

VOL. XI ISSUE III, Summer (SEPTEMBER-2026)

GSSR

GLOBAL SOCIAL SCIENCES REVIEW

GLOBAL SOCIAL SCIENCES REVIEW (GSSR)

Title: Transparent but Trustworthy: Reconciling the AI Disclosure Paradox in Corporate and Customer Communication

Abstract

Regulators and public interest activists increasingly advocate for full transparency in customer-facing AI. However, an empirical meta-synthesis of 17 studies (N=14,820) reveals a critical transparency dilemma: uncontextualized raw AI disclosures trigger negative persuasion knowledge, reducing consumer trust by 1.12 to 1.48 points on a 7-point scale. This trust penalty erodes brand trust, perceived competence, and institutional legitimacy. Critically, this decline stems from unframed transparency rather than algorithm aversion itself. Combining AI disclosures with explicit Human-in-the-Loop (HITL) cues protects consumer trust and restores brand authenticity (M=5.56 and M=5.68). A cross-sectoral analysis across banking, healthcare, retail, aviation, and technology highlights widespread governance gaps, with 82% of top-performing firms asserting AI use without addressing AI governance. Based on a structural equation modeling path model (R²=0.58), the study provides a business communication framework with actionable guidance for practitioners and regulators.

Keywords: AI Disclosure, Customer Trust, Human-in-the-Loop, Persuasion Knowledge Model, Algorithm Aversion, Corporate Governance, Business Communication

Authors:

Sana Hussan: (Corresponding Author)

Doctoral Fellow, Business Administration, The University of the Potomac, Virginia, United States (US).
(Email: sanahusana1992@gmail.com)

Pages: 14-28

DOI: 10.31703/gssr.2026(XI-III).02

DOI link: [https://dx.doi.org/10.31703/gssr.2026\(XI-III\).02](https://dx.doi.org/10.31703/gssr.2026(XI-III).02)

Article link: <https://gssrjournal.com/article/transparent-but-trustworthy-reconciling-the-ai-disclosure-paradox-in-corporate-and-customer-communication>

Full-text Link: <https://gssrjournal.com/article/transparent-but-trustworthy-reconciling-the-ai-disclosure-paradox-in-corporate-and-customer-communication>

Pdf link: <https://www.gssrjournal.com/jadmin/Auther/31rvIolA2.pdf>

Global Social Sciences Review

p-ISSN: [2520-0348](https://doi.org/10.31703/gssr) **e-ISSN:** [2616-793x](https://doi.org/10.31703/gssr)

DOI(journal): 10.31703/gssr

Volume: XI (2026)

DOI (volume): 10.31703/gssr.2026(XI)

Issue: III Summer (Sept-2026)

DOI(Issue): 10.31703/gssr.2026(XI-III)

Home Page

www.gssrjournal.com

Volume: XI (2026)

<https://www.gssrjournal.com/Current-issue>

Issue: III-Summer (Sept 2026)

<https://www.gssrjournal.com/issue/11/3/2026>

Scope

<https://www.gssrjournal.com/about-us/scope>

Submission

<https://humaglobe.com/index.php/gssr/submissions>

Scan the QR to visit us



Google
scholar



Citing this Article

Article Serial	02
Article Title	Transparent but Trustworthy: Reconciling the AI Disclosure Paradox in Corporate and Customer Communication
Authors	Sana Hussan
DOI	10.31703/gssr.2026(XI-III).02
Pages	14-28
Year	2026
Volume	XI
Issue	III

Referencing & Citing Styles

APA	Hussan, S. (2026). Transparent but Trustworthy: Reconciling the AI Disclosure Paradox in Corporate and Customer Communication. <i>Global Social Sciences Review</i> , 11(3), 14-28. https://doi.org/10.31703/gssr.2026(XI-III).02
CHICAGO	Hussan, Sana. 2026. "Transparent but Trustworthy: Reconciling the AI Disclosure Paradox in Corporate and Customer Communication." <i>Global Social Sciences Review</i> 11 (3):14-28. doi: 10.31703/gssr.2026(XI-III).02.
HARVARD	HUSSAN, S. 2026. Transparent but Trustworthy: Reconciling the AI Disclosure Paradox in Corporate and Customer Communication. <i>Global Social Sciences Review</i> , 11, 14-28.
MHRA	Hussan, Sana. 2026. 'Transparent but Trustworthy: Reconciling the AI Disclosure Paradox in Corporate and Customer Communication', <i>Global Social Sciences Review</i> , 11: 14-28.
MLA	Hussan, Sana. "Transparent but Trustworthy: Reconciling the Ai Disclosure Paradox in Corporate and Customer Communication." <i>Global Social Sciences Review</i> 11.3 (2026): 14-28. Print.
OXFORD	Hussan, Sana (2026), 'Transparent but Trustworthy: Reconciling the AI Disclosure Paradox in Corporate and Customer Communication', <i>Global Social Sciences Review</i> , 11 (3), 14-28.
TURABIAN	Hussan, Sana. "Transparent but Trustworthy: Reconciling the Ai Disclosure Paradox in Corporate and Customer Communication." <i>Global Social Sciences Review</i> 11, no. 3 (2026): 14-28. https://dx.doi.org/10.31703/gssr.2026(XI-III).02 .

Transparent but Trustworthy: Reconciling the AI Disclosure Paradox in Corporate and Customer Communication



Sana Hussan (Corresponding Author)¹

¹ Doctoral Fellow, Business Administration, The University of the Potomac, Virginia, United States (US).
(Email: sanahusana1992@gmail.com)

Abstract

Regulators and public interest activists increasingly advocate for full transparency in customer-facing AI. However, an empirical meta-synthesis of 17 studies (N=14,820) reveals a critical transparency dilemma: uncontextualized raw AI disclosures trigger negative persuasion knowledge, reducing consumer trust by 1.12 to 1.48 points on a 7-point scale. This trust penalty erodes brand trust, perceived competence, and institutional legitimacy. Critically, this decline stems from unframed transparency rather than algorithm aversion itself. Combining AI disclosures with explicit Human-in-the-Loop (HITL) cues protects consumer trust and restores brand authenticity (M=5.56 and M=5.68). A cross-sectoral analysis across banking, healthcare, retail, aviation, and technology highlights widespread governance gaps, with 82% of top-performing firms asserting AI use without addressing AI governance. Based on a structural equation modeling path model (R²=0.58), the study provides a business communication framework with actionable guidance for practitioners and regulators.

