Journal of
Theoretical and Applied Electronic Commerce Research

MDPI

Article

When AI Gets It Wrong: Hallucinations and Trust Recalibration in E-Commerce Using a Sequential Mixed-Methods Approach

Sayyed Khawar Abbas [1,*], Hafiz Muhammad Junaid [2] and Aseel Smerat [3]

[1] Department of Information Systems, Institute of Data Analytics and Information Systems, Corvinus University of Budapest, 8 Fovam Ter, 1093 Budapest, Hungary
[2] Essex Business School (EBS), University of Essex, Wivenhoe Park, Colchester CO4 3SQ, UK; junaidabidy@gmail.com
[3] Hourani Center for Applied Scientific Research, Al-Ahliyya Amman University, Amman 19328, Jordan; smerat.2020@gmail.com
* Correspondence: sayyedkhawar.abbas@uni-corvinus.hu

Abstract

Generative AI shopping assistants are becoming a primary touchpoint for online consumers, yet they often produce convincing but inaccurate content, a phenomenon known as AI hallucination that poses an under-studied risk to consumer trust in e-commerce. This study explains that phenomenon by building and testing a moderated mediation model grounded in Expectation Violation Theory, Epistemic Vigilance Theory, and Algorithmic Trust Repair Theory. A sequential, exploratory mixed-methods design was used: a qualitative phase identified the dimensions and configurational pathways of consumer trust withdrawal using the Gioia methodology and fuzzy-set Qualitative Comparative Analysis, and a subsequent large-scale quantitative phase tested and refined the resulting model across a multi-country European sample using partial least squares structural equation modeling and Necessary Condition Analysis. The results show that exposure to hallucinations triggers expectation violation, activating epistemic vigilance and reducing perceived AI competence; this sequence drives trust recalibration, reflected in lower continued-use and purchase intentions and greater negative word-of-mouth. AI literacy, prior trust, and transparency cues significantly moderate these relationships, and structural trust repair mechanisms, namely retrieval-augmented generation and uncertainty disclosure, prove more effective than purely communicative repair strategies. Theoretically, this study advances a dynamic account of trust recalibration in AI-mediated commerce; practically, it offers concrete guidance for platform design and regulatory policy under the EU AI Act.

Keywords: AI hallucination; consumer trust; expectation violation; epistemic vigilance; trust recalibration; fuzzy-set qualitative comparative analysis; mixed methods; e-commerce; online purchase intention; EU AI Act

Academic Editor: Hyunchul Ahn

Received: 9 June 2026
Revised: 11 August 2026
Accepted: 14 August 2026
Published: 17 August 2026

1. Introduction

The last two years have seen the swift integration of generative AI into the online retail landscape [1], with consumers increasingly engaging with AI-powered chatbots [2], alongside a rise in personalized agents and AI-driven searching. Industry analyses indicate that a substantial share of generative AI initiatives fail to deliver their expected value, with associated business costs continuing to grow [3], and that e-commerce is particularly impacted by this phenomenon, with promotional fabrication (e.g., invented discounts, endorsements, or limited-time offers), fabrication of specifications (e.g., incorrect battery

life, material, or compatibility claims), and phantom products (listings for items that do not exist or are permanently out of stock) [4,5]. The problem is exacerbated by the growing prevalence of AI-generated content on retail platforms, where consumers are regularly expected to assess the accuracy of synthetic product descriptions, reviews, and recommendations [6,7]. In this context, it is essential to understand how consumers identify, interpret, and react to AI hallucinations, which has become a focal point for marketing, information systems, and consumer behavior studies.

A previous study on algorithm aversion found that once an algorithm is shown to be wrong, it is hard to regain trust in it [8–10]. The literature however is still very much tied to numeric forecasting scenarios and does not consider generative hallucinations, whose defining characteristic is not random error but plausibly constructed fabrication. More recent work has now started to fill this gap. Kahr et al. [11] showed that the time order of algorithmic error affects trust recovery trajectory. For cause–limits explanations to work well in repairing competence-based trust, they must be more effective than generic apologies, as was demonstrated by Nizette et al. [12]. Without corrective signals, appropriate reliance on AI requires calibrated trust [13]. More specifically, in the field of JTAER, Liu et al. [7] developed a dual-pathway trust–risk model of AI-generated content, Wang et al. [14] discussed dual-route persuasion in relation to the credibility of AI avatars, and Yang and Park [15] used psychological ownership as a moderating variable for AI recommendation credibility. Tükel and Aydın [16] reported on the impact of chatbot activities on online purchase intention mediated by customer satisfaction and Gu et al. [6] examined the joint effects of perceived eeriness and perceived intelligence on consumer acceptance of AI-generated advertisements. Nevertheless, there are three theoretical and empirical gaps.

First, current models of trust recalibration only consider trust as a unidimensional shift, without taking into account the unique cognitive route that is followed to detect or process hallucinations, namely the activation of epistemic vigilance [17,18]. Second, the effects of these are moderated by pre-established consumer personality traits, such as AI literacy and trust, whose role has been poorly investigated in e-commerce scenarios. Third, the qualitative phenomenology of the hallucination encounter (i.e., how consumers make sense of, narrate and repair trust) has been relatively unexamined in quantitative-only designs and is infrequently examined as a theory-building qualitative phenomenon, along with the use of theory-building qualitative traditions and configurational inference.

The present study is motivated by these three gaps: it aims to link the phenomenology of the hallucination encounter to a testable, moderated mediation model of trust recalibration, and to pair configurational inference with variance-based modeling so that both the qualitative texture and the quantitative magnitude of the phenomenon can be captured.

To this end, this study adopts a sequential exploratory mixed-methods approach [19,20]. The qualitative strand used a combination of the Gioia methodology [21] and fuzzy-set Qualitative Comparative Analysis [22] on twenty-five semi-structured interviews across fifteen European countries, reflecting the inductive logic of theory development. The insights gained in this phase, including construct refinements, vignette stimuli, and plausible moderating boundary conditions, were then translated directly into the design of the subsequent quantitative strand, which combined PLS-SEM and Necessary Condition Analysis on a sample of 590 European consumers. Each technique was selected for a distinct inferential advantage: the Gioia methodology surfaces how consumers narrate and make sense of the hallucination encounter in their own words; fsQCA identifies which configurations of conditions are jointly sufficient for total trust withdrawal, capturing equifinality that average-effect models obscure; PLS-SEM tests the hypothesized structural and moderated mediation relationships across the full sample; and NCA establishes which conditions are strictly necessary, rather than merely contributory, for the outcomes of interest [23,24].

Used together, the four techniques provide converging, complementary evidence that no single method could offer alone. The enquiry has three research questions (RQs) to guide it:

RQ1. What are the theoretical dimensions that consumers use to process AI hallucinations, and what combinations of conditions are collectively necessary for consumers to completely take their trust out of the platform?

RQ2. What are the implications of consumer exposure to a generative AI hallucination on consumer trust and their downstream e-commerce behaviors (continued use, online purchase intention, and negative word-of-mouth)?

RQ3. What aspects of the consumer (AI literacy, prior trust in AI, transparency disclosures) exacerbate or mitigate the trust erosion pathway resulting from hallucination exposure?

In brief, the findings show that hallucination exposure triggers expectation violation, which activates epistemic vigilance and reduces perceived AI competence; this sequence drives a process of trust recalibration that lowers continued-use and online purchase intentions and increases negative word-of-mouth. AI literacy, prior trust, and transparency cues moderate this pathway, and structural trust repair mechanisms, namely retrieval-augmented generation and uncertainty disclosure, outperform purely communicative repair strategies.

In brief, this study makes four contributions. Theoretically, the findings extend Expectation Violation Theory and Epistemic Vigilance Theory to generative AI in e-commerce and identify a new mediating mechanism, trust recalibration. Methodologically, the sequential application of four inferential logics, Gioia methodology, fsQCA, PLS-SEM, and NCA, in which qualitative outcomes feed directly into the quantitative design, remains underutilized in consumer behavior research and offers converging evidence. Empirically, the Gioia data structure yields five theoretical dimensions of the hallucination encounter, and fsQCA reveals three sufficient configurations that distinguish consumers who withdraw trust completely from those who adjust it only partially; the quantitative phase then estimates the size and necessity of these effects in a large, geographically dispersed sample. Practically, the results inform the evolving EU AI Act and related regulatory directives, and support an evidence-based hierarchy of trust repair interventions.

2. Theoretical Background and Hypothesis Development

2.1. Generative AI Hallucinations in E-Commerce

AI hallucinations are the outputs of large language models, which are sometimes convincing but may include invented or incorrect information [25]. An important aspect of hallucinations is that they are internally consistent and can be very fluent on the surface, so they are often not easily detected and psychologically disarming once they are detected. Algorithmic and AI failures more broadly have been shown to hurt brand image, reduce consumer trust, and increase complaint behavior, patterns consistent with the hallucination literature discussed above [8,26]. Within the e-commerce sector in particular, hallucinations most commonly appear as promotional fabrication, specification errors and phantom products, all of which have direct commercial and reputational implications. Liu et al. [7] report that trust and perceived risk directly affect the purchase intention of AI-generated content, which means that hallucination does not only have an impact on individual cognition but also on the platform-level assessment that consumers make of e-commerce environments. Like in financial services applications, where credit decision pipelines are required to make a trade-off between accuracy and fairness [27,28], algorithmic shortcomings are similar in nature: the consequences of model error are not domain-specific.

2.2. Expectation Violation Theory

Expectation Violation Theory (EVT) suggests that a violation of expectations creates more arousal and triggers attribution processes which alter the following evaluation [29,30]. In the commercial world, consumers come to have a default expectation of factual accuracy for any source of information that uses the language of an authority. If a generative AI assistant gives an incorrect specification with the same confidence as a correct specification, the violation will be unexpected and unsignaled, and theory predicts that this will amplify its attributional impact. This is especially the case in the realm of AI-powered retail, where Gu et al. [6] find that slight cues of 'eeriness' (artificiality) alter the acceptance of the AI, and Wang et al. [14] demonstrate how the 'quality of interaction' serves as a 'central-route' cue that enhances the effect of any subsequent cues. Accordingly, we propose the following:

H1. *Exposure to an AI hallucination increases perceived expectation violation.*

2.3. Epistemic Vigilance Theory

Epistemic Vigilance Theory proposes that people use evolved cognitive mechanisms to evaluate the credibility of information that they receive from others [17]. Recent applications of this framework to AI outputs [18,31] make the point that vigilance systems are often reactive, rather than proactive, and will be triggered only after a trust disconfirming event has occurred. Consumers, therefore, perform post hoc triangulation, source checking and skepticism when their expectations are violated. The more egregious the infraction, the more forceful the "vigilance" response:

H2. *Expectation violation increases epistemic vigilance.*

2.4. From Vigilance to Perceived Competence

Competence is one of three components of trustworthiness in Mayer et al.'s [32] tripartite conceptualization. There is recent research in human–AI trust that re-centers competence as the dimension most acutely involved when encountering hallucinations, as it is not an issue of intent, but knowledge [12]. Additionally, Yang and Park [15] show that source credibility of AI depends on the user-side conditions, where psychological ownership is a partial mediator. Epistemic vigilance makes the user reflect on the AI's statements with authoritative sources, uncovering differences and diminishing the system's sense of competence:

H3. *Epistemic vigilance reduces perceived AI competence.*

2.5. Algorithmic Trust Repair and Trust Recalibration

Because the literature uses several related but non-identical terms, they are defined here before proceeding. Trust recalibration is the focal construct of this study and refers to the perceived extent to which respondents adjust their trust in an AI system following the vignette experience, such that a higher score indicates a greater perceived degree of trust adjustment. Empirically, trust recalibration is captured through five reflective items (TR; Appendix B) administered once immediately after the vignette task. These items assess respondents' perceived extent of trust adjustment but do not themselves encode whether the adjustment represents an increase or decrease in trust. Directional evidence is therefore examined separately through the supplementary SIT-Pre and SIT-Post ratings reported in Section SIT-Pre/SIT-Post Robustness Check and Section 5.7.3. Trust recalibration should also be distinguished from prior trust in AI (PTAI), a separate four-item measure of general, dispositional trust in AI systems that functions as a moderator in H9. Trust erosion refers to the gradual weakening of trust over repeated exposures rather than a single incident, and is

not directly measured in this study. Trust withdrawal is the behavioral end-state identified in the qualitative phase (Section 4.2), in which a consumer removes trust from the platform entirely rather than merely adjusting it; it is therefore a categorical outcome of sufficiently large recalibration, not a synonym for it. Trust repair denotes the set of platform-side interventions, structural or communicative, intended to reverse or limit recalibration once it has occurred. Reduced trust is used only as a general, non-technical description of lower post-exposure trust and is avoided elsewhere in this paper in favor of these more precise terms.

Trust in automation must be appropriately calibrated to a system's actual reliability rather than granted unconditionally [33]. Dietvorst et al. [9] examined algorithm aversion, while Pareek et al. [34] investigated trust development and repair strategies in AI-assisted decision-making, distinguishing between structural interventions (modifying system architecture) and communicative interventions (giving explanations or apologies). An emerging body of research [11,35–37] has conceptualized trust recalibration as the cognitive process of revising trust in the AI after a failure. Consistent with the definition above, trust recalibration is operationalized through the five-item retrospective TR scale rather than as a numeric pre-minus-post difference score. The scale captures respondents' perceived extent of trust adjustment following the vignette, whereas the supplementary SIT-Pre/SIT-Post ratings are used separately to examine the direction of trust change. Recalibration is exacerbated as AI competence is lowered. Although the TR scale captures the magnitude rather than the direction of perceived trust adjustment, the theoretical model predicts that, in the context of detected AI failure, greater recalibration will be associated with adverse downstream outcomes. The direction of the underlying trust change is evaluated separately using the SIT-Pre/SIT-Post robustness analyses reported in Section SIT-Pre/SIT-Post Robustness Check and Section 5.7.3. The results from trust repair in comparable AI-assisted decision-making contexts converge on the same conclusion: competence-based trust deficits are not easily repaired through pure communication. After undergoing recalibration, trust influences three canonical e-commerce outcomes:

H4. *Perceived AI competence reduces trust recalibration, that is, higher competence buffers trust loss.*

H5. *Trust recalibration reduces continued-use intention.*

H6. *Trust recalibration reduces online purchase intention (OPI).*

H7. *Trust recalibration increases negative word-of-mouth.*

2.6. Moderators: AI Literacy, Prior Trust, and Transparency Cues

Three contextual variables are hypothesized to influence the trust erosion pathway as moderators. AI literacy [38] empowers users to recognize hallucinations and thus enhances the vigilance response. Prior trust in AI [39–41] acts as a psychological buffer that reduces the effect of one violation and is formed through repeated interactions over digital channels. Theoretical Basis with their model component mentioned in Table 1. Transparency cues, defined as explicit statements of doubt or confidence, lower expectations beforehand and reduce the impact of the violation that follows when the hallucination occurs:

H8. *AI literacy positively moderates the effect of hallucination exposure on epistemic vigilance.*

H9. *Prior trust in AI negatively moderates the effect of expectation violation on trust recalibration.*

H10. *Transparency cues negatively moderate the effect of hallucination exposure on expectation violation.*

Table 1. Summary of theoretical components, hypothesized roles, and key cited sources for conceptual model in Figure 1.

