). A trust-based consumer decision-making model in electronic commerce: The role of trust, perceived risk, and their antecedents. *Decision Support Systems*, *44*(2), 544–564. https://doi.org/10.1016/j.dss.2007.07.001

[67]. Kim, T. W., & Duhachek, A. (2020). Artificial intelligence and persuasion: A construal-level account. *Psychological Science*, *31*(4), 363–380. https://doi.org/10.1177/0956797620904985

[68]. Kizilcec, R. F. (2016). How much information? Effects of transparency on trust in an algorithmic interface. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)* (pp. 2390–2395). ACM. https://doi.org/10.1145/2858036.2858402

[69]. Kline, R. B. (2016). *Principles and practice of structural equation modeling* (4th ed.). Guilford Press. https://openlibrary.org/isbn/9781462523344

[70]. Komiak, S. Y. X., & Benbasat, I. (2006). The effects of personalization and familiarity on trust and adoption of recommendation agents. *MIS Quarterly*, *30*(4), 941–960. https://doi.org/10.2307/25148760

[71]. Kreps, S., McCain, R. M., & Brundage, M. (2022). All the news that's fit to fabricate: AI-generated text as a tool of media misinformation. *Journal of Experimental Political Science*, *9*(1), 104–117. https://doi.org/10.1017/XPS.2020.37

[72]. Kruskal, W. H., & Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, *47*(260), 583–621. https://doi.org/10.1080/01621459.1952.10483441

[73]. Lankton, N. K., McKnight, D. H., & Tripp, J. (2015). Technology, humanness, and trust: Rethinking trust in technology. *Journal of the Association for Information Systems*, *16*(10), 880–918. https://doi.org/10.17705/1jais.00411

[74]. Law, A. M. (2015). *Simulation modeling and analysis* (5th ed.). McGraw-Hill Education. https://www.mheducation.com/

[75]. Liddell, T. M., & Kruschke, J. K. (2018). Analyzing ordinal data with metric models: What could possibly go wrong? *Journal of Experimental Social Psychology*, *79*, 328–348. https://doi.org/10.1016/j.jesp.2018.08.009

[76]. Logg, J. M., Minson, J. A., & Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes*, *151*, 90–103. https://doi.org/10.1016/j.obhdp.2018.12.005

[77]. Long, D., & Magerko, B. (2020). What is AI literacy? Competencies and design considerations. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)* (pp. 1–16). ACM. https://doi.org/10.1145/3313831.3376727

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

[79]. Longoni, C., Fradkin, A., Cian, L., & Pennycook, G. (2022). News from generative artificial intelligence is believed less. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)* (pp. 97–106). ACM. https://doi.org/10.1145/3531146.3533077

[80]. Lou, C., & Yuan, S. (2019). Influencer marketing: How message value and credibility affect consumer trust of branded content on social media. *Journal of Interactive Advertising*, *19*(1), 58–73. https://doi.org/10.1080/15252019.2018.1533501

[81]. Luo, X., Tong, S., Fang, Z., & Qu, Z. (2019). Frontiers: Machines vs. humans: The impact of artificial intelligence chatbot disclosure on customer purchases. *Marketing Science*, *38*(6), 937–947. https://doi.org/10.1287/mksc.2019.1192

[82]. MacCallum, R. C., Widaman, K. F., Zhang, S., & Hong, S. (1999). Sample size in factor analysis. *Psychological Methods*, *4*(1), 84–99. https://doi.org/10.1037/1082-989X.4.1.84

[83]. Malhotra, N. K., Kim, S. S., & Agarwal, J. (2004). Internet users' information privacy concerns (IUIPC): The construct, the scale, and a causal model. *Information Systems Research*, *15*(4), 336–355. https://doi.org/10.1287/isre.1040.0032

[84]. Mathur, A., Acar, G., Friedman, M. J., Lucherini, E., Mayer, J., Chetty, M., & Narayanan, A. (2019). Dark patterns at scale: Findings from a crawl of 11K shopping websites. *Proceedings of the ACM on Human-Computer Interaction*, *3*(CSCW), Article 81. https://doi.org/10.1145/3359183

