Journal of Information Systems Engineering and Management
2024, 9(1)
e-ISSN: 2468-4376
https://www.jisem-journal.com/ **Research Article**

The AI Disclosure Paradox in Marketing: A Persuasion Knowledge Authenticity Framework

Dr Minal Shah[1], Dr Shrawan Pandey[2] Shraddha Gupta[3]

1Assistant Professor, Bhilai Institute of Technology, Durg

28minalshah@gmail.com

ORCID: 0000-0003-3320-0917

[2]Assistant Professor, Bhilai Institute of Technology, Durg

Shrawan.pandey@bitdurg.ac.in

[3]Assistant Professor, Bhilai Institute of Technology, Durg

Shraddhagupta18@gmail.com

ORCID: 0000-0002-2125-7900

ARTICLE INFO

Received: 03 Feb 2024

Revised: 17 March 2024

Accepted: 27 March 2024

ABSTRACT

Regulators and industry codes increasingly require brands to disclose when consumer-facing content or interactions involve artificial intelligence, yet a growing body of marketing evidence shows that disclosure itself carries a behavioral cost: it can activate consumer skepticism, reduce perceived authenticity, and depress trust and purchase intention. This conceptual paper integrates the persuasion knowledge model (Friestad & Wright, 1994), theories of consumption authenticity (Beverland & Farrelly, 2010), algorithm aversion research (Dietvorst, Simmons, & Massey, 2015), and recent empirical evidence on AI chatbot and AI-generated advertising disclosure (Luo, Tong, Fang, & Qu, 2019; Longoni & Cian, 2022; Kučinskas & Survilaitė,) to develop a Persuasion Knowledge–Authenticity Framework explaining when AI disclosure helps versus harms marketing outcomes. The framework specifies five interacting constructs — disclosure salience, task-type congruence, perceived authenticity threat, persuasion knowledge activation, and track-record sensitivity — and advances six propositions describing how these constructs jointly determine the size and direction of disclosure effects on consumer trust and purchase intention. The paper contributes to marketing theory by reframing AI transparency as a contingent design problem rather than a uniformly positive or negative practice, and it offers marketers a diagnostic vocabulary for calibrating disclosure strategy across utilitarian and hedonic contexts, brand familiarity levels, and consumer AI-literacy segments. To illustrate the framework's plausibility ahead of purpose-built testing, the paper reanalyzes published results from a nationally representative U.S. survey (Pew Research Center's American Trends Panel, June, N = 5,023) covering seven AI-disclosure vignettes, finding a pattern of disclosure penalties broadly consistent with the framework's emphasis on persuasion-relevant and consequential contexts. Boundary conditions, managerial implications, and a future research agenda are discussed.

Index Terms: artificial intelligence disclosure; persuasion knowledge; brand authenticity; algorithm aversion; consumer trust; AI-generated advertising

Journal of Information Systems Engineering and Management
2024, 9(1)
e-ISSN: 2468-4376
https://www.jisem-journal.com/ **Research Article**

1. Introduction

Artificial intelligence is now embedded across the marketing communications stack: chatbots conduct sales conversations, generative systems draft advertising copy and imagery, and recommendation engines personalize offers at scale. As this embedding has deepened, so has pressure — from regulators, platform policies, and professional codes of conduct — for brands to disclose when a consumer-facing message, interaction, or recommendation involves AI rather than a human. Disclosure is typically framed in policy discussions as an unambiguous good: a matter of transparency and consumer protection. Marketing research on the actual behavioral consequences of disclosure tells a more complicated story.

Field experimental evidence shows that undisclosed AI chatbots can match the sales performance of experienced human agents, yet disclosing the chatbot's non-human identity before a sales conversation reduces purchase rates by more than three-quarters (Luo, Tong, Fang, & Qu, 2019). Complementary laboratory evidence shows that consumers do not react uniformly to AI involvement: they prefer AI-generated recommendations in utilitarian domains but resist them in hedonic domains, a pattern termed the “word-of-machine” effect (Longoni & Cian, 2022). Most recently, experimental work on AI-generated advertising finds that disclosing AI involvement in ad creation reduces perceived authenticity, and that this authenticity loss — rather than any direct dislike of AI — is the primary mechanism through which disclosure damages brand image, trust, and purchase intention, with the effect more pronounced for well-established brands and for consumers who know more about AI (Kučinskas & Survilaitė,).