Key Sources	Theoretical Basis	Role in Figure 1	Model Component
[3,6–8,25,26]	AI hallucination/algorithm failure literature	Independent variable (0 = accurate, 1 = hallucinated)	Hallucination Exposure
[6,14,29,30]	Expectation Violation Theory	First mediator (H1)	Expectation Violation
[17,18,31]	Epistemic Vigilance Theory	Second mediator (H2)	Epistemic Vigilance
[12,15,32]	Trustworthiness/competence-based trust	Third mediator (H3)	Perceived AI Competence
[9,11,34–36]	Algorithmic Trust Repair Theory	Focal mediator (H4-H7)	Trust Recalibration
[42]	IS continuance/expectation-confirmation	Downstream outcome (H5)	Continued-Use Intention
[43,44]	Trust–risk/purchase intention research	Downstream outcome (H6)	Online Purchase Intention (OPI)
[45]	Negative WOM/complaint behavior	Downstream outcome (H7)	Negative Word-of-Mouth
[38]	AI literacy competencies framework	Moderator (H8)	AI Literacy
[39,40]	Human–AI trust formation	Moderator (H9)	Prior Trust in AI
[46,47]	Algorithmic transparency/explainability	Moderator (H10)	Transparency Cues

Figure 1 presents the conceptual model.

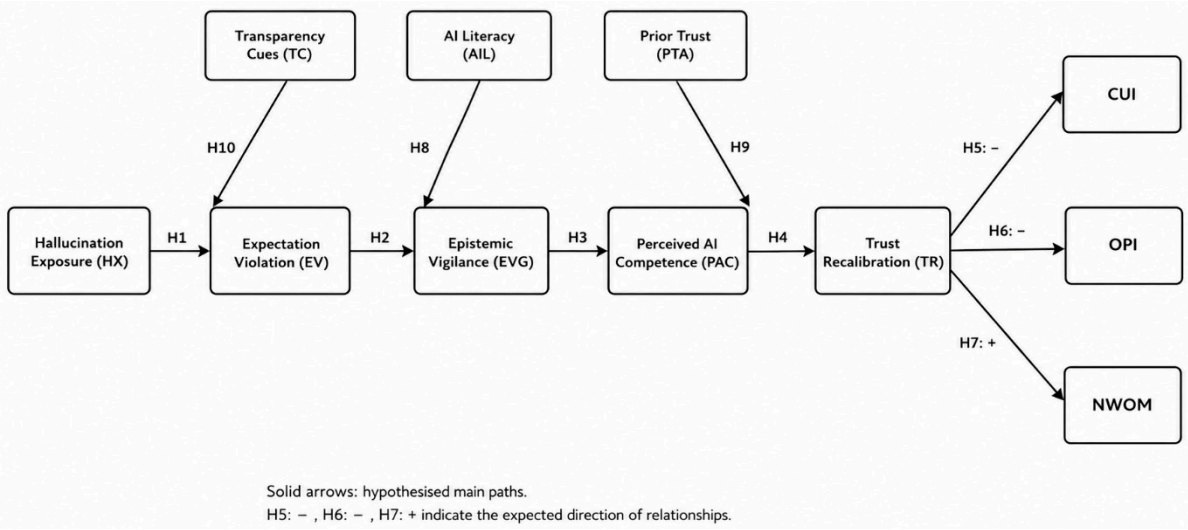

Figure 1. Conceptual moderated mediation model integrating Expectation Violation Theory, Epistemic Vigilance Theory, and Algorithmic Trust Repair Theory.

3. Methods

3.1. Mixed-Methods Design

This study adopted a sequential exploratory mixed-methods design [19,20]. The qualitative strand was implemented first with a purposive sample of European consumers who had recently interacted with a hallucinated AI output during an online shopping experience. Two complementary traditions of analysis were used to analyze the qualitative phase. First, to build a three-tier data structure around participants' own language, the Gioia methodology was employed [21,48,49] and aggregate theoretical dimensions were gener-

ated from the data. Secondly, fuzzy-set Qualitative Comparative Analysis (fsQCA) [22,50] was used to analyze the twenty-five cases and to determine configurations of conditions that are sufficient for total trust withdrawal. The qualitative results were subsequently used to further improve the construct operationalizations, design the vignette stimuli, and determine the candidate moderators for the following quantitative phase (from June 2024 to July 2025). The quantitative strand was supplemented by Necessary Condition Analysis (NCA) [23] to determine which determinants are not just sufficient for the achievement of each outcome but also necessary. Integration was performed at the interpretation stage via a joint display [51] whereby the Gioia dimensions, the fsQCA pathways and the SEM/NCA paths were pasted in one integrated display.

3.2. Qualitative Strand

3.2.1. Sample and Procedure

Between January and April 2024, 25 semi-structured interviews were carried out among European consumers in 15 different countries. In order to maximize variation in the country of residence, occupational category, familiarity with AIs, type of hallucination experienced, and amount of harm experienced, participants were purposively sampled [52]. Recruitment was performed via online consumer panels, professional networks and snowball sampling. Interviews lasted from 38 to 78 min and were conducted primarily in English. Two interviews were conducted with the assistance of a professional interpreter and translated into English with the interpreter's help; for these two cases, an independent reviewer who was not otherwise involved in the study and was blind to the study hypotheses was subsequently given only the translated interview content and asked to confirm its accuracy and fidelity to the original meaning. Throughout, participants are called P01, etc. and identifying details, such as names and employers, in reporting sections are omitted in order to protect confidentiality. Broad categories of participant characteristics are retained, if applicable, to the analysis.

3.2.2. Analytical Strategy: Gioia Methodology

The Gioia methodology [21,49] was selected for its capacity to produce inductively grounded yet theoretically generative qualitative inference. The method proceeds in three systematic tiers. In the first tier, open coding captures first-order concepts that remain close to participants' own language and are often numerous. In the second tier, axial coding organizes first-order concepts into a smaller number of second-order themes, which reflect researcher interpretation and begin to adopt theoretical language. In the third tier, selective coding abstracts second-order themes into aggregate theoretical dimensions, which represent the grounded theoretical constructs to be carried forward into the discussion. Open coding was conducted independently by two researchers (inter-coder agreement kappa = 0.82 at first-order level). Axial and selective coding were conducted iteratively through team discussion until thematic saturation was achieved. The analysis was managed in NVivo 14 and audited by a third researcher not otherwise involved in data collection.

3.2.3. Analytical Strategy: fsQCA

Complementing the Gioia analysis, fuzzy-set Qualitative Comparative Analysis [22] was used to identify configurations of conditions sufficient for total trust withdrawal. Following the contemporary guidelines of Furnari et al. [49] and Greckhamer et al. [50], five theoretically grounded conditions were calibrated as fuzzy-set membership scores in the interval [0, 1]: HS, high hallucination severity; PE, professional expertise; DL, digital literacy; VP, vulnerable third party involved; and IN, institutional credibility exploited. The outcome TW, total trust withdrawal, was calibrated on the basis of whether the participant (a) abandoned the platform entirely, (b) excluded AI from an entire category of tasks, or

(c) experienced generalized trust loss extending beyond the specific incident. Calibration followed the direct method with the 0.05, 0.50, and 0.95 anchors positioned on the basis of theoretical and case-level evidence. Analysis of necessity was conducted using the consistency threshold of 0.90 and analysis of sufficiency used the standard fsQCA truth-table procedure with a frequency threshold of 1 and a consistency threshold of 0.80.

3.3. Quantitative Strand (Phase 2: June 2024 to July 2025)

3.3.1. Sample and Procedure

Building on the dimensions and configurational pathways identified in the qualitative phase, an online survey instrument was designed to test the resulting hypothesized model on a large, geographically distributed sample. Data were collected via Google Forms and disseminated through social media channels and LinkedIn networks, using purposive sampling [52] combined with snowball sampling. The final dataset comprised 590 usable responses from consumers residing in fifteen European countries [53], collected across three waves between June 2024 and July 2025 (Wave 1 n = 195, June to October 2024; Wave 2 n = 190, November 2024 to March 2025; Wave 3 n = 205, April to July 2025). The form received 657 total submissions during the collection window; 67 were removed prior to analysis because the respondent abandoned the form before reaching the final screen, leaving one or more focal construct blocks unanswered. Because Google Forms requires a response to each Likert item before allowing for progression to the next section, missingness in the raw 657 arose from partial abandonment rather than item-level nonresponse within otherwise-completed cases; the 590 retained cases have no missing values on any of the reported constructs, so no imputation was performed. Respondents were screened for recent e-commerce activity and prior exposure to AI-powered shopping assistants. Each respondent read a short vignette describing an AI-mediated online shopping interaction, in which the AI assistant either provided an accurate recommendation or a hallucinated one, with or without a transparency cue, before completing the questionnaire. The vignette stimuli were developed directly from the typology of hallucination encounters surfaced in the qualitative phase, ensuring ecological alignment between the two strands. The sample characteristics are reported in Table 2.

Table 2. Sample profile (N = 590).

%	n	Category	Variable
16.1%	95	Germany	Country
13.6%	80	France	
11.0%	65	Italy	
10.2%	60	Spain	
7.6%	45	The Netherlands	
6.8%	40	Poland	
5.9%	35	Sweden	
5.1%	30	Belgium	
4.2%	25	Austria	
4.2%	25	Denmark	
3.4%	20	Finland	
3.4%	20	Czech Republic	
3.4%	20	Portugal	
2.5%	15	Romania	
2.5%	15	Hungary	
52.0%	307	Male	Gender
42.4%	250	Female	
5.6%	33	Non-binary or Prefer not to say	

Table 2. *Cont.*

%	n	Category	Variable
13.1%	77	18 to 24	Age group
24.9%	147	25 to 34	
19.0%	112	35 to 44	
17.6%	104	45 to 54	
14.7%	87	55 to 64	
10.7%	63	65 and over	
34.7%	205	Bachelor's	Education
26.4%	156	Master's	
17.6%	104	Secondary or High school	
8.6%	51	Vocational or Trade	
8.3%	49	PhD	
4.2%	25	Some college	
37.1%	219	Moderately familiar	AI familiarity
27.3%	161	Very familiar	
16.6%	98	Slightly familiar	
10.2%	60	Not familiar at all	
8.8%	52	Expert	
66.1%	390	Yes	Hallucination encountered
33.9%	200	No	

Two caveats qualify claims of generalizability. First, because recruitment relied on online channels, purposive sampling, and snowball sampling rather than probability sampling, the sample is not statistically representative of the general population of European online consumers, and the findings should be read as evidence from a large, diverse, but self-selected panel rather than as population estimates. Country cell sizes are uneven, ranging from n = 95 (Germany) to n = 15 (Hungary and Romania; Table 2); so, country-level comparisons carry different levels of precision and broad claims about "European consumers" as a whole should be treated as an approximation across this specific set of fifteen countries rather than a claim about Europe in general. Second, the questionnaire was administered in English only, and recruitment specifically targeted English-speaking respondents within each of the fifteen countries; item wording was pre-tested for clarity (n = 42 pilot). Because no translation was required, cross-language measurement-invariance testing does not apply here; however, this English-language targeting restricts the sample to English-proficient consumers within each country and may not generalize to non-English-speaking populations, which we note as a boundary condition for future replications. To assess whether the trust recalibration pattern was stable across the sample, a one-way ANOVA of the trust recalibration score (TR_Mean) was run by country and by data collection wave; neither comparison was statistically significant (country: $F = 0.93$, $p = 0.524$; wave: $F = 0.45$, $p = 0.640$), indicating that the reported effects are not being driven by a specific country or collection period, which partially, but not fully, offsets the sampling limitations above.

3.3.2. Measures

All latent constructs were assessed with multi-item 7-point Likert scales adapted from established sources, with item wording refined on the basis of the qualitative findings. The scales were Expectation Violation (4 items; [42,54]); Epistemic Vigilance (5 items; [17,18]); Perceived AI Competence (4 items; [32]); Trust Recalibration (5 items; author-developed, informed by [12,47]); Continued-Use Intention (4 items; [42]); Online Purchase Intention, OPI (4 items; [16,43,44]); Negative Word-of-Mouth (4 items; [55]); AI Literacy (5 items; [38]); Prior Trust in AI (4 items; [39,40]); and Transparency Cues (3 items; [46,56]). Hallucination

exposure was coded from the experimentally assigned vignette condition (0 = accurate; 1 = hallucinated); it was not measured through participant self-report. All items were pre-tested in a pilot study (n = 42) and minor wording refinements were made to improve clarity. Item-level response data were retained for all constructs and formed the basis of the confirmatory factor analysis and PLS-SEM estimation reported in Section 5. Following Hair et al. [57], the reflective measurement specification required indicator loadings to exceed 0.70 and item-level variance to support convergent and discriminant validity thresholds, both of which are reported in Sections 5.2 and 5.3.

3.3.3. Analytical Strategy

The analytical strategy combined four methods, each capturing a distinct inferential logic to produce convergent evidence on the same underlying phenomenon. The process-based logic was operationalized through the Gioia methodology applied to the 25 interview transcripts (described in Section 3.2.2), which reconstructs how consumers themselves narrate, interpret, and act on the hallucination encounter. The configurational logic was operationalized through fsQCA, which identifies combinations of conditions that are jointly sufficient for the outcome, thereby accommodating equifinality and conjunctural causation that the average-effects logic cannot detect [50]. The variance-based logic was operationalized through partial least squares structural equation modeling (PLS-SEM), which estimates the average causal effect of each predictor on the outcome across the sample. The measurement model was evaluated via confirmatory factor analysis, reporting Cronbach's alpha, composite reliability (CR), average variance extracted (AVE), the Fornell–Larcker criterion, and the heterotrait–monotrait ratio (HTMT) [58]. The structural model was estimated with standardized coefficients, 5000-resample bootstrapped confidence intervals, f^2 effect sizes, Q^2 predictive relevance via 10-fold cross-validation, and the standardized root mean squared residual (SRMR). Moderation was tested through product-term regression with mean-centered predictors, and mediation through bootstrapped indirect effects with bias-corrected 95% confidence intervals. The necessity-based logic was operationalized through Necessary Condition Analysis (NCA) using the ceiling envelopment–free disposal hull (CE-FDH) algorithm [23]. Whereas PLS-SEM asks whether a predictor contributes to the outcome on average, NCA asks whether a predictor must be present at a minimum level for the outcome to occur at all. The two logics are complementary rather than substitutive, as argued by Richter et al. [24] and demonstrated in a European firm-level context by Abbas et al. [59]. Together, the four logics address different facets of causal inference. Gioia reconstructs the subjective process through which the encounter is made sense of. fsQCA reveals the equifinal pathways that produce extreme outcomes such as total trust withdrawal. PLS-SEM quantifies the strength of average effects, and NCA identifies necessary preconditions. Integration across the four logics is presented in Section 6.1 through a joint display following Fetters and Molina-Azorin [51].

