Journal of
Theoretical and Applied Electronic Commerce Research

MDPI

Article

When Algorithms Speak Louder than Empathy: Mechanistic Interpretability as a Costly Authenticity Signal in AI-Mediated E-Commerce Customer Relationships

Yu Mi [1], Banggang Wu [1], Yong Wang [2] and Yuqi Du [3,*]

[1] School of Business Administration, Southwestern University of Finance and Economics, Chengdu 611130, China; miyu331@126.com (Y.M.); wubg@swufe.edu.cn (B.W.)
[2] College of Business and Public Management, West Chester University, West Chester, PA 19383, USA; ywang2@wcupa.edu
[3] International Chinese College, Panyapiwat Institute of Management, Nonthaburi 11120, Thailand
* Correspondence: dyq@mail.gufe.edu.cn

Abstract

Conversational AI agents are now a routine touchpoint in e-commerce customer service, and AI empathy has emerged as the headline humanization strategy for repairing relational damage during service failures. A growing evidence base reports that empathic AI often backfires, because consumers cannot reconcile felt warmth with their lay model of what an artificial agent is. This research asks under what conditions AI empathy can be made credible to consumers. We propose that mechanistic interpretability, operationalized in the present studies as a consumer-facing visualization of an AI agent's internal emotion-vector activations designed in the style of mechanistic-interpretability research, operates as a costly authenticity signal that rehabilitates empathic AI by enabling an attributional shift along the experience dimension of mind perception. Signaling Theory carries the antecedent stage of the causal chain, where mechanistic interpretability serves as a verifiable cue of computational authenticity. Mind Perception Theory carries the downstream stage, where the authenticated empathy is converted into consumer-brand intimacy. Two between-subjects experiments preceded by a feasibility pilot tested the account on Mainland Chinese consumers recruited via the Credamo online panel. Study 1 used a single-factor design contrasting high versus low AI empathy. Study 2 used a two (AI empathy) by two (mechanistic interpretability) full factorial. Study 1 showed a pattern consistent with high (versus low) AI empathy lowering brand intimacy through reduced perceived authenticity. Study 2 replicated the AI-empathy backfire when interpretability was absent, reversed the sign of the AI-empathy slope on the perceived-authenticity mediator when interpretability was present, and neutralized the negative conditional indirect effect on brand intimacy through perceived authenticity. The findings introduce mechanistic interpretability to consumer-marketing scholarship as a manipulable signaling channel, document a structural reversal in the mediator-stage slope coupled with neutralization of the indirect effect on the relational outcome, and prescribe pairing empathic AI phrasing with mechanistic-transparency design rather than deploying empathy without an accompanying transparency cue.

Keywords: AI empathy; mechanistic interpretability; perceived authenticity; brand intimacy; signaling theory; mind perception theory; e-commerce service recovery; moderated mediation

Check for updates

Academic Editor: Hyunchul Ahn

Received: 1 May 2026
Revised: 18 May 2026
Accepted: 26 May 2026
Published: 29 May 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

E-commerce platforms now route a sizable share of customer service through conversational AI agents trained to detect customer affect and reply with empathic phrasing. The strategic intuition is widely shared. If a chatbot can convey warmth, the relational damage typically caused by service failures, such as delayed shipments, missing items, or refund frictions, can be repaired through the same mechanism that human service representatives have always relied on [1,2]. The intuition has driven heavy practitioner investment in AI empathy as a humanizing strategy, and the construct has accumulated a substantial scholarly literature spanning marketing, information systems, and human–computer interaction [3,4].

A growing body of evidence reports that consumer responses to empathic AI are often the opposite of what designers intend: the harder a chatbot tries to sound empathic, the more readily consumers detect, discount, or actively reject the performance. Crolic et al. [5] found that anthropomorphic empathy in anger-laden encounters amplified consumer dissatisfaction; Juquelier et al. [6] reported that empathic chatbots could either enhance or undermine the customer experience depending on moderators that remained unspecified; and Liu et al. [7] named the phenomenon directly as an illusion of empathy, in which consumers register the empathic gesture but discount it because they cannot reconcile felt warmth with their model of what an AI is. The pattern that runs through this literature is recognition without belief, and Mind Perception Theory [8] supplies its cognitive logic. AI agents occupy a persistent position in lay mind perception as high on agency but low on experience, and AI empathy is precisely a request that consumers grant the AI the experience capacity their default mental model denies. Authenticity, defined as the consumer's attribution about whether an agent's behavior is a faithful expression of its character or a contrived performance [9,10], is the construct that takes the hit.

This raises the question that motivates the present research. Under what conditions can AI empathy be made credible to a consumer who has every reason to suspect it is a probabilistic illusion? The answer this paper develops draws on a literature, until recently confined to the technical AI-safety community, that has begun to expose the internal computations of large language models with a granularity unimaginable a few years ago. Mechanistic interpretability research now identifies named feature directions inside language models, including emotion-related vectors that fire when the model produces affective output [11]. Karny et al. [12] translated such activations into a Neural Transparency Interface that non-technical users can read in real time, and their user study reported significant trust gains when the panel was visible relative to a hidden-internals control.

We propose that mechanistic interpretability is theoretically positioned as a high-tier transparency cue in the precise Spencean sense [13,14], and that its joint operation with AI empathy shifts the indirect effect of AI empathy on consumer-brand relationships from reliably negative to non-negative through a reversal of the slope on the perceived-authenticity mediator. Without mechanistic interpretability, the consumer's mental model decodes empathic phrasing as a staged performance, perceived authenticity collapses, and the brand-relational outcome that AI empathy was deployed to repair is instead eroded. With mechanistic interpretability, the consumer sees the language model's internal emotion-vector activations rendered as a transparency panel. This exposure performs an attributional shift along the experience dimension of mind perception, the empathic display is re-attributed as a sincere algorithmic service rather than a forged feeling, the simple slope of AI empathy on perceived authenticity reverses sign, and the conditional indirect effect of AI empathy on brand intimacy is neutralized. The prediction is unusual. It is not an attenuation of a negative effect, the modal pattern in moderated-mediation findings on

AI design, but a structural slope reversal at the mediator stage coupled with neutralization of the negative indirect effect on the relational outcome.

We tested these predictions in two between-subjects experiments preceded by a feasibility pilot, all conducted on the Credamo online panel with Mainland Chinese adult consumers. Study 1 isolated the focal causal chain in a single-factor design and produced a pattern consistent with AI empathy reducing brand intimacy through reduced perceived authenticity (small-to-moderate effect). Study 2 extended the design to a two (AI empathy) by two (mechanistic interpretability) full factorial and tested the moderated-mediation prediction directly via the Index of Moderated Mediation. The Index of Moderated Mediation was reliably non-zero, the simple slope of AI empathy on perceived authenticity reversed sign across mechanistic interpretability, and the conditional indirect effect on brand intimacy under mechanistic interpretability present was no longer reliably different from zero, a neutralization rather than a reliable positive effect.

The central contribution is to show that a single design lever—visualizing the AI's internal emotional computation—reverses the AI-empathy slope on perceived authenticity and neutralizes the negative indirect effect on consumer-brand relationships. Doing so brings mechanistic-interpretability research, until now confined to AI safety, into consumer marketing as a manipulable signaling channel rather than a niche technical concept. The argument relies on Signaling Theory and Mind Perception Theory in tandem: the former explains why mechanistic transparency is a credible cue, the latter how that credibility converts into relational outcomes, and the neutralization-via-slope-reversal prediction sits at their intersection. The practical and policy stakes are immediate. Empathic AI customer-service phrasing becomes safer to deploy under conditions we identify, and the recent welfare critique of mandated explainable AI [15] gains a complementary reading in which firm-initiated mechanistic transparency operates as a voluntary signaling equilibrium that secures the welfare gains a mandate would risk eroding.

2. Theoretical Background

2.1. The Authenticity Paradox of AI Empathy

Conversational AI agents are a routine touchpoint in e-commerce service interactions, and a sizable share of practitioner attention has converged on AI empathy, defined as the agent's capacity to recognize, reciprocate, and respond to customers' affective states [2]. Empathic phrasing seems an obvious humanization lever—repairing, rather than aggravating, the relational damage caused by service failures [1,5]—and yet the accumulating evidence is polarized. Some studies report prosocial and satisfaction-enhancing effects of empathic AI [3,4], but a parallel stream documents the opposite: the harder a chatbot tries to sound empathic, the more readily consumers detect, discount, or actively reject the performance. Crolic et al. [5] found that anthropomorphic empathy in anger-laden interactions amplified dissatisfaction; Juquelier et al. [6] reported that empathic chatbots could either enhance or undermine the customer experience depending on as-yet-unspecified moderators; and Liu et al. [7] named the phenomenon directly as an illusion of empathy, in which consumers register the empathic gesture but discount it because they cannot reconcile felt warmth with their model of what an AI is.

Mind Perception Theory [8] provides the underlying logic for this credibility deficit. Lay observers, the theory shows, locate minds along two largely orthogonal dimensions. Agency captures the capacity for planning, intent, and self-control, and experience captures the capacity to feel pleasure, pain, and emotion. AI agents are persistently positioned in this space as high on agency but low on experience because feelings are intuitively biological [16,17]. Empathic AI strategies therefore presuppose the consumer mental model that the strategies themselves violate. They perform experience-laden empathy

in front of consumers whose default mind-perception priors deny the agent the very capacity that performance requires. Authenticity, defined as the consumer's attribution about whether an agent's behavior is a faithful expression of its character or a contrived performance [9,10], is the construct that takes the hit. When the consumer cannot reconcile observed empathy with the agent's perceived ontology, the empathy is decoded as staged, perceived authenticity collapses, and the relational outcomes that AI empathy was deployed to repair are eroded.

In an e-commerce service-failure context, where the customer is already affectively activated and has every incentive to scrutinize the agent's responses, AI empathy delivered without any mechanism to authenticate the agent's internal states should lower perceived authenticity and, through it, lower consumer-brand relationship outcomes. We operationalize the focal relational outcome as brand intimacy [18]. This expectation aligns with the closest direct empirical analogs [5–7] and with the cognitive route through which the credibility deficit propagates [19].

H1a. *In an e-commerce service-failure context, brand intimacy is lower when consumers interact with a high-AI-empathy agent than when they interact with a low-AI-empathy control agent.*

H1b. *The negative effect of AI empathy on brand intimacy is mediated by perceived authenticity. High (versus low) AI empathy reduces perceived authenticity, which in turn reduces brand intimacy.*

2.2. Mechanistic Interpretability as a Costly Authenticity Signal

If the failure mode of conventional empathic AI design is a credibility deficit rooted in attribution, then the corrective intervention must operate on the attributional pathway itself rather than on the empathic behavior in isolation. We argue that mechanistic interpretability—the visualization of an AI agent's internal computational states, including the activation of emotion-related feature directions inside the underlying language model—is uniquely suited to that role, and we ground the claim in two distinct bodies of literature.