These findings sit awkwardly alongside two decades of research on how consumers cope with persuasion attempts. The persuasion knowledge model holds that consumers possess an evolving schema for recognizing when, how, and why they are being marketed to, and that activating this schema changes how they interpret a communication and its source (Friestad & Wright, 1994). Disclosure of AI involvement is, in this light, not a neutral transparency cue but a trigger for persuasion knowledge: it invites the consumer to ask what the brand is trying to achieve by using AI, and to discount the message accordingly. Research on consumption authenticity further shows that consumers actively seek and weigh cues of human origin, craft, and intention when judging whether an object or experience is authentic (Beverland & Farrelly, 2010) — precisely the cues that AI disclosure calls into question. And research on algorithm aversion shows that consumer trust in algorithmic agents, once granted, is unusually fragile: people lose confidence in algorithms faster than in humans after observing the same mistake (Dietvorst, Simmons, & Massey, 2015).

Taken together, these literatures point to what this paper terms the AI disclosure paradox: the same transparency that regulators and ethicists prescribe as a remedy for AI-related consumer harm can itself function as a persuasion-knowledge trigger that damages the perceived authenticity, trust, and commercial performance of marketing communications — but only under identifiable conditions. No existing framework specifies these conditions in an integrated way. This paper addresses that gap by asking two questions. First, through what psychological mechanisms does AI disclosure affect consumer trust and purchase intention in marketing contexts? Second, what contextual and design conditions determine whether disclosure effects are large, small, or reversed? To answer these questions, the paper develops a Persuasion Knowledge–Authenticity Framework comprising five constructs — disclosure salience, task-type congruence, perceived authenticity threat, persuasion knowledge activation, and track-record sensitivity — and six propositions. Section 2 reviews the relevant theoretical foundations. Section 3 problematizes the disclosure paradox. Section 4 presents the framework and propositions. Section 5 reanalyzes a nationally representative survey to illustrate the framework's plausibility. Section 6 discusses theoretical and managerial implications and boundary conditions, Section 7 sets out a future research agenda, and Section 8 concludes.

Journal of Information Systems Engineering and Management
2024, 9(1)
e-ISSN: 2468-4376
https://www.jisem-journal.com/ **Research Article**

2. Theoretical Background

2.1 The Persuasion Knowledge Model

Friestad and Wright (1994) propose that consumers develop, over time, a body of persuasion knowledge — beliefs about marketers' goals, tactics, and the appropriateness of particular influence attempts — that they draw upon whenever they recognize a communication as an attempt to persuade them. Once persuasion knowledge is activated, consumers shift from evaluating a message on its face content to evaluating it as a strategic act, attending to the presumed motives of the source and often discounting the message accordingly. Disclosure of AI involvement is a natural activator of this schema: it signals that a non-human, and potentially unaccountable, agent is behind a communication, inviting consumers to ask why the brand chose to use AI and what that choice implies about the sincerity of the message.

2.2 Consumption and Brand Authenticity

Beverland and Farrelly (2010) show that consumers actively seek out cues — of origin, craft, human intention, and connection to place or tradition — that allow them to experience objects and brand communications as authentic, and that this pursuit of authenticity serves identity goals of control, connection, and virtue. Authenticity judgments are therefore not passive reactions to objective facts about a product's origin but active interpretive work performed by consumers using available cues. AI disclosure directly intervenes in this interpretive process by supplying a cue — non-human authorship — that is difficult to reconcile with the human-sourced craft and intentionality that authenticity judgments typically rely upon.

2.3 Algorithm Aversion and the Fragility of Algorithmic Trust

Dietvorst, Simmons, and Massey (2015) document algorithm aversion: even when algorithms outperform human forecasters on average, people lose confidence in algorithms more quickly than in humans after observing an identical error, and often abandon an algorithm after a single mistake despite continuing to tolerate repeated human error. This asymmetry suggests that consumer trust in AI-driven marketing tools, once established, may be unusually vulnerable to visible failures — a dynamic with direct implications for how disclosure interacts with a brand's track record of AI performance.