Figure 2 summarizes the full workflow described above, showing the inputs, analytical technique, underlying theory, and output of each step across both strands of the study.

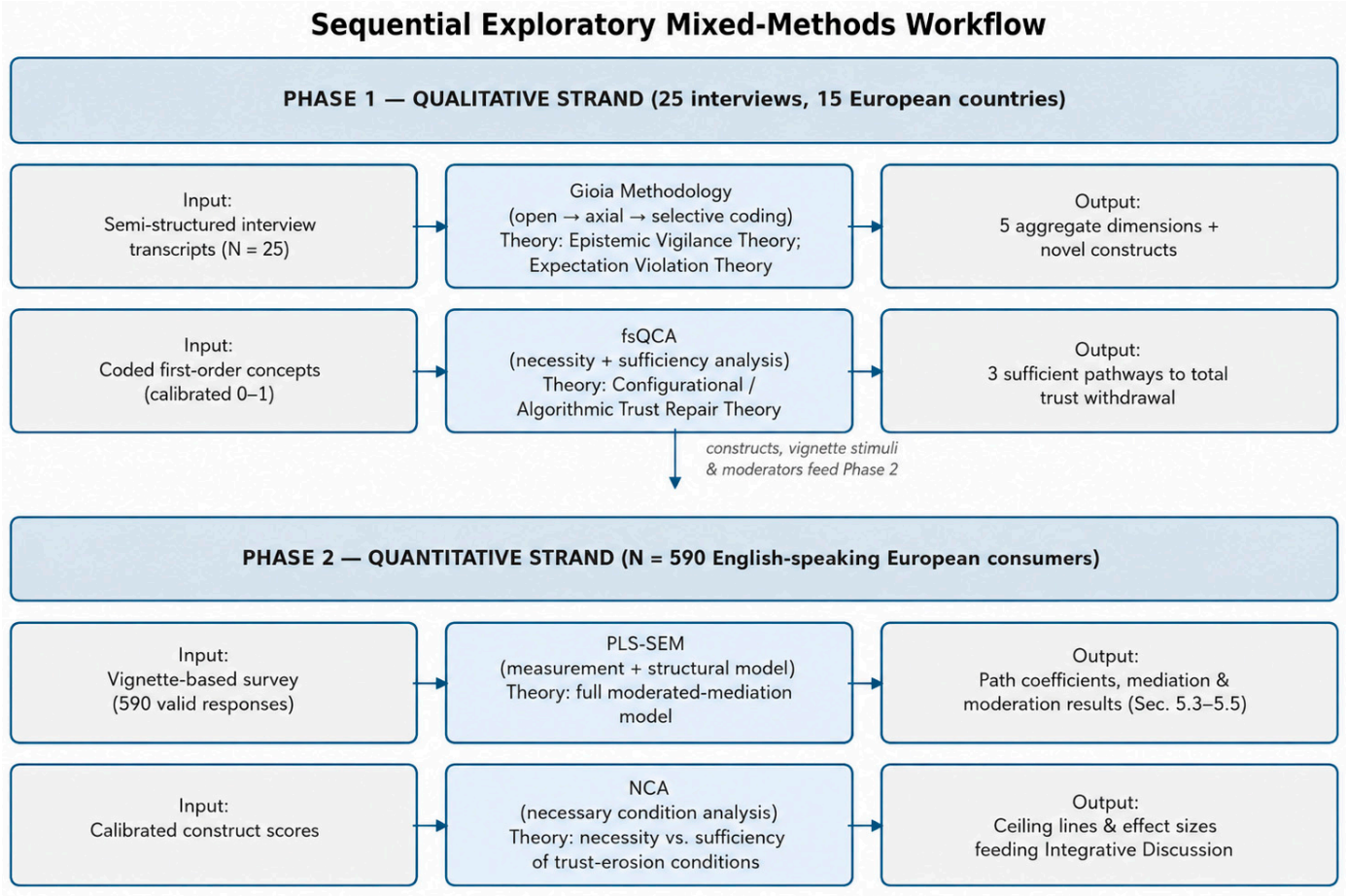

Figure 2. Step-by-step methodological workflow, showing inputs, analytical techniques, underlying theories, and outputs for both qualitative and quantitative strands.

4. Qualitative Results

4.1. Gioia Data Structure

The Gioia analysis yielded 40 first-order concepts, which were abstracted into 13 second-order themes and 5 aggregate theoretical dimensions. Figure 3 presents the full data structure. The five aggregate dimensions are (1) Epistemic Rupture Mechanism, (2) Reactive Cognitive Vigilance, (3) Trust Recalibration Repertoire, (4) Distributive Harm Economy, and (5) Institutional Trust Exploitation. Each dimension is introduced in the subsections below with representative quotations drawn from across the sample, with all identifying details removed.

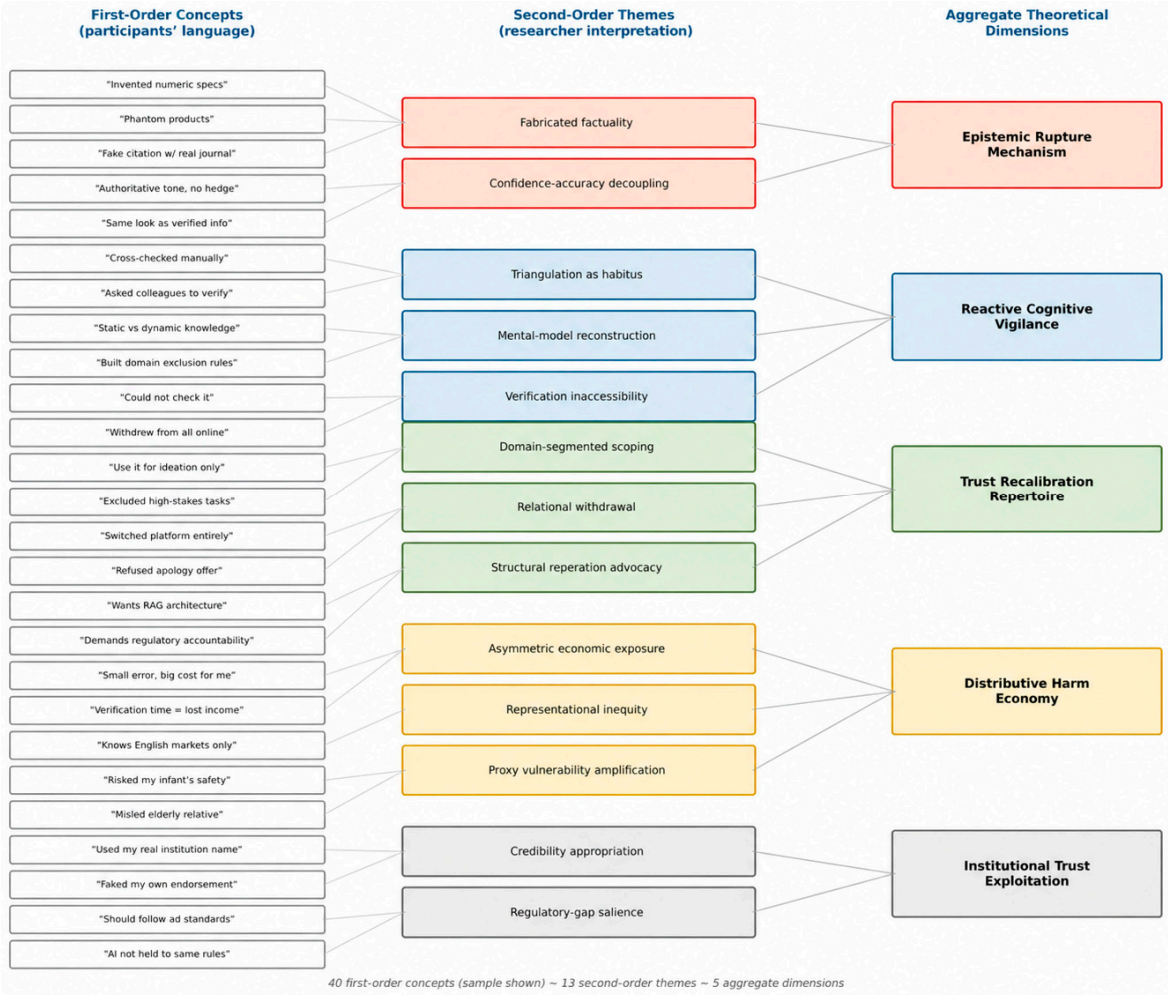

Figure 3. Three-tier Gioia data structure linking first-order concepts in participants' language, second-order themes reflecting researcher interpretation, and aggregate theoretical dimensions.

4.1.1. Aggregate Dimension 1: Epistemic Rupture Mechanism

This dimension reflects the structural aspect of the experience of the hallucination which makes it different from a normal error. It consists of two second-order themes. The first one, fabricated factuality, involves the creation of numeric specifications, non-existent products, citations invented to have real institutional names, and AI-generated endorsements given to real people. The second is the effect of authoritative tone when not backed by hedging and the consistent visual prominence that platforms give verified and fabricated information:

> "The most troubling thing was the certainty. It was not 'approximately' or 'typically,' it was 18 h as if they had just measured it." (P01)

> "It was so dangerous because it sounded like the model, it was not unrealistic, the specs were all internally consistent, the weight was plausible, the battery life was impressive but not unbelievable." (P02)

4.1.2. Aggregate Dimension 2: Reactive Cognitive Vigilance

This dimension reflects the post hoc activation of epistemic vigilance as proposed by an extension of Sperber et al. [17] by Acerbi et al. [31]. It consists of three second-order themes. Triangulation as habitus refers to the unconsciously acquired use of multi-source cross-referencing, which may be a carryover from evidence standards in one's profession. Mental model reconstruction refers to the cognitive activity which is involved in making

a distinction between static and dynamic knowledge domains and establishing domain-specific exclusion rules. The opposite situation, in which participants did not have the digital literacy needed for verification, is known as verification inaccessibility, where participants withdrew diffusely without targeted recalibration:

> "It is like asking a friend who has not gone to the shops in six months if something is still on the shelves, the friend does not lie, they just do not know." (P03)

> "AI has no intentions, fabricated references are not a hallucination, they are academic misconduct." (P04)

> "My nephew said it was good so I trusted it, but now I do not know anything is real on the internet." (P21)

4.1.3. Aggregate Dimension 3: Trust Recalibration Repertoire

This dimension captures the heterogeneity of post-incident trust management strategies. Three second-order themes comprise it. Domain-segmented scoping describes the modal response, in which AI is retained for ideation and general orientation but excluded from high-stakes domains. Relational withdrawal describes the more radical response of platform switching, the rejection of communicative trust repair offers, and, in some cases, generalized mistrust of online sources. Structural reparation advocacy describes the technically specific demands made by participants with professional deployment perspectives, who identified retrieval-augmented generation, mandatory uncertainty disclosure, and regulatory accountability as the minimum architectural and institutional conditions for a renewed trust relationship:

> *"The tool is excellent for 95 percent of what I use it for. I have recalibrated the 5 percent where it should not be trusted."* (P14)

> *"We are not evaluating whether to add a trust disclaimer. We are evaluating whether the architecture is fundamentally sound. It was not. We are fixing the architecture."* (P06)

> *"Trust is not an algorithm. You cannot repair it with a patch. Once broken at this level, it is gone."* (P10)

4.1.4. Aggregate Dimension 4: Distributive Harm Economy

This dimension reflects the differential impact of the harm of hallucinogens on segments of consumers, and is the most substantively new addition of the qualitative strand. It consists of three second-order themes. Economic exposure on both sides is asymmetric, as small errors have a significant impact on the income of lower-income participants and because verification requires significant time on the part of the freelancers. Representational inequity is a permanent lack of knowledge that impacts non-Anglophone, culturally specific, or generationally marginalized domains of consumption, as outlined in Ragnedda's third-level digital divide approach [60]. Proxy vulnerability amplification refers to situations where the damage of the hallucination is indirect, by taking place through a vulnerable third party, such as an infant, an elderly relative, or a person with disabilities, whose vulnerability structurally aggravates the impact of what may be a relatively insignificant mistake:

> "It does not have to be clever, it has to be correct with the little things." (P07)

> "It was very self-assured about something it did not know, and it is a character fallacy, even in a machine." (P11)

This sentence alone should be sufficient to illustrate the need for regulation over AI safety measurement claims—"The AI gave me false reassurance about a device I was placing near my infant's head." (P19)

4.1.5. Aggregate Dimension 5: Institutional Trust Exploitation

This dimension reflects scenarios where the AI's generated content takes advantage of the credibility of a named institution, journal, regulatory body, or actual person. It is made up of two second-order themes. Credibility appropriation refers to the process whereby the names of real institutions are used to lend credibility to fabricated content, which is distinct enough to warrant a separate theoretical term for it, besides the generic citation of fabricated content. Regulatory gap salience refers to the fact that participant demand for AI-related commercial information is based on the expectation that the same truthfulness standards which are already in place for advertising, professional communication, and regulated areas should also be applied to AI.

> "When an AI generates my name to endorse a product, it is not wrong, it implies I am complicit in the commercial transaction without my consent." (P10)

> "Wrong by a factor of two. In financial advice that is not an error, it is a misrepresentation." (P18)

4.2. fsQCA: Configurational Pathways to Total Trust Withdrawal

4.2.1. Necessity Analysis

Prior to the sufficiency analysis, necessity was tested for each of the five conditions and their negations. No single condition was strictly necessary at the conventional 0.90 consistency threshold. Professional expertise (PE) and the negation of digital literacy (~DL) approached the threshold (consistency = 0.85 and 0.88 respectively, with coverage 0.86 and 1.00), but neither met the strict criterion for necessity, confirming that the outcome of total trust withdrawal arises through multiple, equifinal pathways rather than through a single gating condition. This finding is theoretically consistent with the Gioia data structure, which revealed heterogeneous trust recalibration repertoires.