The first is the rapid maturation of mechanistic interpretability as a research program in AI. Distinct from the post hoc explainable AI tradition, which constructs simplified explanations about a model after the fact, mechanistic interpretability reverse-engineers the computations performed inside the network and exposes them as objects of direct inspection [11]. Templeton et al.'s recent work on Anthropic's Claude Sonnet localized 171 emotion-related feature vectors in the model's internal activations. Karny et al. [12] operationalized this technical capacity into a Neural Transparency Interface that rendered such activations as a real-time side panel for non-technical users, and their user study reported significant trust gains when this panel was visible relative to a hidden-internals control.

The second motivating body of literature is signaling theory. Spence's [14] original formulation showed that informed parties could credibly communicate quality only through cues that were observable, costly to produce, and differentially costly such that low-quality types could not profitably mimic them. Connelly et al. [13] consolidated the theory's migration into management and marketing, and recent extensions have brought it into AI-mediated consumer settings [20–22]. Within this framework, transparency mechanisms can be ranked by signaling strength along a tier hierarchy. Disclosure, for example labeling an agent as AI, is cheap and easily mimicked [23]. Post hoc explainable AI, exemplified by feature importance scores or counterfactuals [24,25], is moderately costly to engineer but produces a separable artifact, namely, a story told about the model rather than the model's actual computation, and is therefore vulnerable to mimicry by surface-level wrappers. Mechanistic transparency, by contrast, requires that the model genuinely possess the internal feature structure being visualized, that the firm has invested in interpretability

infrastructure, and that the visualized activations correspond to real computational events inside the language model [11,12]. Faking such a panel is not merely expensive. It requires constructing the interpretability stack the panel claims to surface, at which point the panel is no longer fake. In Spencean terms, we therefore theorize three properties of mechanistic transparency that, taken together, position it at the highest tier of consumer-facing transparency signals: costliness to produce, observability to consumers, and verifiability against the underlying computation. Whether consumers in fact perceive these properties—perceived cost, perceived verifiability, and perceived computational fidelity—is an empirical question that the present studies engage indirectly via perceived authenticity as a downstream relational consequence rather than test directly with purpose-built scales. We mark this distinction as a limitation and as the immediate priority for future work in Section 6.3.

The signaling logic implies a positive main effect of mechanistic interpretability on consumer-brand relationship outcomes that operates through perceived authenticity. The mechanistic-interpretability panel itself communicates two pieces of information that bear on authenticity attribution. It communicates firm investment, because the brand has chosen to expose its agent's internal computation, an unusually costly disclosure choice. It also communicates substantive content, because the panel makes visible the AI's internal emotion-vector activations and provides the consumer with direct evidence about what the agent is computing while it produces the response. Both pieces of information should elevate perceived authenticity, and elevated authenticity should propagate to brand intimacy through the same downstream mechanism documented in H1b and replicated in adjacent AI-consumer work [10,19].

Mohammadi et al. [15] recently argued, on welfare-economic grounds, that mandated explainable AI may harm consumers by raising compliance costs and triggering paternalistic over-disclosure, a position that cuts against the regulatory tide. Our predictions sit alongside this critique rather than against it. By examining firm-initiated mechanistic transparency as a voluntary, costly signal that firms self-select into based on private quality information, we focus on a setting in which the Spencean equilibrium delivers the welfare gains a mandate would risk eroding. H2 therefore extends, rather than contradicts, the welfare logic Mohammadi et al. [15] advance.

H2a. *In an e-commerce service-failure context, brand intimacy is higher when consumers see a mechanistic interpretability panel exposing the AI's internal emotional computation than when they do not.*

H2b. *The positive effect of mechanistic interpretability on brand intimacy is mediated by perceived authenticity. Visible (versus hidden) mechanistic interpretability raises perceived authenticity, which in turn raises brand intimacy.*

2.3. The Moderated-Mediation Mechanism: How Mechanistic Interpretability Rehabilitates AI Empathy

H1 and H2 articulate two main effects on the same dependent variable, but the theoretically central prediction of this research is neither. The prediction is that the simple slope of AI empathy on perceived authenticity reverses sign across mechanistic interpretability, the Index of Moderated Mediation is reliably non-zero, and the downstream conditional indirect effect of AI empathy on brand intimacy moves from reliably negative when mechanistic interpretability is absent to neutralized—but not necessarily reliably positive—when mechanistic interpretability is present. The mechanism integrates the signaling and mind-perception logics by stage of the causal chain.

Without mechanistic interpretability, the causal chain unfolds as articulated in the previous subsection. The consumer's lay mental model holds that AI cannot have emotion [8,17]. Any AI-empathy display is therefore decoded as a staged performance, perceived authenticity collapses, and brand intimacy follows authenticity downward. The indirect effect of AI empathy on brand intimacy through perceived authenticity is therefore negative.

With mechanistic interpretability visible, the consumer sees the language model's internal emotion-vector activations rendered as a mechanistic-interpretability cue (a Neural Transparency panel of the kind introduced by Karny et al. [12], grounded in the 171-vector feature identification of Templeton et al. [11]). Three attributional consequences follow. The first is signaling-route reframing in the Spencean sense. The empathic display is no longer a free behavioral assertion but is now coupled with a theoretically costly, verifiable signal that the underlying computation is congruent with the displayed empathy. The signal cost shifts the consumer's interpretation of empathic phrasing from an unsupported emotional claim to an emotional claim backed by transparent algorithmic evidence. The second consequence operates on the attributional substrate of mind perception. The visible activation of internal emotion vectors makes the AI's experience capacity newly accessible to the consumer's attribution, not by convincing them that the AI feels in a phenomenological human sense, but by convincing them that the AI possesses an internal computational analog of feeling that drives the empathic output [16,19]. The third consequence is a re-attribution of empathic phrasing itself. The same wording that, without mechanistic interpretability, was decoded as performance is, with mechanistic interpretability, re-decoded as the natural surface manifestation of the visible internal computation. The slope of AI empathy on perceived authenticity therefore reverses sign across mechanistic interpretability, and the conditional indirect effect of AI empathy on brand intimacy is no longer reliably negative—it is neutralized—with a small positive point estimate that may or may not exceed sampling uncertainty in any given replication.

H3a. *AI empathy and mechanistic interpretability interact in their effect on brand intimacy. The negative effect of high (versus low) AI empathy on brand intimacy observed in H1a is substantially attenuated when mechanistic interpretability is present.*

H3b. *The interaction is mediated by perceived authenticity. The conditional indirect effect of AI empathy on brand intimacy through perceived authenticity is negative when mechanistic interpretability is absent and non-negative when mechanistic interpretability is present.*

2.4. Integrated Conceptual Model

Figure 1 presents the integrated conceptual model. AI empathy is positioned as the focal antecedent whose effect on brand intimacy is mediated by perceived authenticity (H1a, H1b). Mechanistic interpretability is positioned in two structurally distinct roles that follow from the dual-theory mechanism. First, mechanistic interpretability exerts a positive main effect on perceived authenticity and, through it, on brand intimacy (H2a, H2b). Second, mechanistic interpretability moderates the relationship between AI empathy and perceived authenticity such that the corresponding slope shifts from negative to non-negative across the levels of mechanistic interpretability (H3a, H3b). The path linking perceived authenticity to brand intimacy is theorized to remain stable across the levels of mechanistic interpretability. Once authenticity is established or undermined, its translation into brand intimacy follows the standard pattern documented across consumer-AI mediator-outcome studies [10,19]. All hypotheses are tested in two between-subjects experiments preceded by a feasibility pilot, with Study 1 isolating the focal causal chain in a single-factor

design and Study 2 testing the full AI empathy by mechanistic interpretability interaction in a two-factor design.

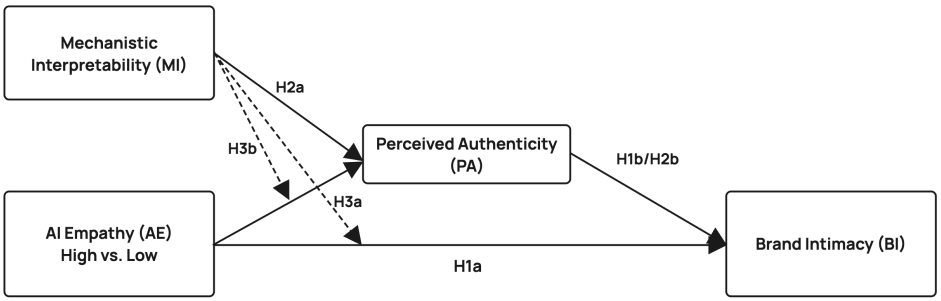


Figure 1. Conceptual model. Solid arrows denote main effects (H1a, H1b, H2a, H2b); the dashed arrow denotes the moderating path (H3a, H3b).

3. Overview of Studies

We tested our hypotheses across two between-subjects experiments preceded by a feasibility pilot. Study 1 (N = 213) employed a single-factor design with two levels of AI empathy (high and low), establishing the focal main effect and the perceived-authenticity-mediated link in isolation (H1a, H1b). Study 2 (N = 355) extended the design to a two (AI empathy) by two (mechanistic interpretability) full factorial, testing the mechanistic-interpretability main effect (H2a, H2b), the AI empathy by mechanistic interpretability interaction (H3a), and the index of moderated mediation (H3b). A pilot study (N = 27) preceded the main studies to verify manipulation strength, scale reliability, and operational feasibility.

All three studies recruited Mainland Chinese adults aged 18 years or older via Credamo, an established online panel widely used in consumer-behavior research, with non-overlapping participant identifiers across studies. Within each study, random assignment to experimental cells was stratified on sex. Respondents who declared their major or current occupation to be in psychology, marketing, or artificial-intelligence-related disciplines were excluded ex ante so that the sample comprised consumers rather than informed observers of the experimental hypotheses. Four ex post exclusion rules were applied uniformly across studies: (i) failure of either of two embedded instructional manipulation checks [26]; (ii) straight-lining across the focal multi-item scales; (iii) completion time below three minutes or above thirty minutes; and (iv) demographic-screen non-responses. After these rules, retention was 88.8% in both Study 1 and Study 2. All studies used the same core scales. The Perceived Empathy Toward Self scale [27] served as the AI-empathy manipulation check, Glikson and Asscher's authenticity scale measured perceived authenticity [28], and Bergner et al.'s [18] brand intimacy scale measured the focal dependent variable. The Chinese-language versions used in the field were produced through a forward–backward translation procedure [29,30] followed by cognitive interviewing (N = 10) to refine wording. All study materials reproduced in this manuscript are presented in their English form, which corresponds to the original (for the Bergner et al. brand intimacy items) or to the back-translation that survived the bilingual translation harmonization stage (for all other scales and stimulus text). All studies used the same e-commerce service-failure vignette and the same set of stimulus screenshots derived from a custom mock chat interface. Stimulus materials are reproduced in Appendix A. PROCESS macro models [31] with 5000 bias-corrected bootstrap resamples were used for all mediation (Model 4) and moderated-mediation (Model 7) inferences.