2.4 The Word-of-Machine Effect and Task-Type Contingency

Longoni and Cian (2022) find that consumer responses to AI recommenders depend systematically on whether the consumption context is utilitarian or hedonic. Consumers tend to view AI as more competent than humans for utilitarian judgments, where performance and functional attributes matter most, but as less competent for hedonic judgments, where taste, emotion, and experiential quality matter most — a pattern they term the word-of-machine effect. This suggests that the reputational and behavioral costs of AI disclosure are unlikely to be uniform across marketing contexts, but should instead vary with the utilitarian-hedonic character of the product or communication at issue.

2.5 Empirical Signals from Chatbot and Advertising Disclosure Research

Two recent empirical studies anchor the practical stakes of this paper's framework. Luo, Tong, Fang, and Qu (2019) show, using a large-scale field experiment with outbound sales chatbots, that disclosing a chatbot's non-human identity cuts purchase rates by roughly 80 percent relative to non-disclosure, because customers treat disclosed bots as less knowledgeable and less empathetic conversational partners. Kučinskas and Survilaitė (), using experimental studies of AI-generated advertising, find that disclosing AI involvement in ad creation reduces perceived authenticity, which in turn mediates reductions in purchase intention, brand image, and brand trust, with the effect stronger for familiar brands and for consumers with greater AI literacy. Together, these studies provide direct evidence that disclosure effects operate through authenticity and persuasion-related mechanisms, and that they are

Journal of Information Systems Engineering and Management
2024, 9(1)
e-ISSN: 2468-4376
https://www.jisem-journal.com/ **Research Article**

moderated by brand- and consumer-level factors — evidence this paper's framework organizes into a coherent set of constructs and propositions.

3. Problematization: The AI Disclosure Paradox

The tension outlined above can be stated precisely. Regulatory and ethical guidance increasingly treats AI disclosure as an unqualified good: consumers are assumed to be better off, and better able to make informed decisions, when they know they are interacting with or receiving content from an AI system. Yet the evidence reviewed in Section 2 shows that disclosure, far from being informationally neutral, functions as a psychological trigger with measurable commercial consequences. It activates persuasion knowledge (Friestad & Wright, 1994), undermines the human-origin cues on which authenticity judgments depend (Beverland & Farrelly, 2010; Kučinskas & Survilaitė,), and interacts with a brand's track record of algorithmic performance in ways that can produce outsized trust losses following even minor errors (Dietvorst et al., 2015).

This is a paradox rather than a simple cost-benefit trade-off because the very features that make disclosure ethically desirable — its capacity to make AI involvement salient and comprehensible to consumers — are the same features that activate the psychological mechanisms responsible for its commercial cost. A brand cannot achieve meaningful transparency without some degree of disclosure salience, yet disclosure salience is precisely what drives persuasion knowledge activation and perceived authenticity threat in the framework developed below. The paradox is not resolved by choosing more or less disclosure in the abstract, but by recognizing that its effects are contingent on task type, brand familiarity, consumer AI literacy, and the presence or absence of complementary design cues — the contingencies that the next section formalizes.

4. A Persuasion Knowledge–Authenticity Framework for AI Disclosure

The Persuasion Knowledge–Authenticity Framework proposes that the commercial consequences of AI disclosure in marketing are governed by five interacting constructs: disclosure salience, task-type congruence, perceived authenticity threat, persuasion knowledge activation, and track-record sensitivity. Table 1 defines each construct, its theoretical anchor, and an illustrative marketing practice through which it is enacted.