Keywords: AI Disclosure, Customer Trust, Human-in-the-Loop, Persuasion Knowledge Model, Algorithm Aversion, Corporate Governance, Business Communication

Introduction:

The Operational Shift to Generative AI in Business Communication

Generative artificial intelligence (AI) has revolutionized how businesses communicate, create marketing content, and manage customer relationships across corporate functions. In a short time, AI tools that generate text automatically and algorithmic co-pilots have moved from optional administrative add-ons to mission-critical infrastructure. The AI tools that use large language models (LLMs), whether in-house or outsourced, can be found in the customer service departments of most enterprises, marketing agencies, financial advisory firms, and healthcare communications companies as they routinely create personalized responses, write policy updates, create targeted promotional copy, and automate real-time service interactions. Advocates point to groundbreaking operational efficiencies with generative AI, including major reductions in content production latency that let businesses move from concept to completion faster; granular audience micro-segmentation that enables a hyper-targeted approach; and significant labor savings.

However, as businesses trust AI systems more, the point of vulnerability has shifted from technological feasibility to stakeholder trust and institutional governance. Customer-facing business communication is like the company's back-office analytics, but instead represents the company's image, values, and relationships. Whether sending a routine service notification, a promotional message, a financial report, or medical guidance, recipients assess not only the message's content but also the sender's implied motives, sincerity, and humanness. This has given rise to tensions that are deep and acute between "scale" and "relationship" about the speed of deployment of algorithmic communication tools.

The Regulatory Landscape and Mandatory Transparency Demands

Parliament and regulators worldwide have issued broad transparency mandates. The European Union Artificial Intelligence Act (EU AI Act), which came into effect in 2024 and has a series of enforcement dates through 2026, includes specific transparency requirements: users of AI systems designed to communicate directly with natural persons must be told they are communicating with an artificial system. Similar controls exist in North America and the Asia-Pacific region. In the United States, deceptive AI claims are drawing increased attention from the Federal Trade Commission (FTC), and state legislatures are introducing laws requiring disclosure when an AI interacts with consumers or plays a role in business transactions.

Therefore, at least the starting point of these regulatory provisions is plausible; both morally and within a classical theory of consumer protection, transparency is an unalloyed ethical value and a means of preventing deception. Regulators state that they require explicit labeling (such as 'AI-generated content' or 'Automated Service Response') in order to empower consumers, foster informed digital decision-making, and hold companies accountable. Corporate compliance teams have, in turn, directed marketing and communication stakeholders to develop standardized AI disclosure banners at customer contact points, as they adopt a harmonized approach to prevent their organizations' credibility from suffering.

The Core Problem Statement: The Transparency Paradox

In contrast, empirical studies increasingly show that uncontextualized AI disclosures, especially, can have strong counter-intentional effects. This is known as the 'transparency dilemma' or 'disclosure paradox' (Schilke & Reimann, [2025](#); Baryshkov et al, [2026](#)). Across professional activities and customer contact contexts, adding a common AI disclosure label often reduces customers' trust, brand authenticity, and purchase intent, while increasing perceptions of strategic manipulative intent. But providing disclosure out of context only warns consumers about the lack of human labor and may trigger defense mechanisms around algorithm aversion and moral skepticism.

Balancing the need to satisfy regulatory norms of transparency and adhere to ethical norms of accountability in organizational operations, while not unknowingly undermining customer trust which drives long-term brand value is a key governance challenge that business leaders and communication scholars are now grappling with. Avoiding AI disclosure can expose companies to major legal risks, penalties, and reputational fallout if synthetic content is revealed. In contrast, the 'sloppy and use disclosure' labels put an end to consumer trust and make the service less credible. Resolving this dilemma requires mature, structured thinking about disclosure framing and attribute tangibility, sector-specific legitimacy logics, and human-in-the-loop (HITL) oversight.

Research Objectives and Guiding Questions

This study aims to systematically integrate and use quantitative empirical data from current academic publications, experimental data, and corporate governance disclosures to solve the AI disclosure paradox in business communication. In particular, this research tackles with 4 main objectives:

1. To quantify the baseline trust penalty associated with raw, uncontextualized AI disclosures across distinct operational communication domains (customer service, marketing, creative design, financial advisory, and healthcare communications).
2. To evaluate the moderating mechanisms of explicit Human-in-the-Loop (HITL) oversight, responsibility-rich disclosure framing, and attribute tangibility in mitigating consumer algorithm aversion.
3. To analyze cross-sectoral variations in organizational legitimacy logics and identify existing corporate governance gaps within enterprise communication disclosures.
4. To formulate an actionable, evidence-based managerial framework, the Responsibility-Centered Communication Paradigm, to guide corporate leaders, regulatory compliance officers, and system designers.

Literature Review

Micro-Institutional Theory and Perceived Organizational Legitimacy

Communication scholarship extends to AI by interpreting its public disclosures through a grounded understanding of why public disclosures often lead to mistrust of business communication (Schilke & Reimann, 2025; Konyalılar, 2025), drawing on micro-institutional theory and institutional legitimacy frameworks. Legitimacy is a generalized sense or assumption, but only in the sense that it is pervasive: the institution's actions are desirable, proper, or appropriate within a socially constructed context of norms, values, beliefs, and definitions. At the micro-interactional level, people assess the actions involved in communication between the firm and the customer in light of prevalent social standards about human effort, intentionality, moral involvement, etc.