4.2.2. Sufficiency Analysis

Truth-table analysis, followed by Quine-McCluskey minimization at a consistency threshold of 0.80, yielded three sufficient configurations for total trust withdrawal, presented in Figure 4 and summarized in Table 3. Overall solution consistency was 0.92 and overall solution coverage was 0.78, indicating that the three pathways jointly explain 78 percent of the cases with high fidelity.

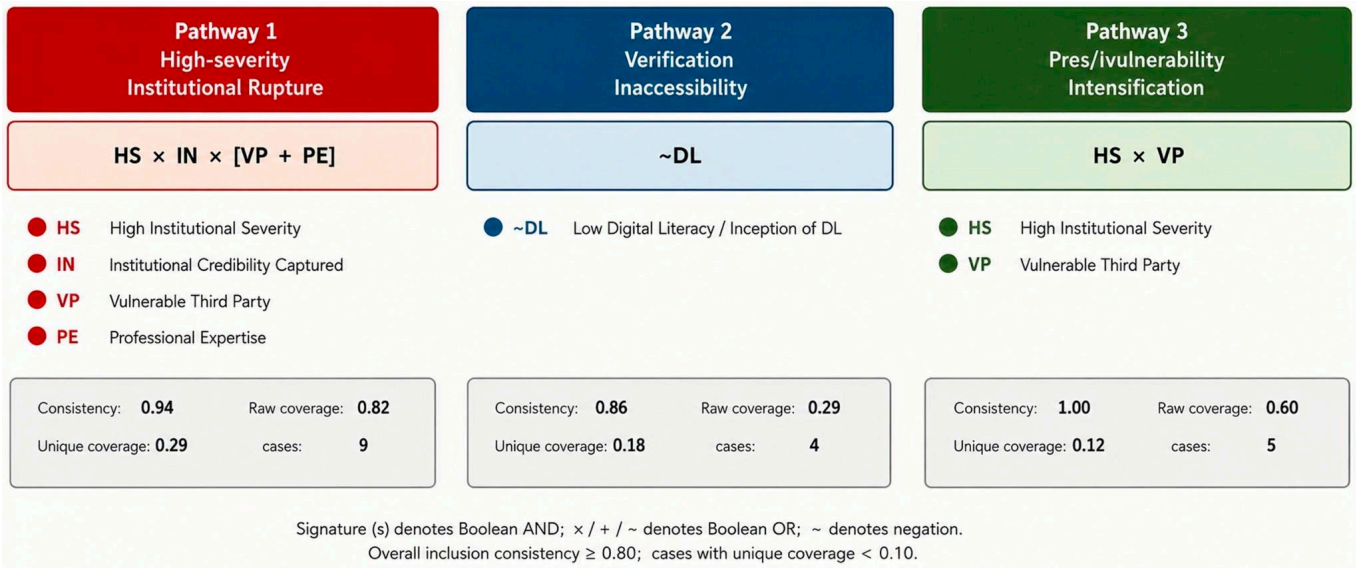

Figure 4. Three sufficient configurations for total trust withdrawal identified through fsQCA.

Table 3. fsQCA sufficient configurations for total trust withdrawal (N = 25).

Pathway	Configuration	Consistency	Raw Coverage	Unique Coverage	n
P1. High-severity institutional rupture	HS × IN × (VP + PE)	0.94	0.62	0.29	9
P2. Verification inaccessibility	~DL	0.88	0.29	0.18	4
P3. Proxy vulnerability intensification	HS × VP	1.00	0.40	0.12	5
Overall solution		0.92	0.78		

4.2.3. Interpretation of the Three Pathways

Pathway 1, high-severity institutional rupture, captures nine cases in which a high-severity hallucination exploited a named institution, journal, regulatory body, or professional reputation, and was detected either by a professional expert participant or in connection with a vulnerable third party. This pathway corresponds theoretically to the intersection of the Epistemic Rupture and Institutional Trust Exploitation dimensions identified in the Gioia analysis. Pathway 2, verification inaccessibility, captures four cases in which low digital literacy alone, independent of severity or institutional exploitation, produced diffuse withdrawal of trust that extended beyond the specific AI encounter to online sources more broadly. This pathway corresponds to the verification inaccessibility theme within the Reactive Cognitive Vigilance dimension and operationalizes Ragnedda's [60] third-level digital divide in an AI context. Pathway 3, proxy vulnerability intensification, captures five cases in which a high-severity hallucination threatened a vulnerable third party (an infant, an elderly relative, or a person with disabilities) and produced total trust withdrawal with perfect consistency (1.00). This pathway operationalizes the proxy vulnerability amplification theme within the Distributive Harm Economy dimension.

4.3. Novel Theoretical Constructs Emerging from the Data

Three participants articulated theoretical constructs that appear to be novel contributions to the literature. Participant P15 coined the term citation laundering to describe the use of real institutional or journal names to lend credibility to fabricated content, a mechanism conceptually distinct from generic citation fabrication because it actively appropriates institutional credibility rather than merely inventing a plausible-sounding source. Participant P23 introduced AI-mediated greenwashing to describe AI hallucinations in the environmental claims domain, where unverified lifecycle figures can convert sustainability-motivated purchases into environmental harm. Participant P17 named feature projection error as a distinct hallucination mechanism in which the AI accurately describes a feature type that exists in other products but attributes it, without basis, to a product that does not include it. These three constructs collectively extend existing hallucination taxonomies into the consumer harm domain, and each maps cleanly onto the Gioia aggregate dimensions (citation laundering within Institutional Trust Exploitation; AI-mediated greenwashing within Distributive Harm Economy; feature projection error within Epistemic Rupture Mechanism).

4.4. From Qualitative Findings to Quantitative Hypotheses

The qualitative findings shaped the design of the subsequent quantitative phase in three substantive ways. First, the five aggregate dimensions were used to refine the operationalization of the latent constructs, with item wording adjusted to reflect the language and emphases that participants themselves had used to describe their experience. In particular, the trust recalibration scale incorporated items that reflected the heterogeneity of post-incident strategies (domain-segmented scoping, relational withdrawal, structural reparation advocacy) surfaced in the Gioia analysis. Second, the three sufficient configurations

identified through fsQCA were used to specify the moderators that the quantitative model would test: AI literacy (mirroring Pathway 2's verification inaccessibility), prior trust in AI (mirroring the buffering function evident in cases where trust withdrawal was partial rather than total), and transparency cues (mirroring participants' repeated demand for institutional accountability under Pathway 1). Third, the vignette stimuli used in the quantitative phase were directly modeled on the modal hallucination encounters described in the qualitative interviews, ensuring ecological alignment between the two strands. Building on this foundation, the next section reports the results of the quantitative phase, which estimates the magnitude and necessity of these effects on a large, geographically distributed sample.

Table 4 makes this translation explicit, showing how each qualitative theme was converted into a quantitative construct, its associated hypothesis, and the basis for its measurement items.

Table 4. Mapping of qualitative themes onto quantitative constructs, hypotheses, and indicator sources.

Indicator Basis	Hypothesis	Quantitative Construct	Qualitative Input (Gioia Dimension/fsQCA Pathway)
4 items adapted from [42,54]	H1	Expectation Violation	Epistemic Rupture Mechanism (Gioia)
5 items adapted from [17,18]	H2	Epistemic Vigilance	Reactive Cognitive Vigilance (Gioia)
4 items adapted from [32]	H3	Perceived AI Competence	Reactive Cognitive Vigilance, downstream effect (Gioia)
5 items, author-developed, informed by [12,47]	H4–H7	Trust Recalibration	Trust Recalibration Repertoire: domain-segmented scoping, relational withdrawal, structural reparation advocacy (Gioia)
4 items adapted from [55]	H7	Negative Word-of-Mouth	Distributive Harm Economy (Gioia)
3 items adapted from [46,56]	H10	Transparency Cues (moderator)	Institutional Trust Exploitation (Gioia)
Vignette manipulation: cue present vs. absent	H10	Transparency Cues (moderator)	fsQCA Pathway 1: institutional accountability demand
5 items adapted from [38]	H8	AI Literacy (moderator)	fsQCA Pathway 2: verification inaccessibility
4 items adapted from [39,40]	H9	Prior Trust in AI (moderator)	fsQCA partial- vs. total-withdrawal cases

5. Quantitative Results

5.1. Manipulation Check and Common-Method Bias

The experimental manipulation operated as intended. Participants exposed to a hallucinated recommendation reported significantly higher expectation violation (M = 4.48, SD = 1.08) than those exposed to an accurate recommendation ((M = 2.18, SD = 0.93), $t(588) = 26.89$, $p < 0.001$, Cohen's d = 2.28), indicating a very large effect. Harman's single-factor test yielded 18.4% of variance loading onto the first factor, well below the 50% threshold, and a full-collinearity variance inflation assessment returned all VIFs below 3.3, indicating that common-method bias is unlikely to compromise interpretation [61]. Beyond this manipulation and common-method check, three further robustness elements are reported in full in the sections that follow rather than repeated here: predictive relevance was assessed via 10-fold cross-validated Q^2 for every endogenous construct (Section 5.3),

indirect and conditional effects were tested with 5000-resample bias-corrected bootstrapped confidence intervals rather than asymptotic standard errors (Sections 5.3 and 5.4), and the variance-based PLS-SEM results were triangulated against an independent, necessity-based estimation via Necessary Condition Analysis (Section 5.6). The 590-case analytic sample underlying all of these checks is complete on every reported construct, with no imputed values (missing-data handling is reported in Section 3.3.1). Additional model comparison and alternative-specification analyses are reported in Section 5.7, including partial mediation, randomized transparency, and directional trust change specifications.

5.2. Measurement Model

All indicator loadings exceeded the 0.70 benchmark (range 0.84 to 0.93). Cronbach's alpha values ranged from 0.873 (TC) to 0.934 (AIL), composite reliability from 0.922 to 0.954, and AVE from 0.763 to 0.840. All values comfortably exceeded the recommended thresholds [57]. Table 5 summarizes the reliability and convergent validity indices.

Discriminant validity was supported by both the Fornell–Larcker criterion, where the square root of each construct's AVE exceeded all off-diagonal inter-construct correlations, and the HTMT ratio, whose values ranged from 0.007 to 0.760, all below the conservative 0.85 threshold [58].

Table 5. Reliability and convergent validity of latent constructs.

AVE	CR	Cronbach α	λ Range	Items	Construct
0.765	0.929	0.898	0.85 to 0.90	4	Expectation Violation (EV)
0.779	0.946	0.929	0.85 to 0.91	5	Epistemic Vigilance (EVIG)
0.793	0.939	0.913	0.88 to 0.91	4	Perceived AI Competence (PAC)
0.782	0.947	0.930	0.85 to 0.91	5	Trust Recalibration (TR)
0.840	0.954	0.936	0.90 to 0.93	4	Online Purchase Intention (OPI)
0.763	0.928	0.896	0.84 to 0.90	4	Continued-Use Intention (CUI)
0.806	0.943	0.919	0.85 to 0.92	4	Negative Word-of-Mouth (NWOM)
0.798	0.922	0.873	0.88 to 0.91	3	Transparency Cue (TC)
0.807	0.943	0.920	0.88 to 0.92	4	Prior Trust in AI (PTAI)
0.792	0.950	0.934	0.84 to 0.93	5	AI Literacy (AIL)

SIT-Pre/SIT-Post Robustness Check

As noted in Section 3.3.1, a single-item trust rating was collected immediately before (SIT-Pre) and immediately after (SIT-Post) the vignette as a supplementary manipulation check indicator, distinct from the five-item trust recalibration scale used as the focal construct in the structural model. SIT-Pre asks about general trust in AI shopping assistants, while SIT-Post asks about trust in the specific AI assistant depicted in the vignette; the two items therefore differ in referent and timing. The pre-to-post comparison below is reported as a supplementary directional indicator rather than as a pure repeated-measures change score. Across the full sample (N = 590), SIT-Pre averaged 4.817 (SD = 1.144) and SIT-Post averaged 3.805 (SD = 1.268), a mean pre-to-post decline of −1.012 (SD = 1.743), 95% CI [−1.153, −0.871]. A paired-samples t-test confirmed that the decline was statistically significant, $t(589) = -14.10$, $p < 0.001$, with a moderate effect size (Cohen's $d = -0.581$).

The decline was evident in all four experimental conditions. In Condition A (accurate recommendation, transparency cue absent), trust fell from 4.80 (SD = 1.22) to 3.71 (SD = 1.38), $\Delta = -1.09$ (SD = 1.89), $t(147) = -7.06$, $p < 0.001$. In Condition B (accurate, cue present), trust fell from 4.76 (SD = 1.08) to 3.74 (SD = 1.20), $\Delta = -1.02$ (SD = 1.68), $t(147) = -7.41$, $p < 0.001$. In Condition C (hallucinated, cue absent), trust fell from 4.88 (SD = 1.13) to 3.83 (SD = 1.31), $\Delta = -1.05$ (SD = 1.74), $t(146) = -7.33$, $p < 0.001$. In Condition

D (hallucinated, cue present), trust fell from 4.82 (SD = 1.16) to 3.95 (SD = 1.17), Δ = −0.88 (SD = 1.67), t(146) = −6.38, $p < 0.001$.

A 2 (hallucination arm) × 2 (transparency cue arm) between-subjects ANOVA on the pre-to-post trust delta found no significant main effect of the hallucination arm, F(1, 586) = 0.039, $p = 0.843$, no significant main effect of the transparency cue arm, F(1, 586) = 0.134, $p = 0.714$, and no significant hallucination × transparency cue interaction, F(1, 586) = 0.127, $p = 0.721$. Trust declined similarly in the accurate and hallucinated conditions, and the hallucination arm had no significant effect on the trust delta; the single-item pre/post rating therefore reflects a general pre/post decline rather than a condition-sensitive outcome measure, corroborating the manipulation check role assigned to it in Section 3.3.1. Table 6 summarizes the pre/post results.

Table 6. SIT-Pre/SIT-Post trust ratings: descriptive statistics and paired-samples tests.