4.1 Core Constructs

Construct	Definition	Theoretical Anchor	Illustrative Marketing Practice
Disclosure Salience (DS)	The prominence and explicitness with which AI involvement in a marketing communication or interaction is revealed to the consumer.	Luo, Tong, Fang, & Qu (2019)	A chatbot's opening line identifying itself as AI versus a discreet disclosure buried in terms of service.
Task-Type Congruence (TTC)	The degree to which the consumption context is utilitarian (functional, performance-based) versus hedonic (experiential, emotion-based).	Longoni & Cian (2022)	AI-generated product-spec comparisons (utilitarian) versus AI-generated lifestyle or brand-

Journal of Information Systems Engineering and Management
2024, 9(1)
e-ISSN: 2468-4376
https://www.jisem-journal.com/ **Research Article**

			story content (hedonic).
Perceived Authenticity Threat (PAT)	The extent to which disclosed AI involvement is perceived to undermine the authentic, human-sourced quality of a brand message or offering.	Beverland & Farrelly (2010); Kučinskas & Survilaitė ()	Consumer reactions to learning that a brand's "handcrafted" storytelling was AI-drafted.
Persuasion Knowledge Activation (PKA)	The degree to which disclosure activates consumers' schema-based coping tactics for recognizing and resisting persuasion attempts.	Friestad & Wright (1994)	A consumer discounting an AI-personalized offer as "just an algorithm trying to sell to me."
Track-Record Sensitivity (TRS)	The asymmetric sensitivity of consumer trust in an algorithmic agent to observed errors relative to equivalent human errors.	Dietvorst, Simmons, & Massey (2015)	A single mis-targeted AI recommendation prompting disproportionate loss of confidence in a brand's AI-driven personalization.

Table 1. Core constructs of the Persuasion Knowledge–Authenticity Framework.

These constructs interact rather than operating in isolation. Disclosure salience is the proximate trigger for persuasion knowledge activation, but its downstream effect on trust and purchase intention is channeled through perceived authenticity threat and is conditioned by task-type congruence: the same disclosure that barely registers in a utilitarian context may substantially damage authenticity perceptions in a hedonic one. Track-record sensitivity operates as a dynamic, time-dependent amplifier, meaning the same disclosed AI system can move from being trusted to being abandoned following a single salient error, independent of its long-run accuracy.

4.2 Propositions

Building on this construct structure and the theoretical foundations in Section 2, the framework advances six propositions linking specific disclosure conditions to consumer trust and purchase-intention outcomes. These propositions are summarized in Table 2 and elaborated below.

Proposition	**Statement**
P1	Higher disclosure salience increases persuasion knowledge activation, which in turn reduces perceived authenticity and downstream trust and purchase intention (Friestad & Wright, 1994; Luo, Tong, Fang, & Qu, 2019; Kučinskas & Survilaitė,).
P2	The negative effect of AI disclosure on perceived authenticity and trust is stronger in hedonic consumption contexts than in utilitarian contexts, because consumers attribute

	human warmth and taste, rather than computational competence, to hedonic judgments (Longoni & Cian, 2022).
P3	The negative disclosure–trust relationship is attenuated by a positive track record of algorithmic performance, but a single subsequent algorithmic error produces a disproportionate loss of consumer trust relative to an equivalent human error (Dietvorst, Simmons, & Massey, 2015).
P4	Brand familiarity moderates the disclosure–authenticity relationship such that established, well-known brands suffer larger authenticity and trust losses upon AI disclosure than unfamiliar or new brands, because consumers hold established brands to higher expectations of human craftsmanship (Kučinskas & Survilaitė,).
P5	Consumers' AI literacy moderates disclosure effects such that more AI-literate consumers react more negatively to AI disclosure, because persuasion knowledge activation depends on the sophistication of consumers' schemas about how AI systems operate (Friestad & Wright, 1994; Kučinskas & Survilaitė,).
P6	Pairing AI disclosure with human-in-the-loop cues — such as named human reviewers, co-creation framing, or visible editorial oversight — reduces the negative effect of disclosure on perceived authenticity and purchase intention relative to disclosing AI involvement alone, because such cues restore the attributions of human agency and intentionality central to authenticity judgments (Beverland & Farrelly, 2010).

Table 2. Propositions of the Persuasion Knowledge–Authenticity Framework.