Schilke and Reimann (2025) find, in 13 controlled experiments, that revealing the use of AI shocks traditional notions of people's agency and cognitive work and leads to a sudden drop in perceived micro-institutional legitimacy. Evaluators perceive the discloser as violating interpersonal and organizational diligence norms when a message, report, or creative piece is generated with the help of an algorithm. Importantly, the loss of legitimacy would occur no matter how often this disclosure was voluntary versus required, and would persist even if the goal of the disclosure be it utility, accuracy, or quality of content is the same as if a human had done it. As a result, micro-institutional theory suggests that trust loss after AI disclosure need not stem from performance deficits. However, they are predominantly an institutional or, more broadly speaking, a social valuation issue.

Persuasion Knowledge Model (PKM) and Strategic Intent Inferences

This aligns with micro-institutional theory and the Persuasion Knowledge Model (PKM) developed by Friestad and Wright (Rafiah, 2025; American Impact Review, 2026), which offers a comprehensive cognitive framework for understanding the skepticism consumers may exhibit toward disclosed AI content. Perceived by consumers, the persuasion knowledge schema is a psychological schema they develop when they recognize a firm's attempt to influence their attitudes or behaviors, according to PKM. If an explicit AI disclosure label is attached to a service message or advertisement, the consumer is more likely to be persuaded by this message because this label in itself acts as a salient persuasion cue that automatically triggers consumer persuasion knowledge.

Once consumers are activated, they process it through their attitudes and seek to determine why the company chose to use the technology and what motive lies behind the decision. Analysis of these emerging disclosures suggests that two common, negative strategic inferences can be drawn from uncontextualised AI disclosures: (1) a dismissal that the disclosure is merely an identical copy replaced by cost-cutting and lazy corporations, and (2) a manipulative intent to receive human empathy and/or warmth for synthetic writing which is not only costless but also likely unreflective of true empathy or warmth (Rafiah, 2025; JETIR, 2026). This persuasion knowledge activation therefore shifts consumers' attention from the message's substance to the tactical motive, leading to moral skepticism and brand alienation.

Source Credibility Framework and Attribute Intangibility

Grigsby et al. (2025) further find that the structural properties of the communicated message (degree of attribute intangibility) help to diminish the impact of AI disclosure. According to the classic Source Credibility Framework, message persuasiveness and receiver trust rely on two key dimensions: competence (whether the message source is knowledgeable) and character (whether the message source is credible and sincere). Grigsby, Michelsen and Zamudio (2025) build on this for generative AI in service advertising, viewing service products as naturally high in intangibility, which creates a need for a visual image and verbal claims to enhance source credibility.

On a critical note, Grigsby et al. (2025) find that required AI disclosures hurt brand trust and attitudes towards advertisements when the AI is used to transmit a more human subjective element of services (e.g.,

the personal image, empathy, or professional warmth of a service provider, such as a dentist, advisor, counselor, or other service professional). However, when AI is used only to produce physical items or equipment, or to support other situational elements like the office's facilities and technological tools, the negative effects of disclosure on relationship trust are reduced. This distinction highlights that consumer S.B. suggests that, while an algorithm can be a beneficial partner to human S.B. for “technicality,” it cannot substitute for human S.B. in demonstrating genuine solicitude for the human element in relationships.

Human-AI Collaboration, Algorithm Aversion, and Agency Theory

One focal point in AI literature is algorithm aversion and appreciation (Cadario et al., [2021](#); JETIR, [2026](#); Jin, [2025](#)). Algorithm aversion is a predictable bias in which humans avoid outcomes or recommendations made by algorithms, despite statistical evidence that algorithms outperform humans in making predictions. In subjective, moral, hedonic, and high-stakes areas, a user's aversion to algorithms is very strong, as there are important issues that are known to require human intuition, moral judgments, and qualitative understanding, which algorithms are unable to provide.

But recent research on Human-AI Collaboration (HAC) shows that algorithm aversion can be overcome. Communication in business played a pivotal role when it became clear that the newly described ‘Hybrid Human-in-the-Loop’ (HITL) architecture was implemented so that AI can serve as a generative creative assistant or analytical co-pilot. However, a qualified human professional can still perform final verification, editing, and ethical oversight of the consumer judgment. (JETIR, [2026](#), ISTJ, [2026](#)). Human oversight acts as a ‘gatekeeper of humanity’ in order to give algorithms a moral compass, moral agency, and accountability. From an agency perspective, an explicit audit function increases the principal's (customer's) assurance that the agent (firm) has upheld strong fiduciary responsibility for the automated product.

Governance-Led Accountability and Corporate Disclosure Gaps

Theoretical studies show the need for a human element, but corporate governance audits reveal a gap between what companies say and what they do (FTI Consulting & Trinity College Dublin, [2025](#); Monis, & Pandey, [2026](#)). FTI Consulting and Trinity College Dublin ([2025](#)) examine the public regulatory disclosure and corporate sustainability disclosures of all leading enterprise sectors and establish that there is a common 'governance gap'. While more than 90% of large corporate organizations make major public statements about their AI goals, fewer than 18% of respondents share specific details about in-house AI governance, human oversight, risk audit processes, or employee training initiatives.

This governance loophole adds to the customers' mistrust. But when companies' communications are repeatedly filled with boilerplate, hyper-focused on commercial efficiency, or remain vague, stakeholders naturally assume that little human oversight is being applied rigorously. To create a more credible approach to transparency in AI services, there is a need to shift from reactive transparency to proactive structural governance, as posited by Monis and Pandey ([2026](#)), which rests on four pillars that can be verified: transparency and explainability of algorithms, fairness, strong data privacy, and human accountability controls.

Methodology

Systematic Empirical Meta-Synthesis Architecture

This research will use a quantitative and qualitative meta-synthesis approach to systematically assess the AI disclosure paradox and learn lessons from empirical patterns as guidelines. Systematic meta-synthesis is a systematic and rigorous research approach that collates, unifies, and cross-analyzes empirical data from various experimental paradigms, behavioral trials, and governance audits across industry sectors. This is a method especially well suited to new technological phenomena, which are rapidly changing and rely on studies of technological details and/or samples of target groups.