[85]. 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

[86]. McKnight, D. H., Choudhury, V., & Kacmar, C. (2002). Developing and validating trust measures for e-commerce: An integrative typology. *Information Systems Research*, *13*(3), 334–359. https://doi.org/10.1287/isre.13.3.334.81

[87]. Morris, T. P., White, I. R., & Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, *38*(11), 2074–2102. https://doi.org/10.1002/sim.8086

[88]. Mozafari, N., Weiger, W. H., & Hammerschmidt, M. (2022). Trust me, I'm a bot: Repercussions of chatbot disclosure in different service frontline settings. *Journal of Service Management*, *33*(2), 221–245. https://doi.org/10.1108/JOSM-10-2020-0380

[89]. Muthén, L. K., & Muthén, B. O. (2002). How to use a Monte Carlo study to decide on sample size and determine power. *Structural Equation Modeling*, *9*(4), 599–620. https://doi.org/10.1207/S15328007SEM0904_8

[90]. Nass, C., & Moon, Y. (2000). Machines and mindlessness: Social responses to computers. *Journal of Social Issues*, *56*(1), 81–103. https://doi.org/10.1111/0022-4537.00153

[91]. Newman, N., Fletcher, R., Robertson, C. T., Ross Arguedas, A., & Nielsen, R. K. (2024). *Reuters Institute Digital News Report 2024*. Reuters Institute for the Study of Journalism. https://doi.org/10.60625/risj-vy6n-4v57

[92]. Norman, G. (2010). Likert scales, levels of measurement and the "laws" of statistics. *Advances in Health Sciences Education*, *15*(5), 625–632. https://doi.org/10.1007/s10459-010-9222-y

[93]. Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). McGraw-Hill.

[94]. Pavlou, P. A. (2003). Consumer acceptance of electronic commerce: Integrating trust and risk with the technology acceptance model. *International Journal of Electronic Commerce*, *7*(3), 101–134. https://doi.org/10.1080/10864415.2003.11044275

[95]. Paxton, P., Curran, P. J., Bollen, K. A., Kirby, J., & Chen, F. (2001). Monte Carlo experiments: Design and implementation. *Structural Equation Modeling*, *8*(2), 287–312. https://doi.org/10.1207/S15328007SEM0802_7

[96]. Pennycook, G., & Rand, D. G. (2021). The psychology of fake news. *Trends in Cognitive Sciences*, *25*(5), 388–402. https://doi.org/10.1016/j.tics.2021.02.007

[97]. Petty, R. E., & Cacioppo, J. T. (1986). The elaboration likelihood model of persuasion. *Advances in Experimental Social Psychology*, *19*, 123–205. https://doi.org/10.1016/S0065-2601(08)60214-2

[98]. Podsakoff, P. M., MacKenzie, S. B., Lee, J.-Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology*, *88*(5), 879–903. https://doi.org/10.1037/0021-9010.88.5.879

[99]. Preacher, K. J., & Hayes, A. F. (2008). Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. *Behavior Research Methods*, *40*(3), 879–891. https://doi.org/10.3758/BRM.40.3.879

[100]. Puntoni, S., Reczek, R. W., Giesler, M., & Botti, S. (2021). Consumers and artificial intelligence: An experiential perspective. *Journal of Marketing*, *85*(1), 131–151. https://doi.org/10.1177/0022242920953847

[101]. Rader, E., Cotter, K., & Cho, J. (2018). Explanations as mechanisms for supporting algorithmic transparency. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)* (Paper 103). ACM. https://doi.org/10.1145/3173574.3173677

[102]. Rousseau, D. M., Sitkin, S. B., Burt, R. S., & Camerer, C. (1998). Not so different after all: A cross-discipline view of trust. *Academy of Management Review*, *23*(3), 393–404. https://doi.org/10.5465/amr.1998.926617

[103]. Sands, S., Ferraro, C., Demsar, V., & Chandler, G. (2022). False idols: Unpacking the opportunities and challenges of falsity in the context of virtual influencers. *Business Horizons*, *65*(6), 777–788. https://doi.org/10.1016/j.bushor.2022.08.002

[104]. Sargent, R. G. (2013). Verification and validation of simulation models. *Journal of Simulation*, *7*(1), 12–24. https://doi.org/10.1057/jos.2012.20

[105]. Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, *52*(3-4), 591–611. https://doi.org/10.1093/biomet/52.3-4.591