Proposition 1 formalizes the core disclosure paradox by specifying persuasion knowledge activation and perceived authenticity threat as sequential mediators between disclosure salience and behavioral outcomes, integrating the mechanism implied by Friestad and Wright (1994) with the empirical disclosure effects documented by Luo et al. (2019) and Kučinskas and Survilaitė (). Proposition 2 imports the word-of-machine effect (Longoni & Cian, 2022) as a boundary condition, predicting that disclosure costs concentrate in hedonic rather than utilitarian marketing contexts. Proposition 3 extends algorithm aversion research (Dietvorst et al., 2015) to a disclosure setting, predicting an asymmetric, error-triggered trust dynamic rather than a stable disclosure penalty. Propositions 4 and 5 operationalize the moderators identified empirically by Kučinskas and Survilaitė () — brand familiarity and consumer AI literacy — within the framework's constructs. Proposition 6 translates the authenticity mechanism (Beverland & Farrelly, 2010) into an actionable design lever, predicting that human-in-the-loop cues can offset, without eliminating, the disclosure penalty.

5. An Illustrative Secondary-Data Analysis

The propositions developed in Section 4 have not, to date, been tested in a purpose-built marketing experiment. To provide a preliminary, real-data illustration of the framework's plausibility ahead of such experiments, this section reanalyzes published results from a nationally representative U.S. survey that, although not designed to test this paper's framework, asked respondents to react to several disclosure vignettes closely resembling the marketing and communication contexts the framework addresses.

5.1 Data Source

The data come from the Pew Research Center's American Trends Panel Wave 173, fielded June 9–15, , among a nationally representative sample of 5,023 U.S. adults (margin of error ±1.6 percentage points;

Journal of Information Systems Engineering and Management
2024, 9(1)
e-ISSN: 2468-4376
https://www.jisem-journal.com/ **Research Article**

Kennedy, Yam, Kikuchi, Pula, & Fuentes,). Respondents were randomly assigned to one of two survey forms. Form 1 (n = 2,505) presented vignettes about a painting, a customer-service chatbot interaction, and a political candidate's speech; Form 2 (n = 2,518) presented vignettes about a song, a medical treatment recommendation, a news article, and a loan decision. In each vignette, respondents first imagined forming a favorable impression of an outcome (liking a painting or song, receiving helpful customer service, learning from a news article, and so on) and were then asked how learning that AI had been involved would change their reaction: more favorably, less favorably, or not at all. This paper reanalyzes the published topline percentages for these seven vignettes; it did not collect new data and had no role in the survey's design.

5.2 Pattern of Disclosure Reactions Across Contexts

For each vignette, Table 3 reports the percentage of respondents who reacted negatively upon learning of AI involvement, the percentage whose reaction did not change, the percentage who reacted positively, and a net disclosure penalty computed as the negative percentage minus the positive percentage. The seven contexts are ordered from the largest to the smallest net penalty.

Vignette (Form; n)	% Reacted Negatively	% No Change	% Reacted Positively	Net Disclosure Penalty
Political speech written by AI (Form 1; n = 2,505)	71	25	3	68
Loan decided by AI (Form 2; n = 2,518)	57	35	8	49
News article written by AI (Form 2; n = 2,518)	56	36	7	49
Painting made by AI (Form 1; n = 2,505)	49	48	3	46
Customer-service reply from an AI chatbot (Form 1; n = 2,505)	41	53	6	35
Song created by AI (Form 2; n = 2,518)	38	58	3	35
Medical treatment recommended by AI, delivered by a human doctor (Form 2; n = 2,518)	45	41	13	32

Table 3. Reactions to learning of AI involvement across seven vignettes, Pew Research Center American Trends Panel Wave 173 (June, N = 5,023). Percentages and sample sizes are drawn from the publicly released topline; net disclosure penalty is calculated by the present author as % negative minus % positive.

Three patterns in Table 3 speak directly to the framework. First, the single largest disclosure penalty by a wide margin occurs for the political speech vignette (net penalty of 68 points, with 71% of respondents

saying the disclosure would make them feel less favorable toward the candidate). Because a candidate's speech is a paradigm case of an explicit persuasion attempt, this pattern is consistent with Proposition 1's claim that disclosure operates substantially through persuasion knowledge activation (Friestad & Wright, 1994): the same disclosure that produces a comparatively modest penalty in a customer-service interaction produces a far larger penalty when the communication's persuasive intent is unmistakable.