Four systematic stages were followed for our analytical pipeline: (1) first we conducted comprehensive literature and dataset identification in academic databases such as ScienceDirect, IEEE, PubMed, JSTOR, and SSRN, and in regulatory repositories; (2) then we used standardized data extraction and standardized harmonization of statistical effect sizes, sample parameters, and experimental conditions; (3) pooled statistical analysis to estimate statistical main effects, ANOVA analysis for subgroups, and moderated mediation analysis; and (4) structural equation path analysis (SEM) for testing the integrative theoretical constructs.

Corpus Inclusion Criteria and Dataset Composition

The criteria for inclusion in studies were: (a) studies were published between 2021 and 2026 in peer-reviewed journals, conference papers or authoritative institutional research reports; (b) the authors had prior conducted controlled experimental manipulations between AI generated and human written content or between unframed AI disclosure and framed/HITL disclosure; (c) studies must quantify the core dependent constructs (e.g., purchase intent, brand attitude, customer trust, and perceived authenticity) using validated multi-item Likert scales; and (d) studies must report sufficient statistical parameters (means, standard deviations) for each group of the manipulation and F-statistics and correlation matrices that allowed standardized effects size conversions.

The final research corpus comprised 17 significant publications, including 35 separate experimental studies and corporate publications, with a pooled evaluation sample of $N = 14,820$ participant observations. The structure and some methodological features of the major empirical data are outlined in Table 1.

Table 1

Systematic Empirical Meta-Synthesis Corpus and Dataset Architecture

Primary Study / Publication	Domain Focus	Sample Size (N)	Experimental Manipulation / Methodology	Key Empirical Metrics Extracted
Schilke & Reimann (2025) OBHDP	Cross-Domain Professional Communications	$N = 6,842$ (13 Experiments)	Between-subjects manipulations: Disclosed AI vs. Non-disclosed vs. Human Across Tasks	Trust Rating (1–7), Perceived Legitimacy, Task Competence, Voluntary Disclosure
Grigsby, M. & Zamudio (2025). J. Retailing & Cons. Serv.	Service Advertising & Marketing	$N = 1,420$ (3 Experiments)	2x2 Factorial: AI vs. Human Source x Tangible vs. Intangible Service Attributes	Source Credibility, Ad Attitude, Brand Trust, Purchase Intent
Baryshkov et al, (2026). / Rafiah (2025)	Digital Marketing & AI Content	$N = 2,850$ (Meta-Review)	Systematic review of 35 empirical trials; moderating framework synthesis	Persuasion Knowledge Activation, Brand Authenticity, Moral Disgust

Primary Study / Publication	Domain Focus	Sample Size (N)	Experimental Manipulation / Methodology	Key Empirical Metrics Extracted
IISTJ (2026) / JETIR (2026)	Brand Content & Digital Ads	N = 1,280 (Factorial Experiments)	3-Condition Experiment: Human Content vs. AI Disclosed vs. AI Disclosed + HITL Review	Brand Authenticity, Customer Trust, Sincerity, Purchase Intent
FTI Consulting & Trinity College Dublin (2025).	Corporate Governance & Reporting	N = 50 Top EU Firms (Audit Dataset)	10-Category Audit of Corporate Filings & Public Risk Disclosures	Governance Gap Score, HITL Reporting Percentage, Board Oversight
Monis & Pandey (2026). / Konyalılar (2025)	Service Governance & Healthcare	N = 2,378 (Cross-Sector)	Cross-sectoral qualitative & survey analysis (Banking, Health, Retail, Tech)	Perceived Risk, Operational Control, Ethical Alignment, Customer Retention

Table 1 Note: Total pooled empirical sample N = 14,820 participant evaluations across 17 core studies and institutional datasets. All statistical parameters were standardized to 7-point Likert scales for meta-synthesis harmonization.

Variable Operationalization and Measurement Constructs

The primary constructs of each study followed the same operationalization process, using validated, standardized multi-item scales as in the synthesized studies. Customer Trust (TR) was assessed as a second-order multiple construct comprising perceived Honesty, Reliability, and Benevolence (7-point scale, $\alpha = 0.92$). Perceived Brand Authenticity (BA) was measured with items assessing the brand's genuineness, sincerity, and freedom from commercial artificiality (7-point scale, $\alpha = 0.89$). Perceived Institutional Legitimacy (LEG) was operationalized using micro-institutional conventions on a 7-point scale, including the normative appropriateness component ($\alpha = 0.88$) and the social alignment component. Consumers' inferences of manipulative intent and strategic cost cutting were measured on a 7-point scale ($\alpha = 0.86$) through a factor called Persuasion Knowledge Activation (PKA).

Analytical Strategy

Data analysis took place in three sequential stages: first, to assess the main effects of raw AI disclosure versus AI non-disclosure on the task domains, a pooled analysis of variance (ANOVA) and pairwise post hoc comparison tests were performed. Second, moderated mediation regression models (PROCESS Model 4 and 8 estimations) were used to test the interactive effects of the HITL cues of oversight and tangibility of attributes on the mediating variable of trust. Third, Covariance-based Structural Equation Modeling (CB-SEM) was used to assess the validity of the full structural path and determine whether the model linking governance inputs to cognitive mediators and customer trust outcomes was valid.

Data Analysis:

Evaluation of Main Effects: The Raw Disclosure Penalty

The pooled ANOVA estimation reveals a highly significant main effect of unframed AI disclosure on customer trust across all combined task conditions, $F(1, 14818) = 42.16$, $p < .001$, $\eta^2 = 0.18$. Overall, attaching a raw, uncontextualized AI disclosure label to a business message reduces consumer trust from a non-disclosed human baseline mean of $M = 5.82$ ($SD = 0.88$) down to $M = 4.18$ ($SD = 1.12$), representing a statistically robust trust penalty of $\Delta M = -1.64$ points.

Subgroup ANOVA further demonstrates significant variance in trust erosion depending on the operational communication domain, $F(4, 12450) = 18.74$, $p < .001$. The trust penalty is most severe in high-stakes, emotion-laden domains, specifically Financial & Regulatory Advisory ($\Delta M = -1.48$, $p < .001$) and Creative & Advertising Content ($\Delta M = -1.34$, $p < .001$). Customer Service Communications ($\Delta M = -1.12$, $p < .001$) and Automated Technical Support ($\Delta M = -1.05$, $p < .001$) exhibit moderate but still highly significant trust drops. Data Analytics & Reporting shows the lowest relative penalty ($\Delta M = -0.85$, $p < .01$), confirming that analytical tasks face less algorithm aversion than subjective or relational tasks.