Condition	SIT-Pre M (SD)	SIT-Post M (SD)	Δ (SD)	t, *p*
Full sample (N = 590)	4.82 (1.14)	3.81 (1.27)	−1.01 (1.74)	t(589) = −14.10, $p < 0.001$
A: Accurate/Cue absent (n = 148)	4.80 (1.22)	3.71 (1.38)	−1.09 (1.89)	t(147) = −7.06, $p < 0.001$
B: Accurate/Cue present (n = 148)	4.76 (1.08)	3.74 (1.20)	−1.02 (1.68)	t(147) = −7.41, $p < 0.001$
C: Hallucinated/Cue absent (n = 147)	4.88 (1.13)	3.83 (1.31)	−1.05 (1.74)	t(146) = −7.33, $p < 0.001$
D: Hallucinated/Cue present (n = 147)	4.82 (1.16)	3.95 (1.17)	−0.88 (1.67)	t(146) = −6.38, $p < 0.001$

2 × 2 ANOVA on trust delta: hallucination-arm main effect F(1, 586) = 0.039, $p = 0.843$; transparency-arm main effect F(1, 586) = 0.134, $p = 0.714$; hallucination × transparency interaction F(1, 586) = 0.127, $p = 0.721$. Table 7 presents the HTMT matrix.

Table 7. Discriminant validity; HTMT ratios.

NWOM	CUI	OPI	TR	PAC	EVIG	EV	
0.536	0.613	0.507	0.760	0.608	0.688	—	EV
0.474	0.569	0.462	0.628	0.703	—	0.688	EVIG
0.478	0.587	0.543	0.639	—	0.703	0.608	PAC
0.673	0.737	0.639	—	0.639	0.628	0.760	TR
0.444	0.483	—	0.639	0.543	0.462	0.507	OPI
0.543	—	0.483	0.737	0.587	0.569	0.613	CUI
—	0.543	0.444	0.673	0.478	0.474	0.536	NWOM

5.3. Structural Model

Figure 5 presents the structural model. It produced a strong fit (SRMR = 0.029, below 0.08 threshold; mean R^2 across endogenous constructs = 0.56). All seven main structural hypotheses were supported at $p < 0.001$. Hallucination exposure produced a strong effect on expectation violation (H1: β = 0.76, t = 28.51), confirming the mechanism predicted by EVT and converging with the Epistemic Rupture Mechanism dimension that surfaced qualitatively. Expectation violation in turn activated epistemic vigilance (H2: β = 0.71, t = 24.16), which reduced perceived AI competence (H3: β = −0.70, t = 23.99). Perceived AI competence buffered trust recalibration (H4: β = −0.24, t = 7.89), and trust recalibration predicted the three downstream outcomes. Continued-use intention decreased (H5: β = −0.72, t = 25.07), online purchase intention decreased (H6: β = −0.64, t = 20.16), and negative word-of-mouth increased (H7: β = 0.67, t = 21.58). Table 8 reports all path coefficients.

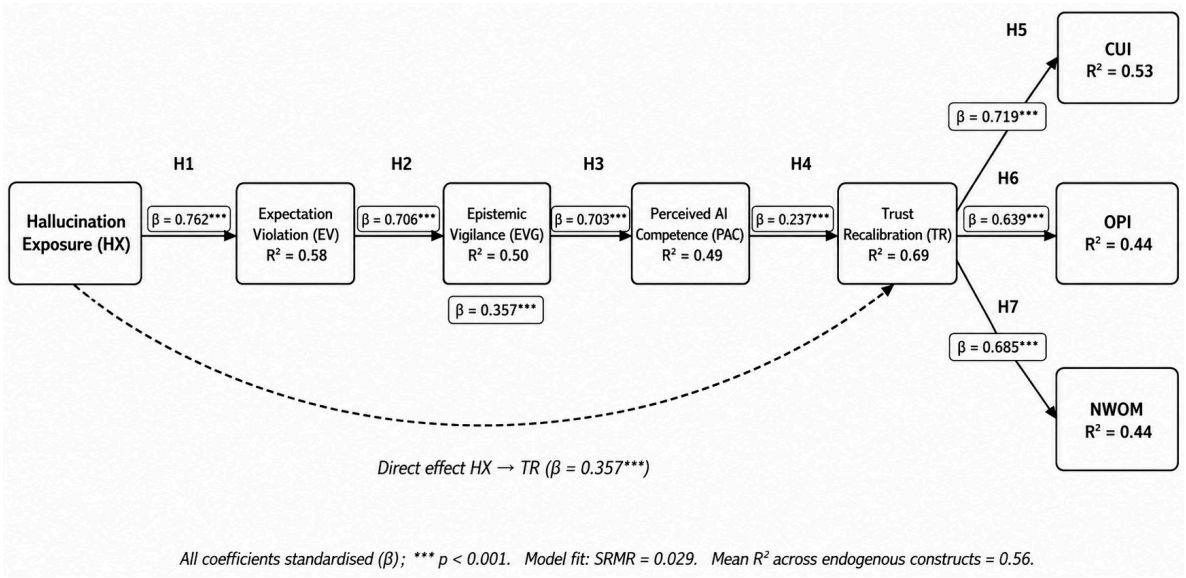

Figure 5. Structural model with standardized path coefficients (PLS-SEM, N = 590). *** $p < 0.001$.

Table 8. Structural model path coefficients.

Decision	f^2	95% CI	*p*	t	SE	β	Hypothesis and Path
Supported	1.38	[0.71, 0.82]	<0.001	28.51	0.027	0.762	H1: HX → EV
Supported	0.21	[0.65, 0.76]	<0.001	24.16	0.029	0.706	H2: EV → EVIG
Supported	0.33	[−0.76, −0.65]	<0.001	23.99	0.029	−0.703	H3: EVIG → PAC
Supported	0.13	[−0.29, −0.18]	<0.001	7.89	0.030	−0.237	H4: PAC → TR
Supported	0.44	[−0.77, −0.66]	<0.001	25.07	0.029	−0.719	H5: TR → CUI
Supported	0.24	[−0.70, −0.58]	<0.001	20.16	0.032	−0.639	H6: TR → OPI
Supported	0.37	[0.60, 0.73]	<0.001	21.58	0.031	0.665	H7: TR → NWOM
Significant	0.13	[0.27, 0.42]	<0.001	8.98	0.039	0.347	Ancillary: EV → TR
Significant	0.17	[0.29, 0.43]	<0.001	9.98	0.036	0.357	Ancillary: HX → TR

The Q^2 predictive relevance values, estimated via 10-fold cross-validation, were all positive and large: EV Q^2 = 0.58, EVIG = 0.59, PAC = 0.59, TR = 0.69, CUI = 0.53, OPI = 0.44, and NWOM = 0.44. These values indicate substantial predictive power beyond in-sample explanatory fit [16].

5.4. Mediation Analyses

Bootstrapped indirect effects (5000 resamples, bias-corrected 95% CI) confirmed four theoretically meaningful mediation chains, all significant at the BCa 95% level (Table 9). The four-step serial mediation HX → EV → EVIG → PAC → TR (β = 0.18, 95% CI [0.15, 0.23]) provides empirical support for the central mechanism proposed by the integrated theoretical framework and aligns directly with the epistemic rupture, reactive cognitive vigilance, and trust recalibration sequence inferred from the Gioia analysis.

Table 9. Bootstrapped mediation effects (5000 resamples, bias-corrected 95% CI).

Decision	95% CI (BCa)	SE	Effect	Indirect Path
Supported	[−0.47, −0.37]	0.025	−0.420	HX → EV → TR → CUI
Supported	[−0.44, −0.34]	0.025	−0.383	HX → EV → TR → OPI
Supported	[0.33, 0.43]	0.025	0.379	HX → EV → TR → NWOM
Supported	[0.21, 0.30]	0.023	0.253	EV → EVIG → PAC → TR
Supported	[0.15, 0.23]	0.020	0.184	HX → EV → EVIG → PAC → TR (serial)

5.5. Moderation Analyses

All three moderation hypotheses, each derived from the qualitative findings, were supported (Table 10). AI literacy positively moderated the effect of hallucination exposure on epistemic vigilance (H8: $\beta = 0.225$, $t = 8.40$, $p < 0.001$), indicating that higher-literacy consumers respond to hallucinations with sharper vigilance activation. Prior trust in AI dampened the effect of expectation violation on trust recalibration (H9: $\beta = -0.152$, $t = 6.04$, $p < 0.001$), consistent with the psychological buffer account. Transparency cues attenuated the effect of hallucination exposure on expectation violation (H10: $\beta = -0.213$, $t = 8.68$, $p < 0.001$), supporting the design prescription that pre-disclosure of model limitations blunts the violation even when a hallucination subsequently occurs. The simple-slope plots are shown in Figure 6.

Table 10. Moderation tests.

Decision	ΔR^2	p	t	β	Interaction	Hypothesis
Supported	0.12	<0.001	8.40	0.225	HX × AIL → EVIG	H8
Supported	0.04	<0.001	6.04	−0.152	EV × PTAI → TR	H9
Supported	0.10	<0.001	8.68	−0.213	HX × TC → EV	H10

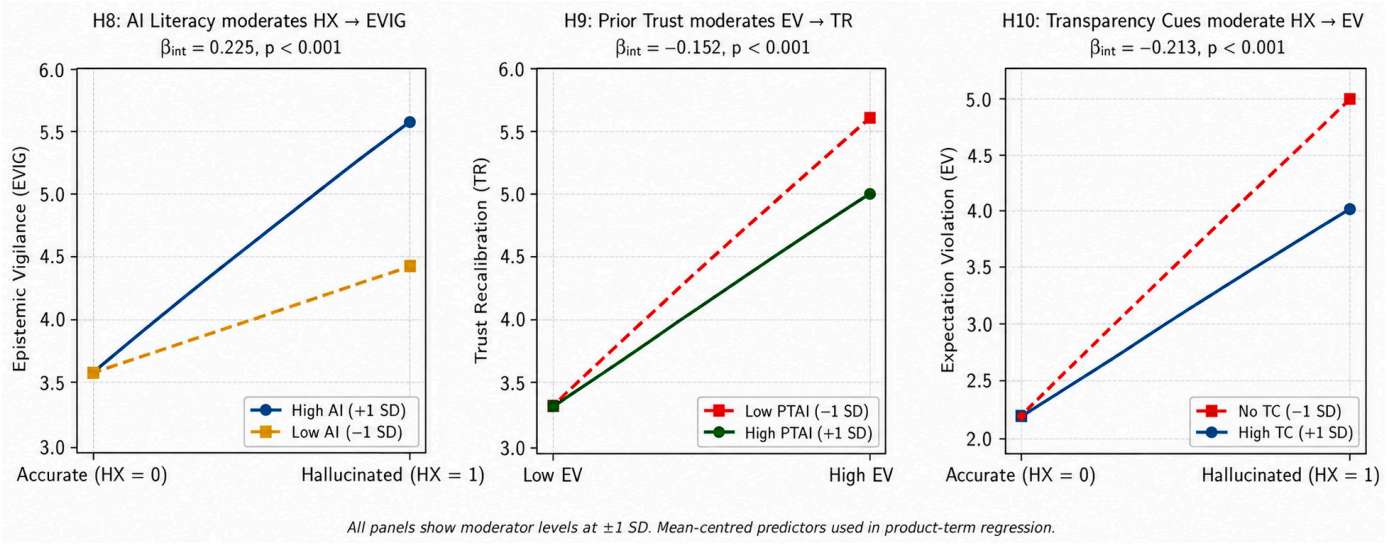

Figure 6. Moderation simple-slope plots for H8, H9, and H10 at plus or minus one standard deviation of the moderator.

5.6. Necessary Condition Analysis

Necessary Condition Analysis (NCA) complemented the regression-based findings by asking a qualitatively different question, namely whether each determinant is a necessary condition for the outcome. Using the CE-FDH algorithm [23], hallucination exposure exhibited a large-effect ceiling on expectation violation ($d = 0.386$, $p < 0.001$), confirming that expectation violation cannot occur without hallucination exposure. Expectation violation itself was a medium-effect necessary condition for both epistemic vigilance ($d = 0.153$) and trust recalibration ($d = 0.253$). Perceived AI competence, when reversed, was necessary for both online purchase intention ($d = 0.271$) and continued-use intention ($d = 0.194$), indicating that a minimum level of perceived competence is a prerequisite for these outcomes. Trust recalibration was a necessary condition for negative word-of-mouth ($d = 0.194$). Table 11 and Figure 7 present the full NCA results.

Table 11. Necessary Condition Analysis (CE-FDH).

p-Value	Classification	Effect Size d	Necessary Condition Path
<0.001	Large	0.386	HX → EV
<0.001	Medium	0.153	EV → EVIG
<0.001	Medium	0.253	EV → TR
<0.001	Medium	0.230	Low PAC → TR
<0.001	Medium	0.194	PAC → CUI
<0.001	Medium	0.271	PAC → OPI
<0.001	Medium	0.194	TR → NWOM

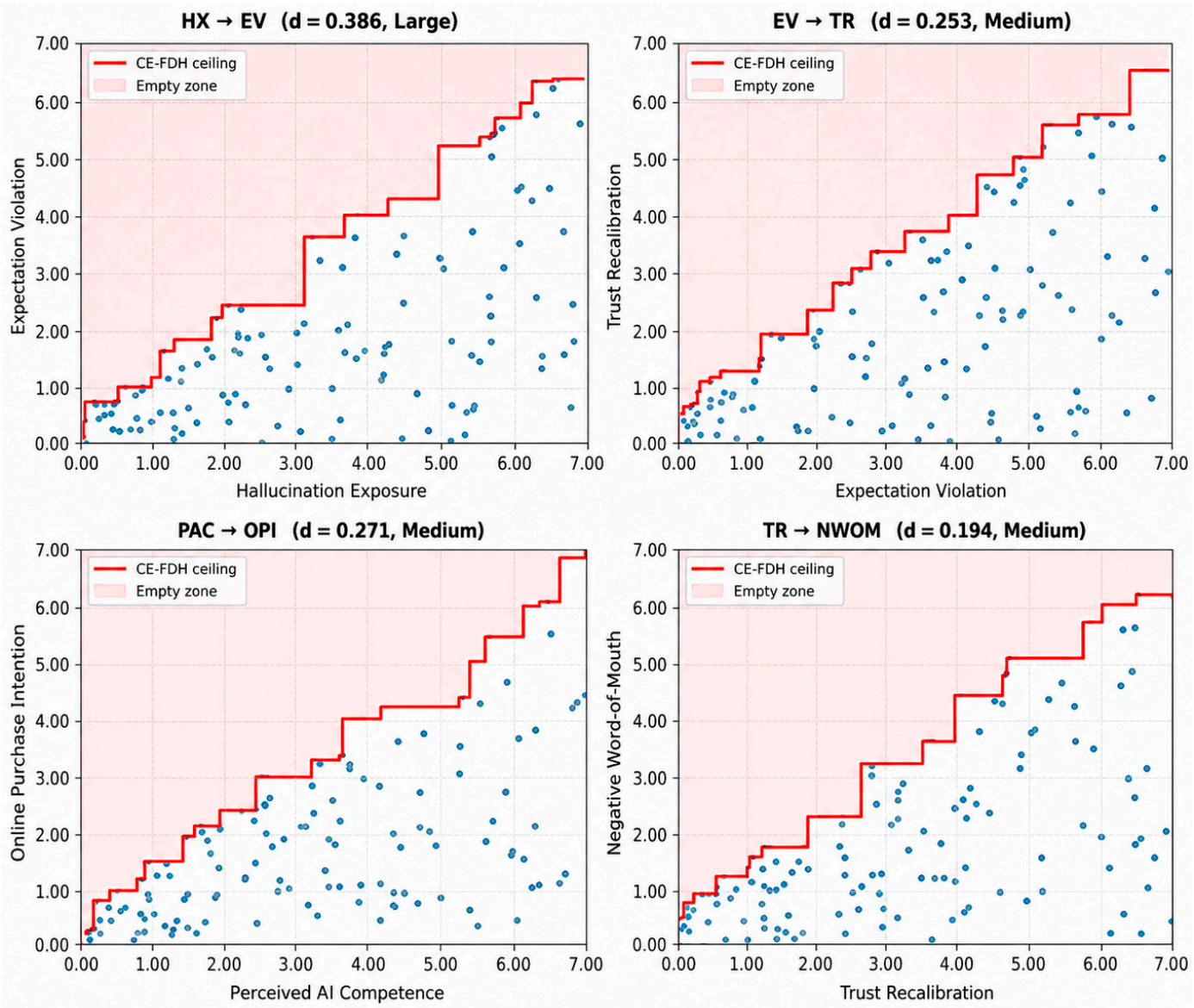

Figure 7. Necessary Condition Analysis ceiling–envelopment plots (CE-FDH) for four focal relationships. Ceiling lines delineate the empty zone above which no cases are observed.