Second, the customer-service chatbot vignette shows a clear net penalty (35 points, with 41% reacting worse and only 6% reacting better), directionally corroborating, in a general-population, nationally representative sample, the basic disclosure penalty that Luo, Tong, Fang, and Qu (2019) documented experimentally in a commercial sales-chatbot setting. This convergence across a controlled field experiment and an independent household survey strengthens confidence that the disclosure penalty central to this paper's framework is not an artifact of one study design.

Third, the results complicate a simple utilitarian-versus-hedonic reading of the word-of-machine effect (Longoni & Cian, 2022): the two nominally utilitarian, high-stakes vignettes — a loan decision and a news article — show net penalties (49 and 49) as large as or larger than the hedonic painting vignette (46), and larger than the hedonic song vignette (35). One plausible reading, consistent with Proposition 3's emphasis on the stakes and finality of algorithmic errors, is that penalties are driven not only by whether a domain is utilitarian or hedonic but by how consequential and difficult to contest the AI-attributed outcome is to the individual — a loan denial or a misleading news article being more consequential than a disliked painting. A second, more tentative pattern concerns Proposition 6: the medical-treatment vignette, in which a human doctor still delivers the AI-informed recommendation, shows the smallest net penalty of the seven contexts (32), broadly consistent with the idea that a visible human-in-the-loop presence can attenuate, without eliminating, the disclosure penalty (Beverland & Farrelly, 2010). Because the medical vignette also differs from the others in domain and stakes, this reading should be treated as suggestive rather than as a controlled test of Proposition 6.

5.3 Limitations of This Illustration

This analysis is offered as an illustrative complement to the conceptual framework, not as a confirmatory empirical test of it, for several reasons. The underlying survey was designed by Pew Research Center to track general public attitudes toward AI, not to operationalize this paper's constructs; disclosure salience, task-type congruence, and human-in-the-loop cues were not experimentally manipulated as independent factors, and the classification of vignettes as more or less utilitarian, hedonic, or persuasion-relevant is a post-hoc interpretation applied by the present author rather than a validated a priori design. The outcome measures are single-item hypothetical reactions rather than incentivized behavior or validated multi-item scales of trust, authenticity, or purchase intention, and the sample is the general U.S. population rather than consumers evaluating a specific brand. The seven vignettes also differ from one another along several dimensions simultaneously — domain, stakes, and the presence of a human intermediary — making it impossible to isolate any single construct's effect from this data alone. The patterns reported here should therefore be read as consistent with, rather than as evidence establishing, Propositions 1, 3, and 6, and they motivate rather than replace the purpose-built experimental and survey studies proposed in Section 7.

6. Discussion

6.1 Theoretical Contributions

This paper makes three contributions to marketing theory. First, it integrates previously separate streams of research — persuasion knowledge (Friestad & Wright, 1994), consumption authenticity (Beverland & Farrelly, 2010), algorithm aversion (Dietvorst et al., 2015), and task-contingent AI evaluation (Longoni & Cian, 2022) — into a single framework specifically addressing AI disclosure, an organizational and regulatory practice that has outpaced its theoretical treatment in marketing. Second, it reframes AI transparency from a uniformly beneficial compliance behavior into a contingent design

variable whose commercial consequences depend on task type, brand familiarity, consumer AI literacy, and the presence of complementary cues, echoing recent experimental evidence that disclosure effects are neither universal nor uniform (Kučinskas & Survilaitė,). Third, it extends algorithm aversion theory, developed primarily in forecasting and decision-support contexts (Dietvorst et al., 2015), to the disclosure moment itself, proposing that the same asymmetric fragility of algorithmic trust documented after forecasting errors should apply to consumer trust following disclosed AI involvement in brand communications.