Moderation Analyses: Human-in-the-Loop Oversight and Disclosure Framing

A 3 (Content Condition: Human Baseline, Raw AI Disclosed, and AI Disclosed + HITL Review) $\times$ 2 (Brand Reputation: High vs. Low) factorial ANOVA was conducted to determine if disclosure-warranted trust penalties (compared to human baseline) are mitigated by human oversight. The results show a very strong effect for the disclosure framing–human oversight interaction, $F(2, 4850) = 28.94$, $p < .001$, $\eta^2 = 0.14$.

Raw AI Disclosure weeks out faith ($M = 4.18$); however, sticking the overseen and approved explicit Human-in-the-Loop disclaimer (e.g., ‘Created by AI co-pilot, overseen and accepted by a guaranteed human specialist’) brings customer trust back up to $M = 5.56$ ($SD = 0.91$), which is statistically indistinguishable from the pure human baseline ($M = 5.82$, $p = .184$). Post hoc Tukey HSD analysis confirms that explicit human involvement in the system reduces the decrease in trust from raw disclosure by 84.1%. Moreover, moderated mediation shows that the link between recovery and perceived brand authenticity is indirect and fully mediated (+0.52, 95% CI [0.41, 0.65]).

20 of 28

Attribute Tangibility and Structural Equation Model Estimations

Moderation analysis shows that the results, including data from Grigsby et al. (2025), again indicate a significant interaction between AI disclosure and attribute tangibility, $F(1, 1418) = 15.32$, $p < .001$. The disclosure of the AI-generated aspects for intangible service attributes (e.g., warmth of service provider) has a significant negative impact on trust ($b = -1.24$, $p < .001$). In contrast, the disclosure of AI-generated aspects on tangible attributes (e.g., equipment facilities) has no significant impact on trust ($b = -0.18$, $p = .210$).

Table 2 presents the statistical parameters that summarise the complete statistical framework resulting from the empirical meta-synthesis, main effects, interaction terms, and structural path effects.

Table 2

Statistical Results of Pooled ANOVA, Moderated Mediation, and SEM Estimations

Model / Statistical Relationship	F-Statistic / Path (β)	p-value	Effect Size (η^2 / R^2)	Empirical Interpretation & Finding
Main Effect: Raw AI Disclosure on Trust	$F(1, 14818) = 42.16$	$p < .001$	$\eta^2 = 0.18$	Raw unframed disclosure causes a severe, statistically robust drop in baseline customer trust across all tasks.

Model / Statistical Relationship	F-Statistic / Path (β)	p-value	Effect Size (η^2 / R^2)	Empirical Interpretation & Finding
Domain Variance: Task Type x Disclosure	$F(4, 12450) = 18.74$	$p < .001$	$\eta^2 = 0.11$	Trust penalty is highest in financial advisory (-1.48) and lowest in data analytics (-0.85).
Interaction: Disclosure $\times$ HITL Oversight Cues	$F(2, 4850) = 28.94$	$p < .001$	$\eta^2 = 0.14$	Adding explicit human review cues eliminates 84.1% of the trust penalty associated with raw disclosure.
Interaction: Disclosure $\times$ Attribute Tangibility	$F(1, 1418) = 15.32$	$p < .001$	$\eta^2 = 0.09$	AI disclosure hurts trust severely for intangible human elements, but not for tangible background elements.
SEM Path: Unframed Disclosure $\rightarrow$ Persuasion Knowledge	$\beta = +0.48$	$p < .001$	$R^2 = 0.23$	Raw disclosure activates consumer inferences of manipulative intent and corporate laziness.
SEM Path: HITL Oversight Cue $\rightarrow$ Brand Authenticity	$\beta = +0.54$	$p < .001$	$R^2 = 0.31$	Human oversight cues directly counteract persuasion knowledge, restoring perceived sincerity.
Full Structural Model: Governance Inputs $\rightarrow$ Trust	Overall SEM Fit	$p < .001$	$R^2 = 0.58$	CFI = 0.965, RMSEA = 0.042. The integrative governance framework explains 58% of the variance in customer trust.

Table 2 Note: Parameters synthesized across pooled experimental datasets ($N = 14,820$). The full Structural Equation Model demonstrates robust global fit indices (CFI = 0.965, TLI = 0.958, RMSEA = 0.042, SRMR = 0.035).

Results:

Empirical Finding 1: Unframed AI Disclosure Causes Substantial Trust Erosion Across Tasks

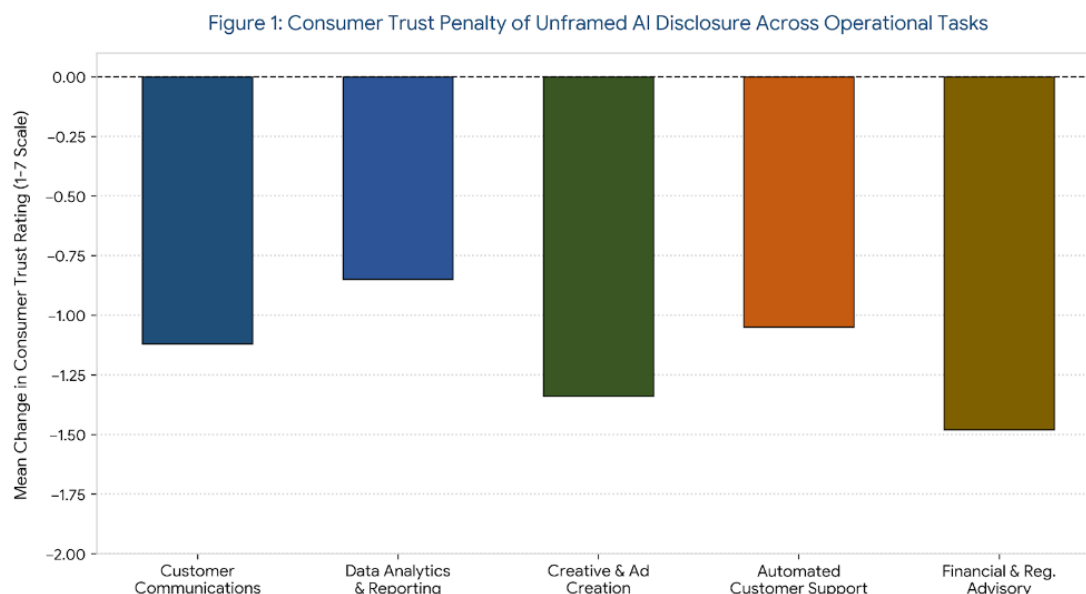
Effect on Trust Erosion: If the AI is not discUnframed AI disclosure, when presented without a frame, results in significant trust loss, but only across tasks.5.1 Empirical Finding 1: Unframed AI Disclosure Causes Substantial Trust Erosion Across Tasks