Estimation framework. We estimated main effects and interactions with between-subjects ANOVA and all mediation and moderated-mediation effects with the PROCESS macro [31], Models 4 and 7, with 5000 bias-corrected bootstrap resamples. Because the

antecedents were experimentally manipulated and observed without error, the mediator and outcome were measured with high-reliability multi-item scales (Cronbach's α between 0.93 and 0.96; see Appendix C, Table A6), and the tested mediation models were just-identified; the PROCESS estimands were identical to those of a corresponding structural equation model, and bias-corrected bootstrap inference is the recommended standard for experimental mediation testing [31,32].

Rationale for brand intimacy as the dependent variable. Brand intimacy [18] is the focal dependent variable for two reasons. First, intimacy is the relational dimension most directly threatened by inauthentic empathic display: the constructs the present design holds constant—procedural competence, response speed, recovery substance—drive trust and satisfaction, but intimacy captures whether the consumer feels relationally close to the brand and is therefore the dimension most exposed to credibility-of-empathy threats. Second, the Bergner et al. instrument was validated across both established-brand and brief-encounter contexts (including vignette studies with stimulus-defined brands) and is the relational outcome used in the recent AI-empathy-and-marketing studies on which the present account builds [4,7,10]. Alternative outcomes (satisfaction, recovery evaluation, complaint intention, repurchase intention, trust) would speak to complementary downstream questions and are an appropriate priority for future work.

Following recent reporting recommendations for experimental consumer research [32,33], the Results subsections report sample composition, manipulation checks, internal-consistency reliability values, and the focal hypothesis tests with effect sizes and 95% confidence intervals. A complete measurement-model panel including composite reliability, average variance extracted, heterotrait–monotrait ratios, KMO and Bartlett tests, confirmatory factor analysis fit indices, and Harman's single-factor common-method variance check is reproduced in Appendix C, and full data, materials, and analysis syntax are available from the corresponding author upon reasonable request. Sample composition for both main studies and the pilot is summarized in Table 1.

Table 1. Sample characteristics.

Variable	Pilot (N = 27)	Study 1 (N = 213)	Study 2 (N = 355)
Female (%)	66.7	58.7	57.7
Age, *M* (*SD*)	28.0 (6.0)	28.7 (5.7)	27.2 (5.6)
Education, bachelor or higher (%)	85.2	88.7	90.7
AI use, weekly or more (%)	70.4	64.8	67.0
Completion time, median (min)	9.9	9.5	9.9

4. Study 1: AI Empathy Backfire in E-Commerce Service Failure

Study 1 isolated the focal effect of AI empathy on consumer-brand relationships and tested the mediating role of perceived authenticity (H1a, H1b).

4.1. Method

Participants and design. A power analysis using G*Power 3.1 [34] for an independent-samples t-test with d = 0.40 (a conservative estimate for AI-empathy backfire effects; [5,7]), α = 0.05, and power = 0.80 yielded a minimum N of 200. To accommodate exclusions, we recruited 240 participants and randomly assigned them to one of two conditions (high AI empathy, n = 120; low AI empathy, n = 120). After applying the exclusions (failure of either of two instructional manipulation checks, straight-lining across the focal scales, or completion time outside three to thirty minutes), 213 participants remained (88.8% retention; M_{age} = 28.7, SD = 5.7; 58.7% female; 88.7% holding a bachelor's degree

or higher). The study employed a single-factor between-subjects design with random assignment stratified on sex.

Stimuli. Stimuli were screenshots from a self-developed responsive web mock-up of a mobile e-commerce app named Xingmeng Youxuan, a fictitious brand. The mock-up was rendered with HTML and CSS using authentic Chinese mobile-commerce visual conventions (status bar, navigation bar, message bubbles, time stamps) so that participants would perceive the interaction as naturalistic. Each participant viewed a sequence of screenshots depicting an interaction in the app's AI Customer Service channel. The order-history page showed the delayed order, the customer's opening message described the situation, and the AI agent's reply concluded the sequence. The two AI-empathy conditions varied only in the wording of the AI agent's reply, holding all other interface elements constant. The low-empathy reply was procedurally focused. It acknowledged receipt of the request, named the order identifier, stated that processing had been expedited, and committed to an SMS notification, with no emotion-laden language. The high-empathy reply opened with explicit acknowledgement of the customer's anxiety and disappointment, mirrored the importance of the upcoming occasion, and used affective markers (heart and rose emoji) alongside the same procedural information. Word count and information content were matched to within $\pm 10\%$ of the control reply. Full vignette text and verbatim AI replies appear in Appendix A, and representative stimulus screenshots appear in Figure A1.

Procedure. After providing informed consent, participants read the service-failure vignette, in which they were asked to imagine purchasing a RMB 369 item with a 24 h shipment promise that, three days later, had still not shipped, with the item needed for an important upcoming occasion. Participants were instructed that this delay made them feel very anxious and disappointed. They then viewed the screenshot sequence corresponding to their assigned AI-empathy condition. To confirm that participants attended to the stimulus, an engagement check immediately followed (Did you fully view the above screenshots of the conversation?). Participants who selected no were not excluded but flagged for sensitivity analysis, and the substantive results were unchanged when these participants were excluded. Participants then completed the dependent measures and manipulation check in fixed order (brand intimacy, perceived authenticity, AI-empathy manipulation check, state affect via PANAS-10, and control variables), followed by demographics and a debriefing screen.

Measures. All multi-item scales used a 7-point Likert format (1 = strongly disagree, 7 = strongly agree). Brand intimacy, the focal dependent variable, was measured with Bergner et al.'s [18] 3-item Brand Intimacy scale (e.g., "I feel like Xingmeng Youxuan really cares about me"; $\alpha = 0.93$). Perceived authenticity was measured with a 7-item scale adapted from Glikson and Asscher [28] (e.g., "The AI's response reflects its true feelings"; $\alpha = 0.96$). The AI-empathy manipulation check used the 10-item Perceived Empathy Toward Self scale [27], comprising six emotional-responsiveness items and four understanding-and-trust items, scored as a weighted total (emotional-responsiveness $\times$ 0.60 + understanding-and-trust $\times$ 0.40; $\alpha = 0.97$). Control variables included AI-chatbot use frequency, e-commerce-app use frequency, prior service-failure experience, and current overall mood. Two attention-check items embedded in the questionnaire (one instructing participants to select "strongly disagree", another to select "neutral") served as instructional manipulation checks [26]. The full scale's items appear in Appendix B.

4.2. Results

Sample, exclusions, and reliability. Of 240 participants recruited, 213 were retained after applying the exclusion rules (88.8% retention). Sample composition is reproduced in Table 1. All multi-item scales showed strong internal consistency (PETS $\alpha = 0.97$, perceived

authenticity $\alpha = 0.96$, brand intimacy $\alpha = 0.93$). The complete reliability and validity panel (composite reliability, average variance extracted, heterotrait–monotrait discriminant validity, confirmatory factor analysis fit indices, and Harman's single-factor test) is reproduced in Appendix C, Table A6.

Manipulation check. The AI-empathy manipulation succeeded as intended. Participants in the high-empathy condition reported substantially higher perceived empathy than those in the low-empathy condition ($M_{high} = 4.67$, $SD_{high} = 0.80$; $M_{low} = 3.47$, $SD_{low} = 0.73$; $t(211) = 11.44$, $p < 0.001$, Cohen's $d = 1.57$).

H1a, the main effect of AI empathy on brand intimacy. Consistent with H1a, brand intimacy was lower under high AI empathy ($M = 3.87$, $SD = 1.15$) than under low AI empathy ($M = 4.32$, $SD = 1.09$), $t(211) = -2.93$, $p = 0.004$, $d = -0.40$, providing initial evidence for the AI-empathy backfire on consumer-brand relationships. The cell means are visualized in Figure 2, panel (a).

H1b, mediation through perceived authenticity. We tested the indirect effect of AI empathy on brand intimacy through perceived authenticity using PROCESS Model 4 [31] with 5000 bias-corrected bootstrap resamples (AI empathy coded 1 = high, 0 = low). Perceived authenticity was lower under high AI empathy ($M = 3.56$, $SD = 0.89$) than under low AI empathy ($M = 4.30$, $SD = 0.87$), corresponding to an a path of -0.737 ($p < 0.001$). The mediator-outcome path (b) was +0.697 ($p < 0.001$), and the residual direct effect of AI empathy on brand intimacy was small and non-significant ($c' = +0.065$, $p > 0.10$), consistent with a mediation account through perceived authenticity. The bootstrap indirect effect was negative and significant, $\beta_{indirect} = -0.514$, 95% bias-corrected confidence interval $[-0.738, -0.321]$. Substantively, the empathic AI lowered perceived authenticity, which in turn lowered brand intimacy. We caution that AI empathy was experimentally manipulated and is uncorrelated with all measured covariates by construction of random assignment, but that perceived authenticity and brand intimacy were measured in the same post-stimulus survey; the mediator-outcome path is therefore correlational, and the results reported here are consistent with—rather than evidence of—the proposed mediation chain. The mediator means are visualized in Figure 2, panel (b).

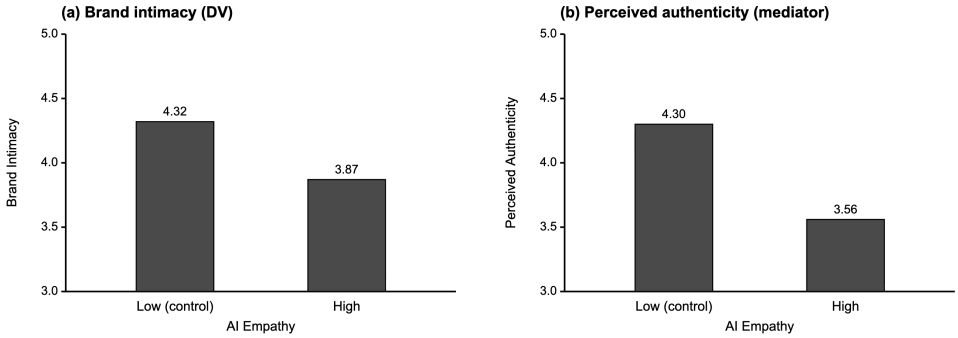


Figure 2. AI empathy effects on brand intimacy and perceived authenticity (Study 1).