6.2 Managerial Implications

For marketing managers, the framework implies that disclosure decisions should not be treated as a binary compliance checkbox but as a design problem with several levers. Marketers can calibrate disclosure salience and framing to context, recognizing that hedonic and identity-relevant communications are more vulnerable to authenticity loss than utilitarian ones (Longoni & Cian, 2022). Brands with strong AI track records can consider building visible performance histories before increasing disclosure salience, while remaining aware that a single high-profile error can trigger disproportionate trust loss regardless of past accuracy (Dietvorst et al., 2015). Established brands, which the evidence suggests suffer larger authenticity penalties from disclosure than lesser-known brands (Kučinskas & Survilaitė,), may need to invest more deliberately in human-in-the-loop cues — named human reviewers, visible editorial processes, or co-creation narratives — to preserve authenticity perceptions (Beverland & Farrelly, 2010). Finally, segmenting disclosure communication by audience AI literacy, rather than using a one-size-fits-all disclosure statement, may help manage the persuasion-knowledge costs identified in Proposition 5.

6.3 Boundary Conditions and Limitations

As a conceptual paper, the framework has not been empirically tested as an integrated whole, and its propositions should be read as a research agenda rather than confirmed findings. The framework draws its empirical grounding primarily from chatbot sales interactions, AI-generated advertising, and algorithmic recommendation contexts; its applicability to other AI-mediated marketing practices, such as AI-driven pricing or fully autonomous shopping agents, remains to be established. Regulatory context is also likely to be a boundary condition: in jurisdictions where AI disclosure is legally mandated with prescribed wording or placement, marketers have less latitude to manage disclosure salience, which may alter the strength of the proposed relationships. Cultural variation in attitudes toward AI and in the cultural salience of authenticity as a consumption value may further moderate the framework's propositions and warrants explicit cross-cultural testing.

7. Future Research Agenda

Several research directions follow from the framework. Experimental studies could manipulate disclosure salience, task type, and the presence of human-in-the-loop cues factorially to test the mediating role of persuasion knowledge activation and perceived authenticity threat proposed in Propositions 1, 2, and 6, extending the experimental paradigms used by Luo et al. (2019) and Kučinskas and Survilaitė . Longitudinal field studies could track consumer trust in disclosed AI systems before and after a salient error to test the asymmetric track-record dynamic proposed in Proposition 3, building on the laboratory evidence for algorithm aversion (Dietvorst et al., 2015). Survey-based scale development could operationalize disclosure salience and track-record sensitivity as measurable constructs, enabling large-sample tests of Propositions 4 and 5 across brand-familiarity and AI-literacy segments. Finally, comparative research across regulatory regimes and cultural contexts would help establish the boundary conditions noted in Section 6.3, and future work might extend the framework to business-to-business marketing contexts, where persuasion knowledge and authenticity dynamics may operate differently than in the business-to-consumer settings that dominate the current evidence base.

8. Conclusion

AI disclosure sits at the intersection of regulatory expectation and commercial risk: it is widely prescribed as a transparency safeguard, yet the evidence reviewed in this paper shows it can activate persuasion knowledge, threaten perceived authenticity, and depress consumer trust and purchase intention. This paper has argued that these effects are neither uniform nor fixed, but are shaped by disclosure salience, task-type congruence, brand familiarity, consumer AI literacy, and the presence of human-in-the-loop cues, and has organized these contingencies into a Persuasion Knowledge–Authenticity Framework comprising five constructs and six propositions. By treating AI disclosure as a contingent marketing design problem, grounded in established theories of persuasion knowledge, consumption authenticity, and algorithm aversion, the framework offers both a research agenda for marketing scholars and a diagnostic vocabulary for practitioners seeking to meet transparency expectations without unnecessarily undermining the trust and authenticity on which brand relationships depend.

References

[1] Beverland, M. B., & Farrelly, F. J. (2010). The quest for authenticity in consumption: Consumers' purposive choice of authentic cues to shape experienced outcomes. Journal of Consumer Research, 36(5), 838–856. https://doi.org/10.1086/615047

[2] 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

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

[4] Longoni, C., & Cian, L. (2022). Artificial intelligence in utilitarian vs. hedonic contexts: The "word-of-machine" effect. Journal of Marketing, 86(1), 91–108. https://doi.org/10.1177/0022242920957347

[5] 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