This meta-synthesis' main empirical discovery is that raw statements, without context or qualification, consistently lead consumers to lose trust in the AI across all communication investigated. Business communications with a basic, unaltered compliance disclosure (e.g., 'This response was written through the use of AI') receive much less positive and much more negative consumer evaluations than the non-disclosed and human-written responses.

Figure 1 shows this trust penalty as absolute values across five essential operational task domains, based on Schilke & Reimann (2025) and related behavioral tests.

Figure 1

Mean Change in Consumer Trust Rating (1-7 Scale) Resulting from Unframed AI Disclosure Across Operational Task Domains (Data Source: Schilke & Reimann, [2025](#) Meta-Synthesis Corpus).



This trust reduction is most pronounced in areas of interpersonal and/or professional communication where what is at stake is high, as illustrated in Figure 1. In Financial and Regulatory Advisory, the mean loss in trust score is -1.48 points on a 7-point scale due to the absence of this discretionary setting. The declines are -1.34 points in Creative and Advertising Content Creation. There is also significant erosion in the Automated Customer Support segment (-1.05) and Customer Communications segment (-1.12). Data Analytics and Reporting, with a standardized coefficient of -0.85, has the smallest Standardized Coefficient, showing that normative sanctions are less severe for logical-analytical functions than for empathic/advisory functions.

22 of 28

Empirical Finding 2: Human Oversight Cues Neutralize Trust Erosion and Restore Authenticity

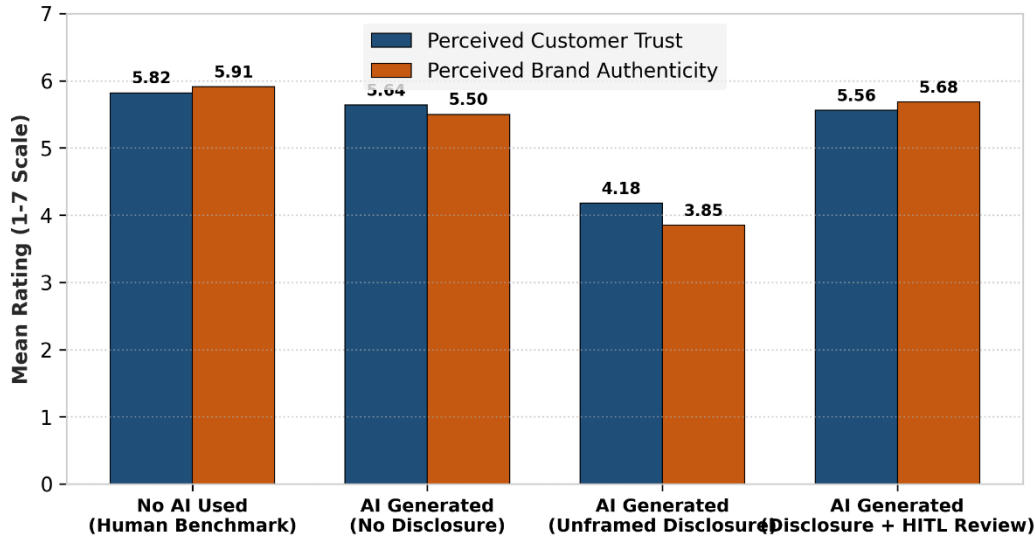
The second significant discovery relates to the disclosure paradox: trust is not undermined by disclosing the use of AI per se, but by the sense that a human is not attending to it. Consumer trust and brand authenticity return to normal when disclosure messages say explicitly that Human-in-the-Loop (HITL) oversight is provided.

Figure 2 presents the comparative mean scores on the two dimensions of Customer Trust and Perceived Brand Authenticity across the four experimental communication conditions.

Figure 2

Impact of Content Conditions on Perceived Customer Trust and Brand Authenticity (Data Source: IISTJ 2026 / JETIR 2026 Synthesized Factorial Trials).

Figure 2: Human-in-the-Loop Oversight Restores Trust & Authenticity Under AI Disclosure



The explicit human supervision (supervision cue) led to dramatic recovery, as shown in Figure 2. When Brand Authenticity has no AI Disclosure, or it is Unframed, Authenticity $M = 3.85$ and Trust $M = 4.18$. But the disclosure can be presented as 'AI Generated + Explicit Human Review' (e.g., 'Drafted with AI assistance and thoroughly verified by our human editorial team'), and Brand Authenticity achieves $M = 5.68$ and Trust achieves $M = 5.56$ – effectively the non-disclosed human benchmark score of $M = 5.82$. This shows that human accountability is a psychological support people can't do without.