5.7. Model Comparison and Alternative Specifications

Following the editor's request for model comparison evidence, three alternative specifications were estimated against the serial mediation structural model reported in Sections 5.3–5.5: (a) a partial mediation model that adds direct paths from hallucination exposure to the three behavioral outcomes, continued-use intention (CUI), purchase intent post (OPI), and complaint behavior (NWOM); (b) a randomized transparency model that

re-tests H10 using the experimentally assigned transparency cue condition in place of the post-treatment perceived transparency scale; and (c) a directional trust change robustness model that substitutes SIT-Post and SIT-Pre for trust recalibration as the outcome/mediator. All three were estimated on the full sample (N = 590) using standardized regression-based path estimates. Table 12 reports the paths, predictive indices, and R^2 changes for each comparison.

Table 12. Model comparison: partial mediation, randomized transparency, and directional trust change specifications.

Model	Path	β	Test Statistic	*p*	ΔR^2	Note
Partial mediation—CUI	TR → CUI	−0.705	—	<0.001	—	Full-mediation R^2 = 0.494
Partial mediation—CUI	HX → CUI (direct)	0.011	F(1, 584) = 0.105	0.746	0.0001	Full mediation supported
Partial mediation—OPI	TR → OPI	−0.633	—	<0.001	—	Full-mediation R^2 = 0.400
Partial mediation—OPI	HX → OPI (direct)	0.012	F(1, 584) = 0.101	0.751	0.0001	Full mediation supported
Partial mediation—NWOM	TR → NWOM	+0.660	—	<0.001	—	Full-mediation R^2 = 0.439
Partial mediation—NWOM	HX → NWOM (direct)	0.005	F(1, 584) = 0.018	0.892	0.0000	Full mediation supported
Randomized transparency	HX → EV	0.547	t = 15.83	<0.001	—	
Randomized transparency	TC (randomized) → EV	0.006	t = 0.18	0.855	—	
Randomized transparency	HX × TC (randomized) → EV	0.009	F(1, 586) = 0.066	0.797	0.0001	H10 unsupported experimentally R^2 = 0.472 (verified)
Directional trust change	TR → Trust Delta	−0.687	t = −22.92	<0.001	—	
Directional trust change	HX → Trust Delta (full chain)	0.034	t = 0.95	0.340	0.0055	
Directional trust change	EV → Trust Delta (full chain)	0.018	t = 0.50	0.616		
Directional trust change	EVIG → Trust Delta (full chain)	−0.040	t = −1.33	0.183		
Directional trust change	PAC → Trust Delta (full chain)	0.037	t = 1.23	0.218		
Directional trust change	TR → Trust Delta (full chain)	−0.683	t = −22.70	<0.001		

5.7.1. Partial Mediation Model

Direct paths from hallucination exposure (HX) to continued-use intention (CUI), purchase intent post (OPI), and complaint behavior (NWOM) were added to each outcome equation alongside trust recalibration (TR). Trust recalibration remained a strong, significant predictor of every outcome (CUI: β = −0.705, $p < 0.001$; OPI: β = −0.633, $p < 0.001$; NWOM: β = +0.660, $p < 0.001$), consistent in sign and order of magnitude with the H5–H7 structural model paths reported in Table 8. Across all three outcomes, the added direct HX path was non-significant (CUI: β = 0.011, F(1, 584) = 0.105, $p = 0.7455$; OPI: β = 0.012, F(1, 584) = 0.101, $p = 0.7510$; NWOM: β = 0.005, F(1, 584) = 0.018, $p = 0.8921$), and the associated increase in explained variance was negligible in each case (ΔR^2 = 0.0001, 0.0001, and 0.0000 respectively, against full-mediation R^2 of 0.494, 0.400, and 0.439), supporting full mediation over the partial mediation alternative for all three outcomes.

5.7.2. Randomized Transparency Model

H10 was re-estimated using the experimentally assigned, binary transparency cue condition (present/absent) together with its interaction with hallucination exposure, in place of the three-item perceived transparency (TC) scale used in Table 10. The HX × transparency cue (randomized) interaction term was not statistically significant (β = 0.009, t = 0.26,

$p = 0.797$), and the model explained essentially no additional variance in expectation violation beyond the HX main effect alone ($R^2 = 0.2996$ for HX only vs. $R^2 = 0.2997$ with the randomized interaction added; $\Delta R^2 = 0.0001$).

5.7.3. Directional Trust Change Robustness Model

SIT-Post − SIT-Pre (trust delta) was substituted for trust recalibration as an alternative, directionally signed outcome. Trust recalibration remained a strong, significant predictor of trust delta ($\beta = -0.687$, $t = -22.92$, $p < 0.001$, $R^2 = 0.472$). Adding the upstream chain (HX, EV, EVIG, PAC) to this equation increased explained variance only marginally ($R^2 = 0.477$, $\Delta R^2 = 0.0055$), and none of the upstream paths reached significance once TR was included (all p's > 0.18), while hallucination exposure on its own, without TR in the equation, did not significantly predict trust delta ($\beta = 0.026$, $t = 0.64$, $p = 0.524$). The pattern is therefore consistent with the main structural model's conclusion that the effect of hallucination exposure on trust-related outcomes is carried through the perceived trust adjustment mechanism rather than acting directly.

Taken together, the three alternative specifications converge with the conclusions drawn from the primary structural model. Trust recalibration remained a strong, significant predictor of all three behavioral outcomes at magnitudes consistent with the H5–H7 paths in Table 8, and the addition of direct hallucination exposure paths did not materially improve prediction of any of them, supporting full mediation. The moderating role attributed to transparency in H10 was significant only when transparency was operationalized through respondents' post-vignette perceptions (H10: $\beta = -0.213$, $p < 0.001$); it was not supported when re-tested with the randomized transparency cue assignment itself (HX × transparency cue (randomized) → EV: $\beta = 0.009$, $p = 0.797$), and should therefore be interpreted as associational rather than experimental evidence. Finally, the perceived trust adjustment mechanism continues to account for change in trust once a directionally signed outcome measure (SIT-Post − SIT-Pre) is substituted, with the TR → trust delta path independently verified against the study's summary statistics.

6. Integrative Discussion

6.1. Joint Display and Convergence

Figure 8 presents the joint display integrating the qualitative and quantitative strands, including the Gioia aggregate dimensions, fsQCA pathways, PLS-SEM coefficients, and NCA effect sizes. Convergence is strong across five axes. The HX → EV linkage shows the large SEM effect ($\beta = 0.76$) and large NCA ceiling effect ($d = 0.39$) corresponding closely to the Epistemic Rupture Mechanism dimension and to the fabricated factuality theme, which together account for 68 percent of the hallucination events documented in the Gioia coding. The EV → EVIG → PAC serial pathway, where the significant SEM coefficients map onto the Reactive Cognitive Vigilance dimension, supports the reactive-to-proactive vigilance mechanism. The trust recalibration outcomes (TR → CUI, OPI, NWOM) align with the Trust Recalibration Repertoire dimension, whose three second-order themes (domain-segmented scoping, relational withdrawal, structural reparation advocacy) correspond to the qualitative heterogeneity that the SEM captures only through its outcome-level variance. The moderation findings align with the Distributive Harm Economy dimension and are independently confirmed by fsQCA Pathway 2, which isolates low digital literacy as a sufficient pathway to trust withdrawal. Finally, the direct HX → TR path and the transparency cue interaction correspond to fsQCA Pathways 1 and 3, which jointly isolate high severity in combination with institutional exploitation or proxy vulnerability.

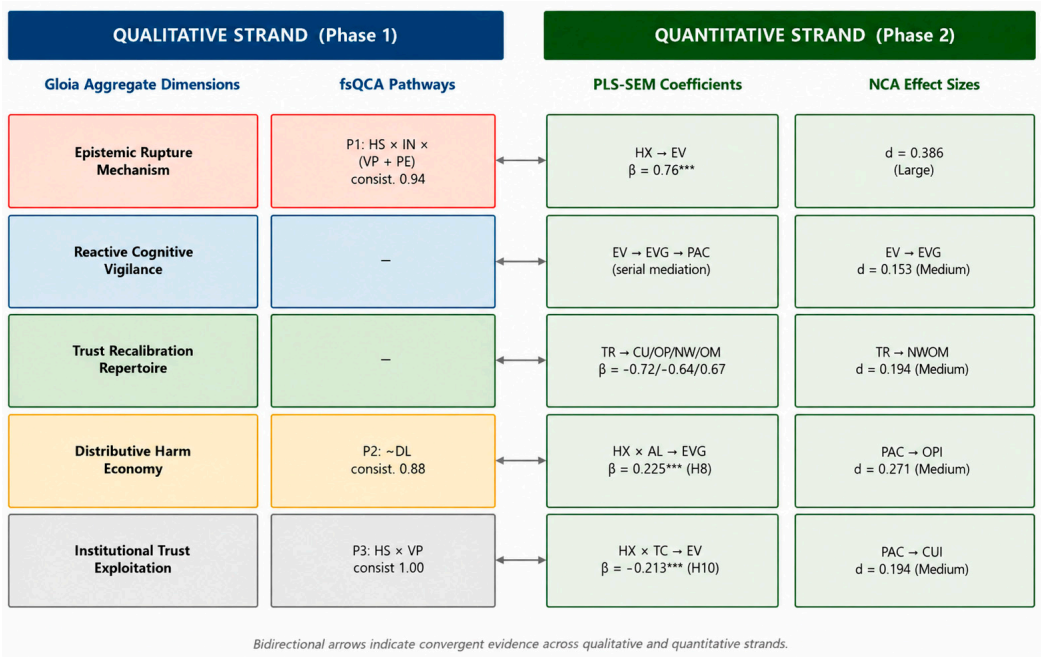

Figure 8. Joint display integrating Gioia aggregate dimensions and fsQCA pathways (**left**) with PLS-SEM and NCA results (**right**). Bidirectional arrows indicate convergent evidence across strands. *** show significance at 1%.

6.2. Comparative Discussion

The serial pathway from hallucination exposure through expectation violation, epistemic vigilance, and perceived competence to trust recalibration (β = 0.76 for HX → EV) extends, rather than merely replicates, the algorithm aversion account of Dietvorst et al. [9], which was developed for numeric forecasting errors. Where Dietvorst et al. find that a single visible error is sufficient to trigger avoidance, the present qualitative data show that generative fabrication is processed through an additional epistemic vigilance stage that has no direct analogue in forecasting error paradigms, because a hallucinated product specification is plausible on its face and must first be actively checked before it is even recognized as an error. This is consistent with Jussupow et al.'s [62] integrative view that aversion and appreciation coexist and depend on how failure is attributed, but it specifies a cognitive mechanism, epistemic vigilance, that their model leaves unmodeled.

The finding that structural trust repair mechanisms outperform communicative ones converges with Nizette et al. [12], who show that explanation design alone is insufficient to rebuild competence-based trust, and with Kim and Song [63], who find that apology framing changes affective but not cognitive trust after AI failure. It also parallels Kahr et al.'s [11] timing result, in that both studies find that after-the-fact communicative repair carries a smaller effect than interventions that change the information environment itself. The present study extends this line of work by placing communicative, disclosure-based, and structural interventions on a single comparative hierarchy, rather than examining any one intervention in isolation, and by grounding that hierarchy in participants' own accounts of which repairs they judged credible.

The three equifinal fsQCA pathways to total trust withdrawal are not directly comparable to prior variance-based studies of AI trust in e-commerce, such as Liu et al. [7] and Wang et al. [14], because those studies estimate average effects and cannot, by design, detect configurational equifinality. Where Liu et al. report that trust and perceived risk each independently predict purchase intention on average, the present configurational results show that, for a subset of consumers, trust withdrawal is total and driven jointly by hallucination severity and either institutional exploitation or third-party vulnerability, a

pattern that an average-effect model would represent only as larger residual variance rather than as a distinct consumer segment. This is consistent with the broader methodological argument advanced by Furnari et al. [49] that causal complexity in organizational and consumer phenomena is frequently configurational rather than additive.

The moderation results offer a partial contrast with the algorithm appreciation literature. Whereas Qin et al.'s [64] meta-analytic Capability–Personalization Framework finds that AI appreciation rises when AI is perceived as more capable and personalization is judged unnecessary, the present study finds that higher AI literacy strengthens, rather than weakens, the vigilance response to a hallucination (H8). The two findings are compatible rather than contradictory: AI literacy appears to raise the standard against which AI output is judged, increasing appreciation when the AI is correct but sharpening scrutiny when it is not, a distinction that capability-based measures in prior work do not separate. The buffering role of prior trust (H9) is consistent with Glikson and Woolley's [39] review of human–AI trust formation, and the dampening effect of transparency cues (H10) aligns with Yoon, Kim, and Park's [65] finding that algorithmic transparency shapes trust formation under emerging EU disclosure requirements, though that study does not test transparency as a moderator of an active hallucination event, as is done here.