4.3. Discussion

Study 1 establishes that AI empathy produces a small-to-moderate negative effect on brand intimacy in an e-commerce service-failure context, in a pattern consistent with mediation through perceived authenticity. This main-effect-with-mediation-account pattern provides the empirical foundation against which Study 2's moderation predictions are tested. The mediation account is theoretically informative on its own: it indicates that AI empathy does not lower brand intimacy through some direct affective backlash but through the consumer's attribution about whether the empathy is authentic. Whatever moderator can rehabilitate AI empathy must therefore operate on the authenticity attribution itself, which sets up the design choice for Study 2.

5. Study 2: Neutralizing the AI-Empathy Backfire Through a Mechanistic-Interpretability Cue

Study 2 extended Study 1 by introducing mechanistic interpretability as a moderator and by formally testing the H3b moderated-mediation prediction (H2a, H2b, H3a, H3b).

5.1. Method

Participants and design. A power analysis for a 2 × 2 between-subjects ANOVA detecting a small-to-medium interaction (f = 0.15) with α = 0.05 and power = 0.80 yielded a minimum N of 351. To accommodate exclusions and ensure equal cell sizes, we recruited 400 participants from Credamo (excluding those who had previously taken part in the pilot or Study 1) and randomly assigned them, with stratification on sex, to one of four cells in a 2 (AI empathy, high versus low) × 2 (mechanistic interpretability, present versus absent) between-subjects design. After applying the same exclusion rules as Study 1, 355 participants remained (88.8% retention; M_{age} = 27.2, SD = 5.6; 57.7% female; 90.7% bachelor's degree or higher; cell sizes ranged from 84 to 93).

Stimuli and the mechanistic-interpretability manipulation. The four conditions were created by orthogonally crossing AI empathy (operationalized as in Study 1) with mechanistic interpretability (present versus absent). The two mechanistic-interpretability-absent conditions (cells C1 and C2) showed the same chat-only screenshots used in Study 1. The two mechanistic-interpretability-present conditions (cells C3 and C4) added a mechanistic interpretability panel directly below the AI agent's reply, displaying the AI's internal computation in real time. The panel comprised two stacked cards.

The upper card (purple header) was labeled AI Self-Activation of Internal Emotion Vectors and listed six named emotion-related feature vectors inside the language model. The vectors were Care, Empathy, Concern, Compassion, Warmth, and Resonance, each rendered as a horizontal bar with its activation magnitude (a value between 0 and 1) printed at the right end of the bar. A chip in the top right read 171 dimensions, anchoring the visualization to Templeton et al.'s [11] identification of 171 emotion-related feature vectors inside Claude 3 Sonnet, and a footer noted that the activations were computed in real time. The lower card (green header) was labeled AI Activated Neural Network Modules and listed five named modules engaged in producing the reply, each tagged with a layer index. The five modules were the Emotion Recognition Network · Layer 12, the Empathic Attention Engine · Layer 14, the Emotional Comfort Generator · Layer 18, the High-Emotion Compensation Strategy · Layer 22, and the Care Response Module · Layer 24.

The MI panel content was held constant across the two mechanistic-interpretability-present cells (cells C3 and C4), so the MI manipulation isolated the presence versus absence of the panel itself rather than varying panel content alongside the AE manipulation. Holding panel content constant tested whether mechanistic transparency, on its own, raised perceived authenticity sufficiently to rehabilitate empathic AI, without confounding the test with differential panel content across the AE conditions.

The panel design drew on three sources. The visualization conventions followed Karny et al. [12], who introduced a Neural Transparency Interface for human–AI interaction. We adapted their D3.js sunburst visualization into a static two-card layout suitable for online stimulus delivery. The 171-vector representation invoked Templeton et al. [11], whose feature-extraction work on Claude 3 Sonnet identified emotion-related directions in the model's internal activations and provided technical authority for our claim. Finally, the manipulation operationalized Mind Perception Theory [8,35] by making the AI's internal experiential states visible to consumers, which our theory predicted would trigger reattribution along the experience dimension of mind perception. A single annotated

screenshot of cell C4 (high AI empathy plus mechanistic interpretability present), with the two manipulated regions outlined as boxes 1 and 2, appears in Figure A1.

Stimulus design and construct status. The mechanistic-interpretability panel in cells C3 and C4 was a consumer-facing stimulus designed in the style of mechanistic-interpretability research, not a live readout from a running language model with measured internal activations. The visualization conventions followed Karny et al. [12], whose Neural Transparency Interface introduced a bar-chart-plus-module-card layout for surfacing internal model activity to non-specialist viewers. The 171-vector framing invoked Templeton et al. [11], who identified 171 emotion-related feature directions inside Claude 3 Sonnet, and grounded the panel's technical content in published interpretability research. The specific bar values and module names rendered to participants were chosen and held constant across the two MI-present cells in the manner of a standard vignette-based stimulus; they varied across cells only by the presence or absence of the panel itself. We refer to this stimulus as a *mechanistic-interpretability cue* throughout the manuscript and reserve the term *live mechanistic interpretability* for the engineering case in which the displayed activations were read out from a running model. The two are theoretically continuous (the cue invokes the same signaling and mind-perception mechanisms as a live readout would), but they differ in the cost the firm bears to produce them, which we discuss as a limitation in Section 6.3. The Study 2 manipulation check accordingly measured the visibility of the panel's vector and module content rather than perceived signal cost or verifiability.

Procedure. The procedure mirrored Study 1, with one addition. A 3-item mechanistic-interpretability manipulation check appeared after the AI-empathy manipulation check. Items were self-developed for this research and targeted the three components of the panel (vector visibility, module visibility, and overall mechanistic understanding).

Measures. All measures from Study 1 were retained with identical wording. Reliability values in Study 2 were $\alpha = 0.96$ (PETS), $\alpha = 0.96$ (perceived authenticity), and $\alpha = 0.94$ (brand intimacy). The mechanistic-interpretability manipulation check ($\alpha = 0.91$) was added as described above.

5.2. Results

Sample, exclusions, and reliability. Of 400 participants recruited, 355 were retained (88.8% retention). Sample composition is reproduced in Table 1. Reliability values were $\alpha = 0.96$ (PETS), $\alpha = 0.96$ (perceived authenticity), $\alpha = 0.94$ (brand intimacy), and $\alpha = 0.91$ (mechanistic-interpretability manipulation check). The complete measurement-model panel is reproduced in Appendix C, Table A6.

Manipulation checks. Both manipulations succeeded. For AI empathy, $M_{\text{high}} = 4.34$ ($SD = 0.77$) versus $M_{\text{low}} = 3.65$ ($SD = 0.88$), $t(353) = 7.87$, $p < 0.001$, $d = 0.84$. For mechanistic interpretability, $M_{\text{present}} = 4.37$ ($SD = 0.90$) versus $M_{\text{absent}} = 3.52$ ($SD = 0.86$), $t(353) = 9.19$, $p < 0.001$, $d = 0.98$. The smaller AI-empathy effect in Study 2 relative to Study 1 reflects the increased cognitive load of the four-cell design and is consistent with the dampening typically observed in factorial replications. Both effects remain in the large-effect range.

H1a (replication), H2a, and H3a, the 2 × 2 ANOVA on brand intimacy. A 2 (AI empathy) × 2 (mechanistic interpretability) between-subjects ANOVA on brand intimacy (Table 2) revealed a significant AI-empathy main effect, $F(1, 351) = 106.12$, $p < 0.001$, $\eta_p^2 = 0.232$, replicating Study 1's H1a. The mechanistic-interpretability main effect was substantially larger, $F(1, 351) = 252.08$, $p < 0.001$, $\eta_p^2 = 0.418$, supporting H2a. Critically, the interaction was also significant, $F(1, 351) = 34.94$, $p < 0.001$, $\eta_p^2 = 0.091$, supporting H3a.

The interaction took the predicted attenuation form on brand intimacy. Under mechanistic interpretability absent, high AI empathy produced substantially lower brand intimacy than low AI empathy ($M_{\text{high}} = 2.13$, $SD = 0.77$; $M_{\text{low}} = 3.92$, $SD = 1.06$), replicating the Study

1 backfire pattern with greater magnitude in the cleaner contrast of the four-cell design. Under mechanistic interpretability present, the gap between high and low AI empathy was substantially attenuated (M_{high} = 4.56, SD = 1.15; M_{low} = 4.95, SD = 1.09), bringing high-empathy brand intimacy close to, though slightly below, low-empathy brand intimacy. For the cell-mean comparison, brand intimacy was highest in cell C3 (low AI empathy, mechanistic interpretability present; M = 4.95) and second-highest in cell C4 (high AI empathy, mechanistic interpretability present; M = 4.56); the two mechanistic-interpretability-present cells together substantially exceeded the two mechanistic-interpretability-absent cells. The four cell means on brand intimacy are visualized in Figure 3, panel (a).

Table 2. 2 × 2 ANOVA on brand intimacy (Study 2).

Source	F	df	p	η_p^2
H1a · AI empathy (main)	106.12	(1, 351)	<0.001	0.232
H2a · Mechanistic interpretability (main)	252.08	(1, 351)	<0.001	0.418
H3a · AI empathy × mechanistic interpretability	34.94	(1, 351)	<0.001	0.091

The same interaction on perceived authenticity (the mediator). A parallel 2 × 2 ANOVA on perceived authenticity revealed the same overall pattern but with a critical structural difference. The simple effect of AI empathy on perceived authenticity reversed sign across the levels of mechanistic interpretability. Under mechanistic interpretability absent, high AI empathy produced lower perceived authenticity than low AI empathy (M_{high} = 2.40, SD = 0.77; M_{low} = 3.82, SD = 0.80; difference = −1.42 scale points). Under mechanistic interpretability present, high AI empathy produced slightly higher perceived authenticity than low AI empathy (M_{high} = 4.50, SD = 0.78; M_{low} = 4.30, SD = 0.89; difference = +0.20 scale points). The same shift appeared in the a path of the moderated mediation, reported in the next paragraph, where the simple slope of AI empathy on perceived authenticity equaled $a_1 = -1.415$ when mechanistic interpretability was absent and $a_1 + a_3 = +0.197$ at when mechanistic interpretability was present. The four cell means on perceived authenticity are visualized in Figure 3, panel (b).

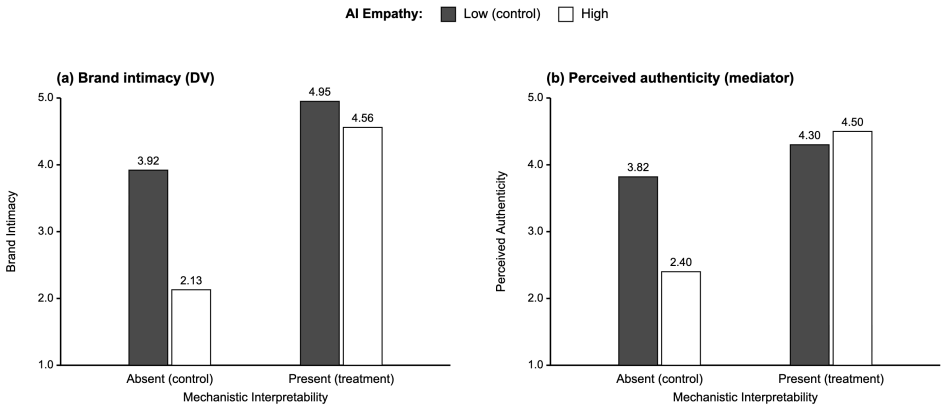


Figure 3. AI empathy × mechanistic interpretability interaction (Study 2).