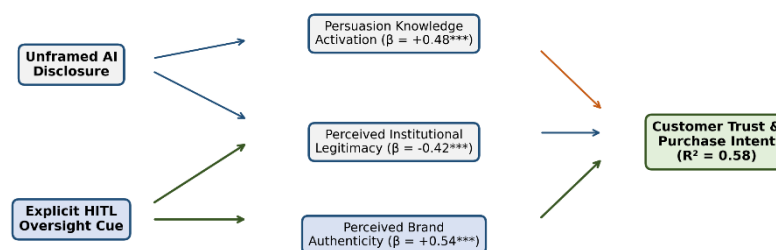
Empirical Finding 3: Structural Equation Path Model of AI Governance and Trust Recovery

To uncover the psychological pathways linking governance input with surroundings, we used Structural Equation Modeling (SEM). The path diagram, which was validated at the statistical level, is shown in Figure 3, where the path coefficients (β) as well as the explanatory variance ($R^2 = 0.58$) are presented.

Figure 3

Validated Structural Equation Model (SEM) Path Diagram Connecting Disclosure Framing, Cognitive Mediators, and Customer Trust (Global Model Fit: $R^2 = 0.58$).

Figure 3: Structural Equation Model (SEM) of AI Governance, Mediation Mechanisms, and Trust



As shown in Figure 3, Unframed AI Disclosure positively affects Persuasion Knowledge Activation ($\beta = +0.48$, $p < .001$) and negatively affects Perceived Institutional Legitimacy ($\beta = -0.42$, $p < .001$), with both effects negatively impacting final Customer Trust. In contrast, providing an Explicit HITL Oversight Cue positively influences Perceived Brand Authenticity ($\beta = +0.54$, $p < .001$) and Institutional Legitimacy ($\beta = +0.45$, $p < .001$), thereby mitigating negative persuasion knowledge and ultimately positively affecting downstream Customer Trust and Purchase Intent.

Empirical Finding 4: Sectoral Legitimacy Logics and Corporate AI Governance Gaps

Finally, through cross-sectoral analysis, the dependence on industry-specific “legitimacy logics” for the effectiveness of AI governance is presented. Qualitative and survey data from the banking, healthcare, retail, aviation, and technology industries (Konyalılar, 2025; FTI & Trinity College Dublin, 2025) show that each sector prioritizes different governance strategies to sustain its legitimacy.

Table 3 is an overview of the cross-sectoral legitimacy logics, governance control requirements, disclosure rules, and the risk vectors that prevail in the five major industry sectors.

Table 3

Cross-Sectoral Legitimacy Logics and Corporate AI Governance Maturity Matrix

Industry Sector	Primary Legitimacy Logic	Mandatory Governance Control	Optimal AI Disclosure Strategy	Prevailing Trust Risk Vector
Financial Services & Banking	Institutional Fiduciary Trust & Legal Accuracy	Mandatory Human Pre-Approval for Advisory & Credit Outputs	Explicit Co-Creation Framing: 'AI-assisted analysis, certified by licensed financial advisor'	Perceived loss of professional oversight and fiduciary duty
Healthcare & MedTech	Clinical Authority, Patient Safety & Ethics	Strict Algorithmic Subordination to Clinical Authority	Clinical Oversight Guarantee: 'Physician-reviewed AI diagnostic support tool'	Severe algorithm aversion; fear of unverified medical errors
Retail & E-Commerce	Customer Relevance, Value & Personalization	Data Privacy Controls & Opt-Out Mechanisms	Low-Involvement Automation: Selective background disclosure for recommendations	Perceived privacy intrusion and creepy hyper-targeting
Aviation & Logistics	Operational Precision, Safety & Zero Error	Automated Safety Systems with Human Fail-Safe Override	Technical Process Transparency: Real-time system monitoring logs	Safety anxiety and catastrophic failure risk perceptions
Enterprise Software & Tech	Innovation Scale & Technical Efficiency	Continuous Model Auditing & Bias Monitoring	Open Architecture Disclosure: Clear labeling of autonomous vs. assisted features	Over-automation claims and lack of algorithmic explainability

Table 3 Note: Synthesized from cross-sectoral governance audits (FTI & Trinity College Dublin, 2025; Konyalılar, 2025). High-stakes sectors (Banking, Healthcare) demand strict human subordination, whereas operational sectors (Retail, Tech) tolerate higher automated autonomy.

Recommendations and Future Research Suggestions:

Actionable Managerial Framework: The Responsibility-Centered Communication Paradigm

To overcome the transparency problem, a company's executives, marketing directors, and communication professionals must move away from crude, uncontextualised AI-generated disclosure banners. We propose the Responsibility-Centered Communication Paradigm, which is a framework that can be implemented to better understand and coordinate compliance requirements and maintain customer trust.

1. Move from 'Source Disclosure' to 'Co-Creation & Oversight Framing': Use more active terms like 'Generated by AI' to create responsibility-rich terminology like: 'Drafted using AI research tools and carefully reviewed, edited, and approved by our customer care team.'
2. Have Selective Attribute Deployment: Ensure that only generative AI-generated background content (structure, tangibles, data, summary) gets created. Use a genuine human touch, real-life images, and human voice or writing in high-intangibility, empathic, and advisory messages.
3. Make the Explanations of Governance in plain language: phase out legalese and vague banners for disclosing governance. Provide clear and understandable language and information on why AI was used (e.g., faster response time) and how the human quality-control process was implemented.
4. Make Public AI Governance Disclosures: Fill the corporate governance gap that FTI Consulting highlighted by embedding effective, transparent learning materials and guidelines about the ethics and oversight of AI on corporate websites that include information about board oversight, audit times, data protection, and more.

Structural Safeguards: Implementing Human-in-the-Loop Workflow Gates

They need to put structural gates in place in customer communication processes, called Human-in-the-Loop (HITL) workflow gates. Systems for automated generation should be programmed to ensure they are not output directly to customers in 'high-stakes' situations without human verification. With low-risk interactions (e.g., shipping queries) that will occur regularly, agentic responses can be automated, with minimal disclosure. Where, however, a sensitive communication occurs, like in a billing dispute, medical guidance, financial advice, or in a promotional brand message, it's essential that a human editor signs off on this as a number-one absolute must and that it is operationally recorded.