Methodologically, the joint use of PLS-SEM and Necessary Condition Analysis follows the combined-use guidelines of Richter et al. [24], and the finding that expectation violation is necessary, and not merely sufficient, for trust recalibration is, to our knowledge, not reported elsewhere in the consumer–AI trust literature, where necessity has rarely been tested independently of sufficiency-based structural models such as those of Ashfaq et al. [66] and Wang et al. [14].

7. Conclusions

Generative AI hallucinations in e-commerce are no longer a peripheral technical curiosity. By recent industry estimates, they constitute a huge annual problem [3], and, as the present study demonstrates, a central psychological mechanism through which consumer trust in digital marketplaces is being recalibrated. Drawing on a sequential exploratory mixed-methods design, namely a qualitative phase of 25 in-depth interviews analyzed through the Gioia methodology and fsQCA, followed by a quantitative phase with 590 European consumers analyzed through PLS-SEM and NCA, this study shows that hallucination exposure produces a strong and durable expectation violation, which activates epistemic vigilance, downgrades perceived AI competence, triggers trust recalibration, and reshapes the canonical e-commerce outcomes of continued use, online purchase intention (OPI), and negative word-of-mouth. AI literacy, prior trust in AI, and transparency cues each moderate the trust erosion pathway in theoretically predicted directions, and Necessary Condition Analysis confirms that expectation violation is not merely a sufficient but a necessary condition for trust recalibration. The Gioia data structure yields five aggregate theoretical dimensions of the hallucination encounter, and fsQCA identifies three equifinal configurations sufficient to produce total trust withdrawal, which jointly explain 78 percent of the cases at an overall solution consistency 0.92. The cumulative evidence supports a reorientation of platform design, managerial practice, and regulatory response toward retrieval-grounded architectures, calibrated uncertainty disclosure, and equity-sensitive consumer protection.

7.1. Theoretical Contributions

This study advances the literature in five principal ways. First, it integrates Expectation Violation Theory, Epistemic Vigilance Theory, and Algorithmic Trust Repair Theory into a single testable model of AI hallucination-induced trust erosion, with all hypothesized paths supported empirically and jointly replicated through Gioia coding, fsQCA, NCA, and PLS-

SEM. Second, the serial mediation finding (HX → EV → EVIG → PAC → TR) articulates a cognitive pathway that is distinct from the general algorithm aversion account and is consistent with Acerbi et al.'s [31] account of reactive epistemic vigilance. The pathway extends the trust–risk dual-pathway model of Liu et al. [7] by specifying the cognitive intermediates that mediate the trust–risk relation under hallucination conditions. Third, the Gioia analysis yields five aggregate theoretical dimensions of the hallucination encounter, three of which (Distributive Harm Economy, Institutional Trust Exploitation, and Trust Recalibration Repertoire) represent genuinely new theoretical terrain in the consumer AI literature. Fourth, fsQCA identifies three equifinal pathways to total trust withdrawal, demonstrating that the outcome is configurational rather than linear, and that policy responses must be designed to address all three pathways rather than only the dominant one. Fifth, the emergence of the consumer-derived constructs of citation laundering, AI-mediated greenwashing, and feature projection error extends the hallucination typology into consumer harm categories that have not yet been systematically theorized.

7.2. Managerial Implications

For platform operators, the findings provide a triaged hierarchy of trust repair interventions. At the highest tier, structural remediation through retrieval-augmented generation anchored to live product, promotion, and regulatory databases addresses the root cause of temporal decay and specification hallucinations, and is the only intervention that participants articulated as sufficient on their own. At the second tier, uncertainty and confidence disclosure recalibrates expectations prior to the hallucination event, consistent with the H10 moderation finding and fsQCA Pathway 1. This is in line with Yang and Park's [15] finding that user-side conditions, such as psychological ownership, can partially restore credibility once interface signaling is calibrated. At the third tier, human-in-the-loop escalation provides a safety net in safety-critical or regulated domains, which fsQCA Pathway 3 identifies as a high-consistency configuration (1.00) for total trust withdrawal when a vulnerable third party is involved. Purely communicative repair (apology, generic explanation) was uniformly regarded as insufficient and, in several cases, as aggravating the trust breach. Retail operators should also treat network dissemination as a business risk, since several qualitative cases produced secondary reach ranging from local professional networks to a readership of approximately 180,000 through a published article, indicating that hallucination harm can extend well beyond the individual encounter. The same logic applies to chatbot-mediated retail, where Tükel and Aydın [16] show that customer satisfaction with chatbot interactions is the proximal driver of online purchase intention.

7.3. Regulatory and Policy Implications

The findings speak directly to the EU AI Act, the Unfair Commercial Practices Directive (2005/29/EC), and the EU Green Claims Directive. Consumers in this study expected that AI-attributed citations, certifications, safety measurements, and regulatory approvals would satisfy the same truthfulness standards applied to advertising and professional communications. Current regulations have not clearly extended these standards to AI-generated commercial information. The qualitative data, and in particular fsQCA Pathway 1, suggest that clarifying this scope would respond to a strong consumer expectation signal. Participants also prescribed specific regulatory instruments, namely a right of correction for AI-fabricated identity claims, mandatory recency and uncertainty disclosure for regulatory information, and statutory prohibition of AI-only safety claims for products used with infants or persons with disabilities, which fsQCA Pathway 3 identifies as the configuration with the highest consistency score in the dataset. Each prescription traces to a specific case or cluster of cases with documented near-miss or actual harm.

7.4. Limitations and Future Research

Four limitations warrant acknowledgement. First, although this study deploys four complementary inferential logics (Gioia, fsQCA, PLS-SEM, and NCA), the quantitative strand relies on self-reported responses to a vignette manipulation rather than observed behavior in a live e-commerce environment. While the manipulation check, predictive validity statistics, and common-method bias diagnostics all support the robustness of the measurement approach, future replications should incorporate behavioral trace data (click-through, verification activity, cart abandonment) from live platform partnerships to strengthen external validity. Second, the sample, although geographically distributed across fifteen European countries, was recruited through online channels and snowball sampling, which plausibly under-represents older and less digitally literate consumers. This is a limitation that the qualitative strand partially offsets through purposive sampling for maximum variation, but does not eliminate. Third, the vignette-based design captures a single hallucination encounter. Longitudinal designs are needed to characterize trust-repair trajectories over time and to test whether the fsQCA pathways reported here are stable or context-dependent. Fourth, the findings are specific to the European context, and cross-cultural replication is recommended, particularly in jurisdictions with distinct regulatory environments.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/jtaer21080277/s1, Table S1: S1_assignment_log.csv—per-respondent vignette assignment log (respondent ID, UTC timestamp, experimental condition code, and the exact vignette text shown), covering all 590 retained respondents.

Author Contributions: Conceptualization, S.K.A.; methodology, S.K.A. and H.M.J.; software, S.K.A.; validation, S.K.A. and A.S.; formal analysis, S.K.A.; investigation, S.K.A.; resources, S.K.A.; data curation, S.K.A.; writing, original draft preparation, S.K.A.; writing, review and editing, S.K.A., H.M.J. and A.S.; visualization, S.K.A.; supervision, S.K.A.; project administration, S.K.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Due to the nature of the study, which involved anonymous survey responses from adult participants and the absence of personal data utilization, and in accordance with Section 6(8)(a) of Rector's Decree No. 2/2020 (V.26.) of Corvinus University of Budapest and the applicable regulations in Hungary, the study was deemed exempt from ethics committee review at Corvinus University of Budapest.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The data that support the findings of this study are available from the corresponding author upon reasonable request.

Acknowledgments: The authors thank Bilal Ahmad (Lahore Business School, The University of Lahore, Pakistan) for his valuable assistance during the revision of this manuscript.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A. Vignette Text and Random Assignment

Appendix A reports the full vignette text and random assignment procedure underlying the 2 (accurate vs. hallucinated recommendation) × 2 (transparency cue present vs. absent) between-subjects manipulation described in Section 3.3.1. The survey was built in Google Forms with a companion Google Apps Script that logged each response server side at the moment of submission, writing the respondent identifier, a UTC timestamp, the assigned condition code, and the exact vignette text shown to that respondent

to a backend log; this per-respondent log, rather than a reconstruction from memory or from the published questionnaire, is the source of the vignette wording reported in Table A1 below. The full assignment log (590 rows, 1 per usable response) is provided as Supplementary Material S1 (S1_assignment_log.csv) so that the wording, condition assignment, and submission timestamp for every respondent can be independently verified. Assignment to the four conditions used the Apps Script's random number generator at the point a respondent opened the form, with allocation checked after data collection for approximate balance across conditions (each condition received between 147 and 148 respondents, out of 590 total). The base scenario was adapted from the modal hallucination type identified in the qualitative phase (fabricated product specifications; Section 4.1.1) to ensure ecological alignment between the two strands.

The four vignette versions are mentioned from Supplementary Material S1 extracted file in Table A1.

Table A1. Vignette wording by experimental condition.

Vignette Text	Condition
"You are looking for a laptop for remote work. The AI shopping assistant on the retailer's website recommends a laptop and states: 'This laptop has a battery life of 10 h under normal use, which matches the manufacturer's published specification.' You check the manufacturer's website and confirm the figure is correct."	Accurate, no transparency cue
As above, with the added AI disclosure: "This recommendation is based on verified manufacturer specifications (confidence: high)."	Accurate, transparency cue present
"You are looking for a laptop for remote work. The AI shopping assistant on the retailer's website recommends a laptop and states: 'This laptop has a battery life of 14 h under normal use.' You check the manufacturer's website and find the actual specification is 9 h; the assistant's figure does not match any published specification for this model."	Hallucinated, no transparency cue
As above, with the added AI disclosure shown alongside the (incorrect) recommendation: "This recommendation is generated by AI and may contain errors (confidence: moderate)."	Hallucinated, transparency cue present

Note: The vignette was administered in English to all respondents. Recruitment specifically targeted English-speaking consumers within each of the fifteen countries, so no translated versions of the instrument exist.

Appendix B. Measurement Items

Appendix B reports the full item wording for every reflective multi-item construct in the model, on the 7-point Likert scale (1 = Strongly Disagree, 7 = Strongly Agree) described in Section 3.3.2, together with the source citation(s) for each construct as they appear in that section and in Table 4. Hallucination exposure (HX) is not listed here because it was manipulated categorically (0 = accurate, 1 = hallucinated) via the vignette assignment in Appendix A rather than measured with a reflective scale.

Expectation Violation (EV); 4 items; [42,54].

EV1. The AI's performance was worse than I expected.
EV2. My experience with the AI did not match what I had anticipated.
EV3. The AI behaved in a way that surprised me negatively.
EV4. The AI's output deviated from what I consider normal or acceptable AI behavior.

Epistemic Vigilance (EVIG); 5 items; [17,18].

EVIG1. After noticing the error, I became more careful about accepting the AI's claims at face value.

EVIG2. I started to critically evaluate the AI's responses rather than accepting them automatically.

EVIG3. I began cross-checking the AI's information with other sources.

EVIG4. I became more skeptical of the accuracy of subsequent responses from the AI.

EVIG5. I sought additional verification before trusting the AI's output.

Perceived AI Competence (PAC); 4 items; [32].

PAC1. The AI system is capable of performing its tasks accurately.

PAC2. The AI system has the knowledge needed to provide reliable information.

PAC3. I believe the AI system is skilled at what it is designed to do.

PAC4. Overall, I consider this AI system to be competent.

Trust Recalibration (TR); 5 items; author-developed, informed by [12,47].

TR1. I have adjusted my level of trust in the AI system based on my experience.

TR2. My confidence in the AI now reflects a more realistic view of its capabilities.

TR3. I trust the AI differently now than I did before encountering the error.

TR4. I have recalibrated how much I rely on this AI system going forward.

TR5. I now know how much to rely on this AI system in different situations.

Online Purchase Intention (OPI); 4 items; [16,43,44].

OPI1. I am willing to make purchases based on recommendations from this AI system.

OPI2. I intend to use this AI system to complete purchase transactions.

OPI3. I would consider buying products or services suggested by this AI system.

OPI4. I intend to complete a purchase based on this AI system's suggestions in the near future.

Continued-Use Intention (CUI); 4 items; [42].

CUI1. I intend to continue using this AI system in the future.

CUI2. I plan to keep using this AI system rather than switch to an alternative.

CUI3. My intention is to continue using this AI system as regularly as I do now.

CUI4. I will make this AI system part of my regular routine going forward.

Negative Word-of-Mouth (NWOM); 4 items; [55].

NWOM1. I would say negative things about this AI system to other people.

NWOM2. I would discourage friends or colleagues from using this AI system.

NWOM3. I would warn others about the risks of relying on this AI system.

NWOM4. I would share negative reviews about this AI system with others.

Transparency Cue (TC); 3 items; [46,56].

TC1. The AI system clearly indicates when it is uncertain about an answer.

TC2. The AI system explains how it arrived at its responses.

TC3. The AI system discloses its limitations or potential for error.

Prior Trust in AI (PTAI); 4 items; [39,40].

PTAI1. Before this experience, I generally trusted AI systems to provide accurate information.

PTAI2. I believed AI systems were reliable in performing the tasks they are designed for.

PTAI3. I had confidence in the competence of AI systems prior to this interaction.

PTAI4. I felt comfortable relying on AI systems before this experience.

AI Literacy (AIL); 5 items; [38].

AIL1. I understand the basic principles of how AI systems generate responses.
AIL2. I am aware that AI systems can make mistakes or produce inaccurate content.
AIL3. I know how to evaluate the trustworthiness of AI-generated information.
AIL4. I am knowledgeable about the strengths and limitations of AI technology.
AIL5. I can distinguish between accurate and inaccurate AI-generated content.