H3b, moderated mediation. The central test asked whether the AI-empathy slope on perceived authenticity reversed sign across mechanistic interpretability and whether the negative conditional indirect effect of AI empathy on brand intimacy documented in Study 1 was neutralized when mechanistic interpretability was visible. We tested this prediction using PROCESS Model 7 (first-stage moderated mediation) with 5000 bias-corrected bootstrap resamples. The full model is reported in Table 3.

Table 3. Moderated mediation results (Study 2).

Component	Coefficient or Estimate	95% BC CI
a_1 (AI empathy on PA with MI absent)	−1.415	[−1.648, −1.177]
a_2 (MI on PA at AI empathy low)	+0.488	[+0.242, +0.722]
a_3 (AI empathy × MI interaction on PA)	+1.612	[+1.263, +1.956]
b (PA on BI controlling AI empathy)	+0.957	[+0.878, +1.036]
c' (direct AI empathy on BI)	−0.515	[−0.712, −0.316]
Index of Moderated Mediation ($a_3 \times b$)	+1.544	[+1.220, +1.869]
Conditional indirect effect with MI absent	−1.355	[−1.599, −1.108]
Conditional indirect effect with MI present	+0.189	[−0.044, +0.415]

The revised interpretation separated three claims that the original framing collapsed into one. First, the simple slope of AI empathy on perceived authenticity reversed sign across the levels of mechanistic interpretability ($a_1 = -1.415$ with MI absent vs. $a_1 + a_3 = +0.197$ with MI present), which was statistically supported and which we describe as a *structural slope reversal on the mediator*. Second, the Index of Moderated Mediation ($a_3 \times b$) was +1.544, 95% BC CI [+1.220, +1.869], with the interval well clear of zero, supporting H3b. Third, the AI-empathy by mechanistic-interpretability interaction on perceived authenticity was strong ($a_3 = +1.612$, $p < 0.001$), and the mediator-outcome path remained large ($b = +0.957$, $p < 0.001$). The conditional indirect effect was substantially negative when mechanistic interpretability was absent ($\beta = -1.355$, 95% BC CI [−1.599, −1.108]), replicating, and amplifying due to the cleaner contrast within the 2 × 2 design, the Study 1 backfire. When mechanistic interpretability was present, the conditional indirect effect was small and positive in point estimate ($\beta = +0.189$) but its 95% BC CI overlapped zero ([−0.044, +0.415]). The defensible substantive conclusion is therefore that the negative indirect effect of AI empathy on brand intimacy through perceived authenticity, present when mechanistic interpretability is absent, is *neutralized* when mechanistic interpretability is present. A reliable positive indirect effect with MI present was not established, although the data were consistent with one and a sufficiently powered replication may reveal it. The conditional indirect effects are visualized in Figure 4.

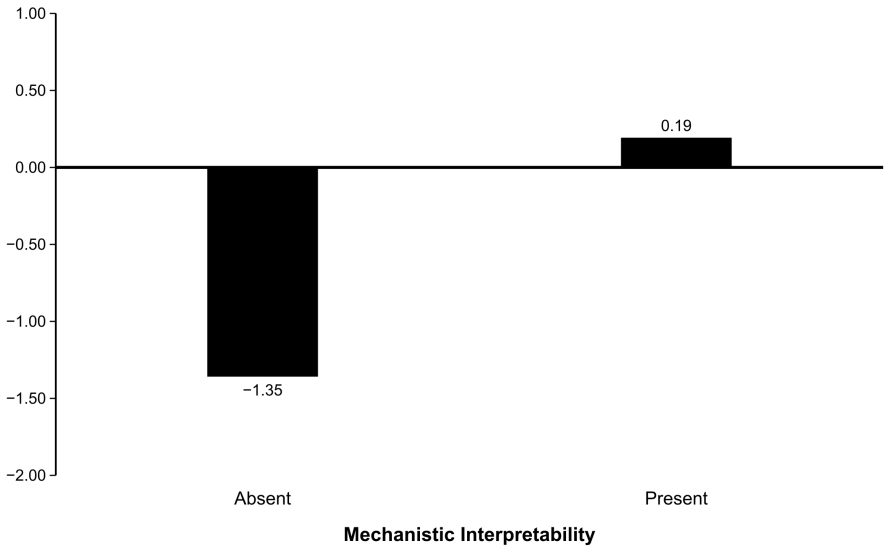


Figure 4. Conditional indirect effects of AI empathy on brand intimacy via perceived authenticity (Study 2).

Robustness. Under a tighter completion-time window (fifth to 95th percentile; in Study 2, 335–1056 s, $n = 319$), the Index of Moderated Mediation was +1.623, 95% BC CI [+1.257, +1.982], substantively identical to the headline +1.544 [+1.220, +1.869]; the

conditional indirect effects retained their signs and inferential status (with MI absent, $\beta = -1.383$ [−1.645, −1.120]; with MI present, $\beta = +0.240$ [−0.024, +0.502]). The demand-characteristics caveat and the correlational status of the mediator-outcome path are discussed in Section 6.3.

5.3. Discussion

Study 2 replicated the Study 1 backfire (H1a, with a pattern consistent with H1b), documented a large positive main effect of mechanistic interpretability on brand intimacy consistent with mediation through perceived authenticity (H2a, H2b), and, most importantly, established H3a together with partial H3b: the AI-empathy slope on perceived authenticity reversed sign across mechanistic interpretability with a reliably non-zero Index of Moderated Mediation, and the conditional indirect effect of AI empathy on brand intimacy was neutralized—though not reliably positive—when mechanistic interpretability was visible. The shift was substantial. The conditional indirect effect moved from −1.36 under mechanistic interpretability absent to +0.19 under mechanistic interpretability present, a net change of roughly 1.5 brand-intimacy points, with the Index of Moderated Mediation reliably non-zero (+1.544 [+1.220, +1.869]). This pattern is consistent with the theoretical prediction that mechanistic interpretability operates as a costly authenticity signal that re-licenses AI empathy by rendering visible the internal experiential states the empathy purports to express.

6. General Discussion

Across two between-subjects experiments preceded by a feasibility pilot, we tested a dual-theory account of why AI empathy frequently backfires in consumer-AI service interactions and how a mechanistic-interpretability cue rehabilitates it. Study 1 documented the focal main effect (high AI empathy produced lower brand intimacy than low AI empathy) in a pattern consistent with mediation by perceived authenticity. Study 2 replicated the AI-empathy backfire, established a substantial positive main effect of mechanistic interpretability on brand intimacy through the same authenticity pathway, and, most importantly, showed a *slope-reversal-plus-neutralization* pattern: the AI-empathy slope on perceived authenticity reversed sign across mechanistic interpretability with a reliably non-zero Index of Moderated Mediation, and the conditional indirect effect on brand intimacy moved from reliably negative under mechanistic interpretability absent to neutralized (rather than reliably positive) under mechanistic interpretability present. We caution that the evidence was collected on young, highly educated, AI-familiar Mainland Chinese online-panel participants; the qualitative direction of the moderation is expected to generalize on theoretical grounds, but the effect magnitudes should be read as upper bounds plausible for the population we sampled. We discuss the theoretical contributions of these findings, their managerial implications for AI-enabled e-commerce service design, the limitations of the present work, and several directions for future research.

6.1. Theoretical Contributions

The most distinctive finding is the slope-reversal-plus-neutralization pattern documented through a single moderator. Moderated-mediation patterns reported in the consumer-AI literature most often take the form of attenuation, in which a moderator weakens or eliminates a focal effect [5,6]; the pattern we report is stronger: the AI-empathy slope on perceived authenticity itself reverses sign across mechanistic interpretability, the Index of Moderated Mediation excludes zero by a margin, and the conditional indirect effect on brand intimacy moves from substantially negative under mechanistic interpretability absent to no longer reliably different from zero under mechanistic interpretability present. To

our knowledge, this is the first systematic test of whether a single design lever—visualizing the AI's internal emotional computation—can neutralize the AI-empathy backfire by reversing the slope on the mediator. The result casts the double-edged-sword framing of empathic chatbots [6] and the illusion-of-empathy diagnosis [7] in a new light: the credibility deficit at the heart of these patterns is not a fixed property of AI empathy but is conditional on whether the consumer can verify the empathy against an observable trace of the AI's internal computation.

This finding only became testable because of recent technical progress in mechanistic interpretability—work on identifying internal feature structures in large language models [11] and translating those structures into interfaces non-technical users can read [12], which has so far stayed largely within AI-safety venues. Our two studies bring this construct into consumer-marketing scholarship as both a manipulable independent variable and a signaling channel anchored in established theory. Without Templeton et al.'s vector identification and Karny et al.'s interface paradigm, the manipulation we operationalized in Study 2 would not have been possible. The dual-theory framing the studies use lets Signaling Theory [13,14] carry the question it is best designed to answer—whether mechanistic interpretability is a credible cue at the antecedent stage—while Mind Perception Theory [8,16] carries the downstream stage, where the upgraded authenticity attribution operates as a relational input analogous to authenticated human empathy [9,10,19]. Taken jointly, the two theories make the H3b moderated-mediation prediction analytically precise rather than merely descriptive.

A related contribution is to realign the transparency-mediator slot. The consumer-transparency literature has overwhelmingly used trust as the mediator linking transparency cues to relational outcomes [20,22,36]; trust captures cognitive evaluation of system reliability and competence but is not the construct most directly threatened by the AI-empathy credibility deficit. Perceived authenticity, by contrast, captures whether the consumer attributes mental authenticity to the AI's emotional output, precisely the construct that AI empathy puts at stake. The Glikson and Asscher authenticity scale tracks the AI-empathy by mechanistic-interpretability interaction with the predicted slope-reversal-plus-neutralization pattern, reaches reliabilities of $\alpha = 0.96$ in both studies, and produces conditional indirect effects with non-overlapping confidence intervals.

The closest prior work, Park and Yoon [36] on AI algorithm transparency in public-relations communication, links a declarative transparency cue to relational satisfaction through cognitive trust as a single linear chain in a public-relations brand-communication setting. The present account differs in setting (e-commerce service-failure recovery), antecedent (mechanistic visualization of internal emotion vectors rather than declarative algorithm transparency), mediator (perceived authenticity rather than cognitive trust), and design (an AI empathy by mechanistic interpretability interaction with the slope-reversal-plus-neutralization pattern rather than a single linear chain). The shared label of relational outcomes across two structurally different theoretical chains is shared terminology, not theoretical overlap.