[106]. Sheehan, B., Jin, H. S., & Gottlieb, U. (2020). Customer service chatbots: Anthropomorphism and adoption. *Journal of Business Research*, *115*, 14–24. https://doi.org/10.1016/j.jbusres.2020.04.030

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

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

[109]. Short, J., Williams, E., & Christie, B. (1976). *The social psychology of telecommunications*. John Wiley & Sons. https://archive.org/details/socialpsychology0000shor

[110]. Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632

[111]. Smith, H. J., Dinev, T., & Xu, H. (2011). Information privacy research: An interdisciplinary review. *MIS Quarterly*, *35*(4), 989–1015. https://doi.org/10.2307/41409970

[112]. Srivastava, R. K., & Gurme, V. M. (2026). Artificial intelligence and consumer behaviour on social media: A study of personalization, trust, and privacy. *Measurement: Digitalization*, *5*, Article 100021. https://doi.org/10.1016/j.meadig.2025.100021

[113]. Sundar, S. S. (2020). Rise of machine agency: A framework for studying the psychology of human-AI interaction (HAII). *Journal of Computer-Mediated Communication*, *25*(1), 74–88. https://doi.org/10.1093/jcmc/zmz026

[114]. Toff, B., & Simon, F. M. (2025). "Or they could just not use it?": The dilemma of AI disclosure for audience trust in news. *The International Journal of Press/Politics*, *30*(4), 881–903. https://doi.org/10.1177/19401612241308697

[115]. Tyson, A., & Kikuchi, E. (2023). *Growing public concern about the role of artificial intelligence in daily life*. Pew Research Center. https://www.pewresearch.org/short-reads/2023/08/28/growing-public-concern-about-the-role-of-artificial-intelligence-in-daily-life/

[116]. U.S. Congress. (2023). *Unfair methods of competition unlawful; prevention by Commission (15 U.S.C. 45, United States Code, 2023 edition)*. U.S. Government Publishing Office. https://www.govinfo.gov/content/pkg/USCODE-2023-title15/html/USCODE-2023-title15-chap2-subchapI-sec45.htm

[117]. Vaccari, C., & Chadwick, A. (2020). Deepfakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news. *Social Media + Society*, *6*(1), Article 2056305120903408. https://doi.org/10.1177/2056305120903408

[118]. Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User acceptance of information technology: Toward a unified view. *MIS Quarterly*, *27*(3), 425–478. https://doi.org/10.2307/30036540

[119]. Venkatesh, V., Thong, J. Y. L., & Xu, X. (2012). Consumer acceptance and use of information technology: Extending the unified theory of acceptance and use of technology. *MIS Quarterly*, *36*(1), 157–178. https://doi.org/10.2307/41410412

[120]. Walther, J. B. (1996). Computer-mediated communication: Impersonal, interpersonal, and hyperpersonal interaction. *Communication Research*, *23*(1), 3–43. https://doi.org/10.1177/009365096023001001

[121]. Wedel, M., & Kannan, P. K. (2016). Marketing analytics for data-rich environments. *Journal of Marketing*, *80*(6), 97–121. https://doi.org/10.1509/jm.15.0413

[122]. Westland, J. C. (2010). Lower bounds on sample size in structural equation modeling. *Electronic Commerce Research and Applications*, *9*(6), 476–487. https://doi.org/10.1016/j.elerap.2010.07.003

[123]. White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, *48*(4), 817–838. https://doi.org/10.2307/1912934

[124]. Xiao, B., & Benbasat, I. (2007). E-commerce product recommendation agents: Use, characteristics, and impact. *MIS Quarterly*, *31*(1), 137–209. https://doi.org/10.2307/25148784

[125]. Zarouali, B., Boerman, S. C., & de Vreese, C. H. (2021). Is this recommended by an algorithm? The development and validation of the algorithmic media content awareness scale (AMCA-scale). *Telematics and Informatics*, *62*, Article 101607. https://doi.org/10.1016/j.tele.2021.101607

[126]. Zuboff, S. (2019). *The age of surveillance capitalism: The fight for a human future at the new frontier of power*. PublicAffairs. https://www.hachettebookgroup.com/titles/shoshana-zuboff/the-age-of-surveillance-capitalism/9781610395694/