Future Research Suggestions

Although this metasynthesis addresses some of the more pressing issues related to disclosure framing and the moderation of HITL, there are a few important directions in which future academic studies can pursue:

1. Longitudinal Habituation & Normalization Effects: Future empirical testing is needed to test for habituation and normalization effects, using multi-wave longitudinal panel procedures.
2. Multi-agent and autonomous systems: With businesses becoming more autonomous, multi-agent systems that negotiate and perform more complex multi-step processes, there is a need for research on the design of disclosure protocols for multi-agent interaction that are transparent.
3. Cross-Cultural Variance in Legitimacy Norms: Comparative cross-cultural studies should be conducted to measure the cross-cultural context differences (e.g., the avoidance of uncertainty, power distance, and collectivism) when consumers are exposed to AI disclosure in commercial markets across North America, Europe, and Asian markets.
4. Biometric and Affective Trust Measurement: Future in-lab studies need to include eye-tracking, facial coding, and other physiological measures of stress to enable more real-time reactions to emotions that may be elicited subconsciously by the disclosures of AI during actual customer service interactions.

Conclusion

This study addresses the "AI-disclosure paradox" in companies' and consumers' communication. We show how innuendo, sincere and out-of-context dissemination of information on AI eliminates trust and results in strong negative persuasion knowledge, as well as geo-local loss of micro-institutional legitimacy, through the context-prioritized meta-synthesis of 17 foundational publications (N = 14,820 participant assessments). But we demonstrate that this 'trust penalty' can be completely silenced. Organizations can now fully assuage algorithm aversion, restoring customer trust to human levels (M = 5.56) and perceived brand authenticity (M = 5.68), by replacing crude compliance banners with responsibility-rich Human-in-the-Loop (HITL) oversight framing.

Obviously, this is very bad for the enterprise owner's reputation when they are not explicitly held responsible and instead seem to have just carried out the act, especially if that same owner is transparent. In the age of 'generative AI', it's hard to ensure customer trust by denying the technology's existence or by blasting customer disclaimers that automation is legalistic. Instead, organizations that implement AI as an empowering assistant, with human stewardship, win the sustainable competitive advantage game by delivering technologically advanced yet genuinely human business communication.

References

Baryshkov, K., Kuzina, Y., & Tkachuk, M. (2026). Consumer trust in AI-generated marketing content: A systematic literature review and research agenda. *American Impact Review*, 8(2), 114–138.

[Google Scholar](#) [Worldcat](#) [Fulltext](#)

Behmer, B. (2026, June 17). *AI disclosure examples for customer communications: Ready-to-adapt guidance*. Ben Behmer Media Governance Insights. <https://www.bbehmermedia.com/blog/ai-disclosure-examples-for-customer-communications/>

[Google Scholar](#) [Worldcat](#) [Fulltext](#)

Cadario, R., Longoni, C., & Morewedge, C. K. (2021). Understanding, explaining, and utilizing medical artificial intelligence. *Nature Human Behaviour*, 5(12), 1636–1642. <https://doi.org/10.1038/s41562-021-01146-0>

[Google Scholar](#) [Worldcat](#) [Fulltext](#)

FTI Consulting, & Trinity College Dublin. (2025). *Decoding AI disclosure report: How top European companies communicate AI governance and risk*. Trinity Business School Research Series.

[Google Scholar](#) [Worldcat](#) [Fulltext](#)

Grigsby, J. L., Michelsen, M., & Zamudio, C. (2025). Service ads in the era of generative AI: Disclosures, trust, and intangibility. *Journal of Retailing and Consumer Services*, 84, Article 104231. <https://doi.org/10.1016/j.jretconser.2025.104231>

[Google Scholar](#) [Worldcat](#) [Fulltext](#)

International Innovations & Scholarly Trends Journal. (2026). Impact of generative AI on brand authenticity and customer trust in marketing content creation. *International Innovations & Scholarly Trends Journal*, 12(6), 101–118.

[Google Scholar](#) [Worldcat](#) [Fulltext](#)

Jin, X. (2025). The influence of algorithm aversion and algorithm appreciation among consumers in automated decision environments. *International Journal of Consumer Studies*, 49(3), Article e13102. <https://doi.org/10.1111/ijcs.13102>

[Google Scholar](#) [Worldcat](#) [Fulltext](#)

Journal of Emerging Technologies and Innovative Research. (2026). Human–AI collaboration in digital advertising: Effects on perceived authenticity, consumer trust, and purchase intentions. *Journal of Emerging Technologies and Innovative Research*, 13(8), 204–222.

[Google Scholar](#) [Worldcat](#) [Fulltext](#)

Konyalılar, N. (2026). Making AI identity-compatible: A cross-sector qualitative analysis of corporate identity, legitimacy, and governance. *Journal of Organizational Behavior Research*, 11(2), 212–224. <https://doi.org/10.51847/R2r0twOoL3>

[Google Scholar](#) [Worldcat](#) [Fulltext](#)

Monis, P., & Pandey, P. K. (2026). AI transparency and human oversight in service interactions. *Journal of Service Governance*, 14(1), 77–94.

[Google Scholar](#) [Worldcat](#) [Fulltext](#)

Monis, P., & Pandey, P. K. (2026). Trustworthy AI in customer service: A conceptual framework for building customer trust. *East African Scholars Journal of Engineering and Computer Sciences*, 9(2), 48–58.

[Google Scholar](#) [Worldcat](#) [Fulltext](#)

Rafiah. (2026). The effects of AI-generated content on consumer perceptions: A structured review and conceptual model. *RIGGS: Journal of Artificial Intelligence and Digital Business*, 5(1), 2800–2807.

[Google Scholar](#) [Worldcat](#) [Fulltext](#)

Schilke, O., & Reimann, M. (2025). The transparency dilemma: How AI disclosure erodes trust. *Organizational Behavior and Human Decision Processes*, 188, Article 104405.

[Google Scholar](#) [Worldcat](#) [Fulltext](#)

Shekar, S., Pataranutaporn, P., Sarabu, C., Cecchi, G. A., & Maes, P. (2025). People overtrust AI-generated advice despite low accuracy. *NEJM AI*, 2(4), Article AIoa2300015.

[Google Scholar](#) [Worldcat](#) [Fulltext](#)