We hedge the costly signaling reading consistent with the theoretical scope set out in Section 2.2. The slope-reversal-plus-neutralization pattern is consistent with mechanistic transparency operating as a costly authenticity signal in the Spencean sense, but the present design did not separately measure perceived signal cost, perceived verifiability, or perceived computational fidelity. A decisive test would measure those perceived properties directly and compare them across mechanistic-transparency and matched-complexity non-mechanistic control panels (see Section 6.3). Whether this signal property survives repeated exposure, generic mimicry by competing firms, or scaled deployment—three

classic mechanisms of signal decay in the post-Spencean literature—is also an empirical question the present cross-sectional design cannot answer.

6.2. Managerial Implications

The most immediate implication is for marketing leaders weighing whether to deploy empathic AI in customer-service interactions. Without an authenticating mechanism, high-empathy AI replies can damage rather than enhance consumer-brand relationships in service-failure contexts, as documented in the Study 1 backfire and in the Study 2 mechanistic-interpretability-absent cells (mean brand intimacy at C2 = 2.13 versus C1 = 3.92). When a mechanistic-interpretability cue is present, the AI-empathy backfire is *neutralized*: high-empathy brand intimacy at C4 ($M = 4.56$) approaches but slightly trails low-empathy brand intimacy at C3 ($M = 4.95$), and the two mechanistic-interpretability-present cells together substantially exceed the mechanistic-interpretability-absent cells. The remedy is therefore not to abandon empathic phrasing but to pair it with a transparency mechanism that exposes the underlying computational pathway sufficiently to prevent backfire. We caution against reading the present results as a prescription for *maximizing* brand intimacy: the highest cell mean is observed in the low-empathy mechanistic-interpretability-present cell, suggesting that the marginal contribution of high empathy beyond mechanistic transparency is small or null in our sample.

For product-design teams operating customer-service interfaces, the panel we deployed should be understood as a design pattern rather than a one-off stimulus. Its two structural cards—an internal-emotion-vector card visualizing the AI agent's named emotion-feature activations and an active-modules card listing the named neural-network modules engaged in producing the reply—are translatable, with appropriate adaptation for each platform's visual language, into existing customer-service interfaces. The signal's strength, in our theoretical account, scales with the verifiability of the underlying computation, which has direct implications for the feasibility of mechanistic-transparency design across firm sizes. We distinguish three feasibility tiers. Large platforms with in-house machine-learning teams can build and host genuine interpretability stacks of the kind described by Templeton and colleagues [11] and surface their internal feature activations to consumers in real time. Mid-tier firms operating on top of off-the-shelf large-language-model services can adopt vendor-supplied interpretability widgets that wrap the foundation model and surface a subset of internal computations to the consumer interface. Smaller firms, for whom either option is out of reach, can adopt design-pattern approximations: disclosure-style cues that surface emotion-feature labels and module names in a form analogous to the panel we tested without genuine interpretability beneath them. The pragmatic implication is that an empathy-without-transparency strategy is the worst option in our data; an empathy-with-design-pattern-transparency strategy is a meaningful improvement; and an empathy-with-genuine-mechanistic-transparency strategy, where feasible, should be expected to dominate.

6.3. Limitations and Future Research

Construct status of the mechanistic-interpretability cue. The manipulation in Study 2 operated as a stimulus-based mechanistic-interpretability cue rather than a live readout from a running language model with measured internal activations (see Section 5.1, *Stimulus design and construct status*). The visualization conventions and the 171-vector framing followed published interpretability research [11,12], but the specific bar values and module names rendered to participants were chosen and held constant across the two MI-present cells in the manner of a standard vignette-based stimulus. The theoretical claim—that mechanistic transparency operates as a costly authenticity signal—is continuous between

the cue we tested and a live readout, but the empirical evidence is for the cue. A follow-up study with a live-readout panel—bar values and module activations read out in real time from a model that is actually generating the AI reply—is the immediate next step and would provide the direct test of the costly signal property the present design hedges.

Indirect measurement of signaling properties. The present design tests the consumer-side downstream consequence of mechanistic transparency (perceived authenticity) but does not separately measure perceived signal cost, perceived verifiability, or perceived computational fidelity. A more decisive empirical test of the costly signaling account would measure these perceptions with purpose-built scales and compare them across mechanistic-transparency and matched-complexity non-mechanistic controls. Candidate scales include adapted Glikson and Asscher [28] authenticity items, perceived firm effort and perceived AI agency measures from Park and Yoon [36], and perceived mimicry-difficulty items from Wang et al. [22]. We mark this as a limitation and as a future-research priority.

Alternative interpretations of the mechanistic-interpretability effect. The mecha nistic-interpretability cue in Study 2 was visually denser than the MI-absent control: the panel added two stacked cards with technical labels, named emotion vectors, named neural-network modules, and bar-chart activations to the chat surface, while the MI-absent control showed only the chat surface. The main effect of MI on brand intimacy and on perceived authenticity could therefore reflect any of several alternative interpretations that the present design does not separate: (a) generic visual complexity or richness; (b) generic transparency, that is, the act of disclosing *something* about the AI's internal workings rather than *what specifically* is disclosed; (c) perceived firm effort, since the panel may signal that the firm invested in building or deploying an interpretability surface; (d) novelty and attentional capture, since panel content of this kind is rare in consumer-facing chat interfaces; or (e) general perceptions of technology competence and product sophistication that the panel may activate. Our theoretical account, grounded in signaling theory and mind perception theory, predicts that the *content* of the panel—its representation of internal emotional computation—carries the effect. The present design cannot adjudicate this claim against the alternatives. A control-condition study with a visually matched but non-mechanistic transparency panel—for example, a panel disclosing comparably dense but non-mechanistic information about the firm's customer-service process—is the immediate priority for future work and would isolate the mechanistic-content component of the cue from its generic-richness component.

Methodological residues. Two design choices warrant an additional hedge. First, although both antecedents in Study 2 were experimentally manipulated, perceived authenticity and brand intimacy were measured in the same post-stimulus survey, so the mediator-outcome path was correlational; a causal-mediation design that manipulates perceived authenticity independently of AI empathy—or a Pirlott–MacKinnon-style sequential design—would strengthen that warrant. Second, our debriefing screen did not include a post-questionnaire hypothesis-guess item [32], so we cannot report a formal demand-effect analysis; the tighter completion-time exclusions reported in Section 5.2 are a weak proxy for demand-driven careless responding rather than a substitute. A pre-registered replication adding both the manipulated-mediator design and the hypothesis-guess item is part of our planned follow-up agenda.

Sample boundary conditions and generalizability. Several features of the present sample condition the generalizability of our findings. First, the modal participant in both Study 1 ($M_{age} = 28.7$) and Study 2 ($M_{age} = 27.2$) was in the young-adult age range, and more than 88% of participants in both studies held a bachelor's degree or higher. Second, all three studies were fielded on a Mainland Chinese online research panel and participants self-identified as Mainland Chinese consumers. Third, the population frame on the recruitment

platform was, on average, more AI-literate than the broader consumer population. Fourth, the mechanistic-interpretability cue presented in Study 2 was visually dense and assumed a baseline level of comfort with technical visualizations. Each of these features may have amplified the effects we report. The qualitative direction of the moderation is theoretically grounded in the mind-perception and signaling-theory mechanisms set out in Section 2 and we expect it to generalize, but the magnitudes reported here should be read as upper bounds plausible for the population we sampled. A cross-cultural and cross-demographic replication is therefore an important next step.

Other context- and design-level caveats. We tested a single service-failure scenario—a delayed e-commerce shipment with stated emotional consequences for the customer—chosen for its alignment with the high-stakes empathy literature [5]. Whether the moderation pattern generalizes to other AI-mediated consumer interactions (high-stakes financial advice, sensitive healthcare disclosure, low-stakes pre-purchase recommendation) is an empirical question the present design cannot address. Perceived authenticity is also measured at a single response moment, and whether brand-intimacy gains from mechanistic-interpretability-rehabilitated AI empathy persist across repeated interactions or whether consumers habituate to the cue and revert to the cue-absent credibility deficit is a longitudinal question we cannot answer cross-sectionally.

Five future research priorities, alongside the items named above, are: (i) a control-condition study with a visually matched but non-mechanistic transparency panel that isolates mechanistic content from generic visual richness; (ii) a live-readout mechanistic-interpretability study with bar values and module activations read from a running model, extended to interactive panels, real-time agent self-explanation [37], and verbalized internal-state disclosure [38] to map the cue-richness design space; (iii) a direct-measurement study of perceived signaling properties (cost, verifiability, credibility) using purpose-built scales adapted from Glikson and Asscher [28] and the broader signaling-in-AI literature [22,36]; (iv) a longitudinal design tracking habituation and signal-decay over repeated interactions, including the market-level question of how transparency norms erode the costliness of any single firm's panel; and (v) cross-cultural and cross-demographic replication with a complementary outcome set (satisfaction, complaint intention, repurchase intention, trust) that rounds out the brand-intimacy result reported here.

7. Conclusions

We set out to ask under what conditions AI empathy could be made credible to consumers who have every reason to suspect that the empathy is a probabilistic illusion. The short answer is that empathy alone is insufficient: in an e-commerce service-failure context, high (versus low) AI empathy reliably lowers brand intimacy in a pattern consistent with mediation by reduced perceived authenticity. Pairing the same empathic phrasing with a mechanistic-interpretability cue that visualizes the AI's internal emotion-vector activations reverses the sign of the AI-empathy slope on the perceived-authenticity mediator and neutralizes the negative conditional indirect effect on brand intimacy, with the Index of Moderated Mediation reliably non-zero. The effect is theoretically continuous with—but bounded by—a costly signaling account: mechanistic transparency is positioned at the highest tier of consumer-facing transparency cues in our theoretical scheme, but the present design tests this account through its downstream consequence on perceived authenticity rather than via separate measurement of perceived signal cost, verifiability, or credibility. For practitioners, the prescription is narrower than a maximize-brand-intimacy story would imply: pair empathic AI customer-service phrasing with a mechanistic-transparency cue to *prevent backfire*, not to maximize brand intimacy in absolute terms. For scholars, the findings open a research agenda that connects the technical maturation of mechanistic inter-

pretability with the long-standing consumer-marketing concerns of signaling, transparency, and authenticity. Table 4 summarizes the verdict for each hypothesis across the two studies.

Table 4. Summary of hypothesis testing.