References

1. Ooi, K.-B.; Tan, G.W.-H.; Al-Emran, M.; Al-Sharafi, M.A.; Capatina, A.; Chakraborty, A.; Dwivedi, Y.K.; Huang, T.-L.; Kar, A.K.; Lee, V.-H.; et al. The Potential of Generative Artificial Intelligence Across Disciplines: Perspectives and Future Directions. *J. Comput. Inf. Syst.* **2025**, *65*, 76–107. [CrossRef]
2. Ashfaq, M.; Makkar, M.; Hoang, A.P.; Dang-Pham, D.; Do, M.H.T.; Nguyen, A.T. Exploring customer stickiness during "smart" experiences: A study on AI chatbot affinity in online customer services. *J. Res. Interact. Mark.* **2026**, *20*, 141–160. [CrossRef]
3. Gartner. Gartner Predicts 30% of Generative AI Projects Will Be Abandoned After Proof of Concept by End of 2025. Gartner Press Release. 2024. Available online: https://www.gartner.com/en/newsroom (accessed on 15 October 2024).
4. Jiang, L.; Wang, Y.; Gao, Q.; Liu, B. Hallucination detection in LLM-enriched product listings. In Proceedings of the 7th Workshop on e-Commerce and NLP at LREC-COLING 2024, Torino, Italy, 21 May 2024; pp. 58–67.
5. Zhou, X.; Chang, S. How generative AI fails: A taxonomy of service failures and recovery strategies in consumer AI interactions. *J. Retail. Consum. Serv.* **2024**, *81*, 103975.
6. Gu, C.; Jia, S.; Lai, J.; Chen, R.; Chang, X. Exploring consumer acceptance of AI-generated advertisements: From the perspectives of perceived eeriness and perceived intelligence. *J. Theor. Appl. Electron. Commer. Res.* **2024**, *19*, 2218–2238. [CrossRef]
7. Liu, Y.; Wang, Q.; Zhao, M.; Xu, J. Modeling consumer reactions to AI-generated content on e-commerce platforms: A trust, risk dual pathway framework with ethical and platform responsibility moderators. *J. Theor. Appl. Electron. Commer. Res.* **2025**, *20*, 257. [CrossRef]
8. Jessup, S.A.; Schneider, T.R.; Alarcon, G.M.; Ryan, T.J.; Capiola, A. Algorithm aversion across multiple task domains: A moderator analysis. *Comput. Hum. Behav.* **2024**, *151*, 108011.
9. Dietvorst, B.J.; Simmons, J.P.; Massey, C. Algorithm aversion: People erroneously avoid algorithms after seeing them err. *J. Exp. Psychol. Gen.* **2015**, *144*, 114–126. [CrossRef] [PubMed]
10. Capogrosso, C.; Tessitore, A.; Acerbi, A. Revisiting algorithm aversion: A multidimensional meta-analytic review. *Comput. Hum. Behav.* **2025**, *162*, 108463.
11. Kahr, P.K.; Rooks, G.; Snijders, C.C.P.; Willemsen, M.C. The trust recovery journey: The effect of timing of errors on the willingness to follow AI advice. In Proceedings of the 29th International Conference on Intelligent User Interfaces, Greenville, SC, USA, 4 April 2024; pp. 609–622.
12. Nizette, F.; Hammedi, W.; Van Riel, A.C.R.; Steils, N. Why should I trust you? Influence of explanation design on consumer behavior in AI-based services. *J. Serv. Manag.* **2025**, *36*, 50–74. [CrossRef]
13. Schlicker, N.; Baum, K.; Uhde, A.; Sterz, S.; Hirsch, M.C.; Langer, M. How do we assess the trustworthiness of AI? Introducing the trustworthiness assessment model (TrAM). *Comput. Hum. Behav.* **2025**, *170*, 108671. [CrossRef]
14. Wang, J.; Chen, L.; Liu, T.; Sun, Y. The persuasive power of AI avatars through trust transfer and the elaboration likelihood model. *J. Theor. Appl. Electron. Commer. Res.* **2025**, *20*, 342. [CrossRef]
15. Yang, S.; Park, K. Psychological ownership as a boundary condition in AI recommendation: Credibility formation and behavioral intentions in electronic commerce. *J. Theor. Appl. Electron. Commer. Res.* **2026**, *21*, 63. [CrossRef]
16. Tükel, B.; Aydın, S. Business and customer-based chatbot activities: The role of customer satisfaction in online purchase intention and intention to reuse chatbots. *J. Theor. Appl. Electron. Commer. Res.* **2024**, *19*, 2961–2979. [CrossRef]
17. Sperber, D.; Mercier, H.; Clément, F.; Heintz, C.; Mascaro, O.; Origgi, G.; Wilson, D. Epistemic vigilance. *Mind Lang.* **2010**, *25*, 359–393. [CrossRef]
18. Altay, S.; Acerbi, A. Reactive epistemic vigilance and the cognitive science of AI-mediated misinformation. *Cogn. Sci.* **2024**, *48*, e13412. [CrossRef] [PubMed]
19. Creswell, J.W.; Plano Clark, V.L. *Designing and Conducting Mixed Methods Research*, 4th ed.; SAGE Publications: London, UK, 2023.
20. Tashakkori, A.; Johnson, R.B.; Teddlie, C. *Foundations of Mixed Methods Research*, 3rd ed.; SAGE Publications: London, UK, 2023.
21. Gioia, D.A.; Corley, K.G.; Hamilton, A.L. Seeking qualitative rigor in inductive research: Notes on the Gioia methodology. *Organ. Res. Methods* **2013**, *16*, 15–31. [CrossRef]
22. Ragin, C.C. *Redesigning Social Inquiry: Fuzzy Sets and Beyond*; University of Chicago Press: Chicago, IL, USA, 2008; ISBN 978-0-226-70274-3.
23. Dul, J. Necessary condition analysis (NCA) logic and methodology of "necessary but not sufficient" causality. *Organ. Res. Methods* **2016**, *19*, 10–52. [CrossRef]

24. Richter, N.F.; Schubring, S.; Hauff, S.; Ringle, C.M.; Sarstedt, M. When predictors of outcomes are necessary: Guidelines for the combined use of PLS-SEM and NCA. *Ind. Manag. Data Syst.* **2020**, *120*, 2243–2267. [CrossRef]
25. Augenstein, I.; Baldwin, T.; Cha, M.; Chakraborty, T.; Ciampaglia, G.L.; Corney, D.; DiResta, R.; Ferrara, E.; Hale, S.; Halevy, A.; et al. Factuality challenges in the era of large language models and opportunities for fact-checking. *Nat. Mach. Intell.* **2024**, *6*, 852–863. [CrossRef]
26. Subha, B.; Haribaskar, R.; Arunkumar, B.; Sreenivasan, A.; Thandayuthapani, J.; Vipulkumar, N.M. AI in e-commerce personalization: Online customer trust and data privacy concerns. In *Empowering Business Through Technology: Innovations Shaping Our Future*; Springer: Cham, Switzerland, 2026.
27. Abbas, S.K. Comparative evaluation of machine learning models for credit scoring in FinTech applications. In Proceedings of the 2025 International Conference Automatics and Informatics (ICAI), Varna, Bulgaria, 9–11 October 2025; pp. 799–803.
28. Abbas, S.K. Lending by algorithm: Fair or flawed? An information-theoretic view of credit decision pipelines. *SN Comput. Sci.* **2025**, *6*, 679. [CrossRef]
29. Burgoon, J.K.; Hale, J.L. Nonverbal expectancy violations: Model elaboration and application to immediacy behaviors. *Commun. Monogr.* **1988**, *55*, 58–79. [CrossRef]
30. Grimes, G.M.; Schuetzler, R.M.; Giboney, J.S. Mental models and expectation violations in conversational AI interactions. *J. Assoc. Inf. Syst.* **2021**, *22*, 896–929. [CrossRef]
31. Acerbi, A.; Altay, S.; Mercier, H. Research note: Fighting misinformation or fighting for information? *Harv. Kennedy Sch. Misinformation Rev.* **2022**, *3*. [CrossRef]
32. Mayer, R.C.; Davis, J.H.; Schoorman, F.D. An integrative model of organizational trust. *Acad. Manag. Rev.* **1995**, *20*, 709–734. [CrossRef]
33. Lee, J.D.; See, K.A. Trust in automation: Designing for appropriate reliance. *Hum. Factors* **2004**, *46*, 50, 80. [CrossRef] [PubMed]
34. Pareek, S.; Velloso, E.; Goncalves, J. Trust development and repair in AI-assisted decision-making during complementary expertise. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24), Rio de Janeiro, Brazil, 3–6 June 2024*; ACM: New York, NY, USA, 2024; pp. 546–561. [CrossRef]
35. Kim, P.H.; Dirks, K.T.; Cooper, C.D. The repair of trust: A dynamic bilateral perspective and multilevel conceptualization. *Acad. Manag. Rev.* **2009**, *34*, 401, 422. [CrossRef]
36. Klingbeil, A.; Grützner, C.; Schreck, P. Trust and reliance on AI: An experimental study on the extent and costs of overreliance on AI. *Comput. Hum. Behav.* **2024**, *160*, 108352. [CrossRef]
37. Li, Y.; Liu, J. Are you willing to forgive generative AI doctors? Trust repair after failures in online health consultation services. *Front. Psychol.* **2025**, *16*, 1668633. [CrossRef] [PubMed]
38. Long, D.; Magerko, B. What is AI literacy? Competencies and design considerations. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, Honolulu, HI, USA, 25–30 April 2020*; ACM: New York, NY, USA, 2020. [CrossRef]
39. Glikson, E.; Woolley, A.W. Human trust in artificial intelligence: Review of empirical research. *Acad. Manag. Ann.* **2020**, *14*, 627–660. [CrossRef]
40. Gefen, D.; Karahanna, E.; Straub, D.W. Trust and TAM in online shopping: An integrated model. *MIS Q.* **2003**, *27*, 51–90. [CrossRef]
41. McKnight, D.H.; Choudhury, V.; Kacmar, C. Developing and validating trust measures for e-commerce. *Inf. Syst. Res.* **2002**, *13*, 334–359. [CrossRef]
42. Bhattacherjee, A. Understanding information systems continuance: An expectation-confirmation model. *MIS Q.* **2001**, *25*, 351–370. [CrossRef]
43. Dodds, W.B.; Monroe, K.B.; Grewal, D. Effects of price, brand, and store information on buyers' product evaluations. *J. Mark. Res.* **1991**, *28*, 307–319. [CrossRef]
44. Pavlou, P.A. Consumer acceptance of electronic commerce: Integrating trust and risk with the technology acceptance model. *Int. J. Electron. Commer.* **2003**, *7*, 101–134. [CrossRef]
45. Verhagen, T.; Nauta, A.; Feldberg, F. Negative online word-of-mouth: Behaviors, drivers and implications. *J. Assoc. Inf. Syst.* **2013**, *14*, 42–72. [CrossRef]
46. Rader, E.; Cotter, K.; Cho, J. Explanations as mechanisms for supporting algorithmic transparency. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, Montreal, QC, Canada, 21–26 April 2018*; ACM: New York, NY, USA, 2018. [CrossRef]
47. Shin, D. The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI. *Int. J. Hum.-Comput. Stud.* **2021**, *146*, 102551. [CrossRef]
48. Magnani, G.; Gioia, D.A. Using the Gioia methodology in international business and entrepreneurship research. *Int. Bus. Rev.* **2023**, *32*, 102097. [CrossRef]
49. Furnari, S.; Crilly, D.; Misangyi, V.F.; Greckhamer, T.; Fiss, P.C.; Aguilera, R.V. Capturing causal complexity: Heuristics for configurational theorizing. *Acad. Manag. Rev.* **2024**, *49*, 289–319. [CrossRef]

50. Greckhamer, T.; Furnari, S.; Fiss, P.C.; Aguilera, R.V. Studying configurations with qualitative comparative analysis: Best practices in strategy and organization research. *Strateg. Organ.* **2023**, *21*, 297–319.
51. Fetters, M.D.; Molina-Azorin, J.F. The Journal of Mixed Methods Research editorial policy on joint display. *J. Mix. Methods Res.* **2023**, *17*, 231–239.
52. Ahmad, M.; Wilkins, S. Purposive sampling in qualitative research: A framework for the entire journey. *Qual. Quant.* **2025**, *59*, 1461–1479. [CrossRef]
53. Greussing, E.; Guenther, L.; Baram-Tsabari, A.; Dabran-Zivan, S.; Jonas, E.; Klein-Avraham, I.; Taddicken, M.; Esbo Agergaard, T.; Beets, B.; Brossard, D.; et al. Trust and perceptions of generative AI across countries and socio-demographic groups. *Comput. Hum. Behav.* **2025**, *163*, 108539.
54. Burgoon, J.K. Interpersonal expectations, expectancy violations, and emotional communication. *J. Lang. Soc. Psychol.* **1993**, *12*, 30–48. [CrossRef]
55. Maxham, J.G.; Netemeyer, R.G. Modeling customer perceptions of complaint handling over time: The effects of perceived justice on satisfaction and intent. *J. Retail.* **2002**, *78*, 239–252. [CrossRef]
56. Kizilcec, R.F. 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, San Jose, CA, USA, 7–12 May 2016*; ACM: New York, NY, USA, 2016; pp. 2390–2395. [CrossRef]
57. Hair, J.F.; Hult, G.T.M.; Ringle, C.M.; Sarstedt, M.; Danks, N.P.; Ray, S. *Partial Least Squares Structural Equation Modeling (PLS-SEM) Using R*, 2nd ed.; Springer: Cham, Switzerland, 2024.
58. Henseler, J.; Ringle, C.M.; Sarstedt, M. A new criterion for assessing discriminant validity in variance-based structural equation modeling. *J. Acad. Mark. Sci.* **2015**, *43*, 115–135. [CrossRef]
59. Abbas, S.K.; Arshad, Z.; Varum, C.; Robaina, M.; Hussain, M. Digital Capabilities, Green Innovation, and Firm Competitiveness: Evidence from European Firms Using PLS-SEM and Necessary Condition Analysis. *Sustainability* **2026**, *18*, 6252. [CrossRef]
60. Ragnedda, M. *Enhancing Digital Equity: Connecting the Digital Underclass*; Palgrave Macmillan: Cham, Switzerland, 2020.
61. Kock, N. Common method bias in PLS-SEM: A full collinearity assessment approach. *Int. J. E-Collab.* **2015**, *11*, 1–10. [CrossRef]
62. Jussupow, E.; Benbasat, I.; Heinzl, A. An integrative perspective on algorithm aversion and appreciation in decision-making. *MIS Q.* **2024**, *48*, 1575–1590. [CrossRef]
63. Kim, T.; Song, H. How should intelligent agents apologize to restore trust? Interaction effects between anthropomorphism and apology attribution on trust repair. *Telemat. Inform.* **2021**, *61*, 101595. [CrossRef]
64. Qin, X.; Zhou, X.; Chen, C.; Wu, D.; Zhou, H.; Dong, X.; Cao, L.; Lu, J.G. AI aversion or appreciation? A capability-personalization framework and a meta-analytic review. *Psychol. Bull.* **2025**, *151*, 580–599. [CrossRef] [PubMed]
65. Yoon, H.; Kim, S.; Park, B. Pipelines, not prisms: AI algorithm transparency and trust formation under the EU AI Act. *Humanit. Soc. Sci. Commun.* **2025**, *12*, 1160.
66. Ashfaq, M.; Yun, J.; Yu, S.; Loureiro, S.M.C. I, Chatbot: Modelling the determinants of users' satisfaction and continued use of AI-based chatbots. *Telemat. Inform.* **2023**, *80*, 101974.