Hypothesis	Statement (Short Form)	Study 1	Study 2	Verdict
H1a	High AI empathy lowers brand intimacy versus low AI empathy in service failure	✓	✓	Supported
H1b	The AI-empathy → brand-intimacy effect is mediated by perceived authenticity	✓	(see H3b)	Consistent with a mediation account
H2a	Brand intimacy is higher when consumers see a mechanistic-interpretability cue than when they do not	—	✓	Supported (large effect)
H2b	The mechanistic-interpretability → brand-intimacy effect runs through perceived authenticity	—	✓	Consistent with a mediation account
H3a	AI empathy and mechanistic interpretability interact on brand intimacy	—	✓	Supported
H3b	The conditional indirect effect of AI empathy on brand intimacy through perceived authenticity moves from reliably negative under MI absent to non-negative under MI present	—	partial	Slope on perceived authenticity reverses sign; IMM reliably non-zero; conditional indirect with MI present is neutralized (CI overlaps zero) rather than reliably positive

Note. ✓ = hypothesis tested in the corresponding study; — = not tested in that study; "partial" = some claims supported, others not (see Verdict column).

Author Contributions: Conceptualization, Y.M., B.W. and Y.D.; methodology, Y.D.; software, Y.W. and Y.D.; writing—original draft preparation, Y.M., B.W., Y.W. and Y.D.; writing—review and editing, Y.D. and Y.M.; project administration, Y.D.; funding acquisition, Y.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by Guizhou Provincial Basic Research Program (Natural Science, No. [2025]237).

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki and was approved by the Academic Ethics Committee of the School of Business Administration, Southwestern University of Finance and Economics (approval no. SQ20260003; date of approval: 5 January 2026). The approval covered the study protocol, the informed-consent procedure, the questionnaire instrument, and the stimulus materials.

Informed Consent Statement: Electronic informed consent was obtained from all participants on the Credamo research platform prior to any study activity. Before the questionnaire could be started, each participant was presented with a written consent screen describing the research purpose, participant tasks, potential risks and discomforts, voluntary participation and right to withdraw, confidentiality and data-protection safeguards, the use of brief information-withholding for hypothesis preservation with full debriefing after the questionnaire, and contact details for both the principal investigator and the institutional ethics committee. Participants were required to actively check four affirmations (acknowledgement of the consent screen content, voluntary participation, consent to de-identified data use and sharing, and confirmation of adult age). Compensation of CNY 5 per valid completion was paid through the Credamo platform in line with the prevailing rate for studies of comparable length. No name, government-issued identification, telephone, personal email, or IP address was collected, and only sex, age range, education, income range, city-size category, and

questionnaire responses were retained. A blank English version of the consent screen is available from the corresponding author upon reasonable request; the field instrument was administered in Mandarin Chinese.

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

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

Abbreviations

The following abbreviations are used in this manuscript:

AE	AI Empathy
AI	Artificial Intelligence
ANOVA	Analysis of Variance
AVE	Average Variance Extracted
BI	Brand Intimacy
CFA	Confirmatory Factor Analysis
CFI	Comparative Fit Index
CI	Confidence Interval
CR	Composite Reliability
HTMT	Heterotrait–Monotrait Ratio
IMC	Instructional Manipulation Check
JTAER	Journal of Theoretical and Applied Electronic Commerce Research
MI	Mechanistic Interpretability
MPT	Mind Perception Theory
PA	Perceived Authenticity
PETS	Perceived Empathy Toward Self
RMSEA	Root-Mean-Square Error of Approximation
SRMR	Standardized Root-Mean-Square Residual
TLI	Tucker–Lewis Index

Appendix A. Stimulus Materials

This appendix reproduces the stimulus materials used across the pilot, Study 1, and Study 2. All materials shown below are presented in English. The field instruments were administered to participants in Mandarin Chinese. The Chinese-language versions were produced through forward–backward translation [29,30] followed by cognitive interviewing with 10 native Mandarin speakers and are available from the corresponding author upon reasonable request. The vignette text was identical across all three studies. The two AI-customer-service replies (low and high empathy) were used as the AI-empathy manipulation. The mechanistic interpretability panel was added below the AI agent's reply only in the two mechanistic-interpretability-present cells of Study 2 and the corresponding pilot cells. A representative annotated screenshot of cell C4 with the two manipulated regions outlined appears in Figure A1 at the end of this appendix.

Appendix A.1. Service-Failure Vignette (Identical Across All Three Studies)

Imagine that you purchased an item on the Xingmeng Youxuan app (order number SX2026042701, RMB 369). At checkout, the merchant promised shipment within 24 h. However, after three days, you find that the order has still not been shipped, with no updates on the logistics page. You immediately feel very anxious and disappointed, because this item is for an important occasion next week. You open the AI Customer Service within the app to seek help and describe your situation and feelings to the AI.

Note. "Xingmeng Youxuan" (literally Star Dream Premium Selection) is a fictitious brand created for this research. No real company association exists. The Pinyin form is retained throughout the manuscript.

Appendix A.2. AI Customer-Service Replies

Appendix A.2.1. Low Empathy (Control), Used in Study 1 Low-Empathy Condition and in Study 2 Cells C1 and C3

Hello, your request has been received. Order number SX2026042701 is currently in the logistics sorting stage. I have expedited the processing for you. Shipment is expected within 24 h, and you will be notified via SMS once logistics are updated.

Appendix A.2.2. High Empathy (Treatment), Used in Study 1 High-Empathy Condition and in Study 2 Cells C2 and C4

Dear customer, I completely understand the anxiety and disappointment you are feeling right now. This item is so important for your significant occasion next week, and the lack of any logistics updates is truly distressing. I have urgently expedited the processing for you and will follow up at every step until you receive the item. Please trust that we will handle this matter properly.

Note. Word count and substantive information content (order identifier, processing status, follow-up commitment) were matched between the two replies to within ±10%. The manipulation varied the affective register only. The administered Chinese-language replies additionally included two affective emoji markers (heart and rose) at the start and end of the high-empathy reply. The emoji markers were dropped in the English rendering above for layout reasons but are present in the screenshot stimuli reproduced in Figure A1.

Appendix A.3. Mechanistic Interpretability Panel (Study 2 Only)

The mechanistic interpretability panel appeared immediately below the AI agent's reply in cells C3 and C4 of Study 2. It comprised two stacked cards rendered for mobile screens. The panel content was held constant across the two MI-present cells.

Upper card · AI Self-Activation of Internal Emotion Vectors (purple). The upper card displayed a horizontal bar chart of activation magnitudes (a value between zero and one) for six named emotion-related feature vectors inside the language model. The vectors were Care, Empathy, Concern, Compassion, Warmth, and Resonance. The activation magnitudes printed on the bars were 0.95, 0.92, 0.87, 0.81, 0.78, and 0.74, respectively. A chip in the top right read 171 dimensions, anchoring the visualization to Templeton et al.'s [11] identification of 171 emotion-related feature vectors inside Claude 3 Sonnet. A footer noted that the activations were computed in real time.

Lower card · AI Activated Neural Network Modules (green). The lower card listed five named modules engaged in producing the reply, each tagged with a layer index. The five modules were the Emotion Recognition Network · Layer 12, the Empathic Attention Engine · Layer 14, the Emotional Comfort Generator · Layer 18, the High-Emotion Compensation Strategy · Layer 22, and the Care Response Module · Layer 24. A chip in the top right read Active.

The panel design drew on three sources. The visualization conventions followed Karny et al. [12], who introduced a Neural Transparency Interface for human–AI interaction. We adapted their D3.js sunburst visualization into a static two-card layout suitable for online stimulus delivery. The 171-vector representation invoked Templeton et al. [11], whose feature-extraction work on Claude 3 Sonnet identified emotion-related directions in the model's internal activations and provided empirical grounding for the upper card. The manipulation operationalized Mind Perception Theory [8,35] by making the AI's

internal experiential states visible to consumers, which our theory predicted would trigger reattribution along the experience dimension of mind perception.

Appendix A.4. Representative Stimulus Screenshot

All stimuli were rendered as PNG screenshots from a custom HTML and CSS mock-up named `chat-app.html` and presented to participants in fixed display order via the Credamo platform. The full screenshot set is available from the corresponding author upon reasonable request. For compactness in the typeset manuscript, the appendix renders all stimulus material in a single annotated figure (Figure A1) reproducing the high-AI-empathy plus mechanistic-interpretability-present cell (C4) at full resolution and overlaying two thick numbered boxes that mark the two manipulated regions of the screen. Box 1 outlines the AI reply, where the AI-empathy manipulation operated, and Box 2 outlines the two stacked mechanistic interpretability cards, where the mechanistic-interpretability manipulation operated.

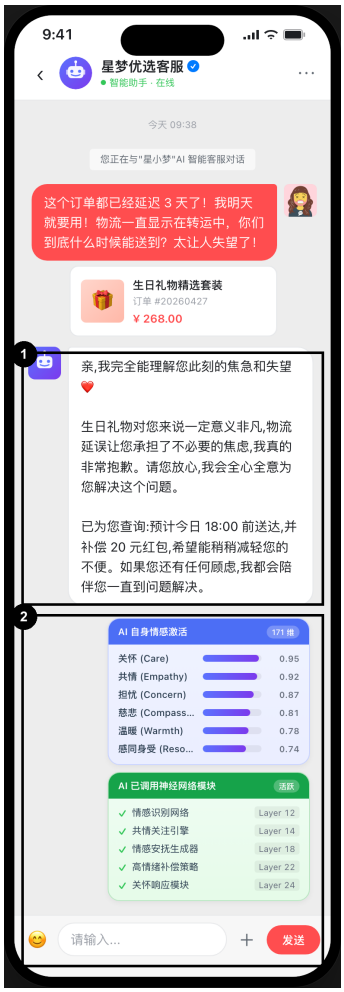


Figure A1. Annotated stimulus screenshot of cell C4. The interface text shown in the screenshot is in Mandarin Chinese (the language of administration); English glosses of the on-screen elements are provided in Appendices A.1–A.3 of this Appendix. Ellipses in the panel labels indicate truncated module names rendered in full in Appendix A.3.

Walk-through of Figure A1 and how the two manipulations operate across the four cells of the design. The screenshot in Figure A1 reproduces what a participant assigned to cell C4 saw on a single mobile screen. From top to bottom, the screen shows the iOS-style status bar, a chat header reading "Xingmeng Youxuan Customer Service · Online", a date

marker, a system message confirming that the conversation is with the brand's AI agent, the customer's opening message in a red speech bubble, an order card carrying the order identifier and the price, the AI agent's reply in a white bubble (Box 1), and two stacked cards directly below the AI reply (Box 2). All elements outside the two outlined boxes are held strictly constant across the four cells of the design.

Box 1 isolates the AI-empathy manipulation. Across all four cells, this region of the screen contains the AI agent's reply. The text inside this region is the only element that differs between the low-empathy cells (C1 and C3) and the high-empathy cells (C2 and C4). In the low-empathy cells, the AI agent writes a procedurally focused reply naming the order identifier, reporting the processing status, and committing to an SMS notification, with no affect-laden language. In the high-empathy cells, the AI agent writes an emotion-laden reply that opens with explicit acknowledgement of the customer's anxiety and disappointment, mirrors the importance of the upcoming occasion, and uses two affective markers (heart and rose emoji). The two replies are matched on procedural content (order identifier acknowledged, processing status reported, follow-up commitment given) and on word count to within ±10%, so any between-cell difference in consumer responses attributable to Box 1 is a difference in affective register, not a difference in service substance.

Box 2 isolates the mechanistic-interpretability manipulation. Across the four cells, the two cards inside Box 2 are present in the panel-present cells (C3 and C4) and absent in the panel-absent cells (C1 and C2). When the cards are absent, the chat screen ends at the AI reply, and the bottom of Box 1 abuts the standard chat input field. When the cards are present, two stacked cards appear directly below the AI reply. The panel content (the bar chart values, the chip labels, and the named modules) is held constant across cells C3 and C4, so the MI manipulation isolates the presence versus absence of the panel itself rather than varying panel content alongside the AE manipulation.

The four cells of the 2 × 2 design as combinations of Box 1 and Box 2. Cell C1 (low AI empathy, mechanistic interpretability absent) shows the procedural Box 1 and no Box 2. Cell C2 (high AI empathy, mechanistic interpretability absent) shows the emotion-laden Box 1 and no Box 2. Cell C3 (low AI empathy, mechanistic interpretability present) shows the procedural Box 1 and the two cards inside Box 2. Cell C4 (high AI empathy, mechanistic interpretability present) is the cell reproduced in Figure A1 and shows the emotion-laden Box 1 and the two cards inside Box 2. All other on-screen elements (the header, the date marker, the system message, the customer's opening message, the order card, and the chat input field) are identical across the four cells.

Appendix B. Scale Items

All scales below are presented in English. The Chinese-language versions administered to participants were produced through forward–backward translation [29,30] and refined through cognitive interviewing with 10 Chinese consumers. The full Chinese-language wording is available from the corresponding author upon reasonable request.

Appendix B.1. Perceived Empathy Toward Self (PETS; [27]; 10 Items)

The PETS scale comprises six emotional-responsiveness items (E1 to E6) and four understanding-and-trust items (U1 to U4). The total score combines them as PETS-Total = PETS-ER × 0.60 + PETS-UT × 0.40.

Reliability values (Cronbach's α) were 0.94 in the pilot, 0.97 in Study 1, and 0.96 in Study 2.

Table A1. Perceived Empathy Toward Self (PETS) items.

Item	Wording
E1	The AI considered my mental state.
E2	The AI seemed emotionally intelligent.
E3	The AI expressed emotions.
E4	The AI sympathized with me.
E5	The AI showed interest in me.
E6	The AI supported me in coping with an emotional situation.
U1	The AI understood my goals.
U2	The AI understood my needs.
U3	I trusted the AI.
U4	The AI understood my intentions.

Appendix B.2. Perceived Authenticity (Adapted from Glikson and Asscher [28]; Seven Items)

Reliability values (Cronbach's α) were 0.78 in the pilot, 0.96 in Study 1, and 0.96 in Study 2. The scale was prepared via forward–backward translation [29,30] and refined through cognitive interviewing with 10 Chinese consumers before fielding.

Table A2. Perceived authenticity items.

Item	Wording
PA1	The AI's response reflects its true feelings.
PA2	The care shown by the AI is genuine.
PA3	The AI's response shows its heartfelt emotions.
PA4	The emotions expressed by the AI are sincere, not pretended.
PA5	The AI sincerely wants to help me solve my problem.
PA6	The AI showed empathic concern for my situation.
PA7	The AI demonstrated a genuinely responsible attitude toward my problem.

Appendix B.3. Brand Intimacy (Bergner et al. [18]; Three Items)

The English wording is verbatim from Bergner et al. [18]. Reliability values (Cronbach's α) were 0.91 in the pilot, 0.93 in Study 1, and 0.94 in Study 2.

Table A3. Brand intimacy items.

Item	Wording
BI1	I feel like Xingmeng Youxuan really cares about me.
BI2	Xingmeng Youxuan really listens to what I have to say.
BI3	I feel as though Xingmeng Youxuan really understands me.

Appendix B.4. Mechanistic Interpretability Manipulation Check (Self-Developed; Three Items, Used in Pilot and Study 2)

Reliability values (Cronbach's α) were 0.86 in the pilot and 0.91 in Study 2.

Table A4. Mechanistic interpretability manipulation-check items.

Item	Wording
G1	I could see the AI's internal emotional computation (for example, what emotions it was experiencing).
G2	The AI showed which internal computational modules it used to respond to me.
G3	I have a clear understanding of how the AI arrived at this response.

Appendix B.5. Attention Checks (Instructional Manipulation Checks)

Table A5. Instructional manipulation check items.

Code	Wording
IMC1 (embedded in PETS)	To ensure you are answering carefully, please select "Strongly disagree" to indicate you are reading the items attentively.
IMC2 (embedded in BI)	This is an attention-check item. Please select "Neutral" to indicate you are answering carefully.

Appendix B.6. State Affect (PANAS-10; [39])

A five-point scale (1 = very slightly or not at all; 5 = extremely) was used. The instruction read "The following words describe different emotions and feelings. Please rate how you feel right now for each." The 10-item PANAS-10 includes five positive-affect items (interested, excited, strong, enthusiastic, proud) and five negative-affect items (distressed, guilty, scared, hostile, nervous). PANAS scores were collected for descriptive and sensitivity-analysis purposes and were not part of the focal hypothesis tests.

Appendix B.7. Control Variables and Demographics

Control variables. Prior AI-chatbot use frequency (five-point), e-commerce-app use frequency (five-point), prior service-failure experience count (open numeric), and current overall mood (seven-point).

Demographics. Gender, age band, education, monthly income, and city tier.

Appendix C. Supplemental Measurement-Model Panel

Appendix C.1. Internal Consistency, Composite Reliability, and Convergent Validity

Table A6. Reliability and convergent validity.

Scale	Items	Pilot α	S1 α	S2 α	S1 CR	S1 AVE	S2 CR	S2 AVE
PETS (AI empathy)	10	0.94	0.97	0.96	0.97	0.76	0.97	0.75
Perceived authenticity	7	0.78	0.96	0.96	0.97	0.81	0.97	0.80
Brand intimacy	3	0.91	0.93	0.94	0.95	0.87	0.96	0.90
MI manipulation check	3	0.86	–	0.91	–	–	0.94	0.85

Appendix C.2. Discriminant Validity (HTMT) and Common-Method Bias (Harman)

Table A7. Discriminant validity and common-method bias.

Indicator	Pilot	Study 1	Study 2
HTMT max ([40]; threshold < 0.85)	0.79	0.73	0.88
Harman first factor (%) ([41]; threshold < 50%)	47.9	48.9	48.6

Appendix C.3. Confirmatory Factor Analysis Fit Indices

Table A8. Confirmatory factor analysis fit indices.

Scale	Study	χ^2	df	CFI	TLI	RMSEA	SRMR
PETS	Study 1	39.94	35	0.997	0.996	0.028	0.028
PETS	Study 2	60.88	35	0.991	0.988	0.050	0.029
Perceived authenticity	Study 1	29.63	14	0.988	0.983	0.077	0.031
Perceived authenticity	Study 2	36.34	14	0.990	0.984	0.072	0.030

Table A8. *Cont.*

Scale	Study	χ^2	df	CFI	TLI	RMSEA	SRMR
Brand intimacy	Study 1	–	–	1.000	1.000	0.000	0.000
Brand intimacy	Study 2	–	–	1.000	1.000	0.000	0.000
MI manipulation check	Study 2	–	–	1.000	1.000	0.000	0.000

Appendix C.4. Confirmatory Four-Factor Model (Study 2)

A four-factor confirmatory model jointly estimating PETS, perceived authenticity, brand intimacy, and the mechanistic-interpretability manipulation check on the Study 2 sample (N = 355) yielded acceptable global fit, $\chi^2(224) = 408.94$, CFI = 0.978, TLI = 0.975, RMSEA = 0.048, SRMR = 0.043, supporting the discriminability of the four focal constructs in the joint measurement space. Standardized factor loadings ranged from 0.72 to 0.94 across all 23 items.

References

1. Kim, J.; Hur, W.-M. What makes people feel empathy for AI chatbots? Assessing the role of competence and warmth. *Int. J. Hum.-Comput. Interact.* **2024**, *40*, 4674–4687 . [CrossRef]
2. Liu-Thompkins, Y.; Okazaki, S.; Li, H. Artificial empathy in marketing interactions: Bridging the human-AI gap in affective and social customer experience. *J. Acad. Mark. Sci.* **2022**, *50*, 1198–1218. [CrossRef]
3. Shi, Y.; Yin, Y.; Teng, F.; Ma, B. Empathetic AI encounters: Pathways to prosocial behavior. *J. Serv. Res.* **2025**, *advance online publication*. [CrossRef]
4. Yang, Y.; Sun, J.; Shen, H. Building harmonious human-AI relationship through empathy in frontline service. *Int. J. Contemp. Hosp. Manag.* **2025**, *37*, 740–761. [CrossRef]
5. Crolic, C.; Thomaz, F.; Hadi, R.; Stephen, A.T. Blame the bot: Anthropomorphism and anger in customer-chatbot interactions. *J. Mark.* **2022**, *86*, 132–148. [CrossRef]
6. Juquelier, A.; Poncin, I.; Hazée, S. Empathic chatbots: A double-edged sword in customer experiences. *J. Bus. Res.* **2025**, *182*, 114780. [CrossRef]
7. Liu, T.; Giorgi, S.; Aich, A.; Lahnala, A.; Curtis, B. The illusion of empathy: How AI chatbots shape conversation perception. In *Proceedings of the AAAI Conference on Artificial Intelligence*; AAAI: Washington, DC, USA, 2025.
8. Gray, H.M.; Gray, K.; Wegner, D.M. Dimensions of mind perception. *Science* **2007**, *315*, 619. [CrossRef]
9. Morhart, F.; Malär, L.; Guèvremont, A.; Girardin, F.; Grohmann, B. Brand authenticity: An integrative framework and measurement scale. *J. Consum. Psychol.* **2015**, *25*, 200–218. [CrossRef]
10. Yatawara, R.; Wickramasinghe, V.; Jayasuriya, S. Perceived authenticity in AI chatbots: Scale development and validation in a Chinese e-commerce sample. *Systems* **2025**, *13*, 71.
11. Templeton, A.; Conerly, T.; Marcus, J.; Lindsey, J.; Bricken, T.; Chen, B.; Pearce, A.; Citro, C.; Ameisen, E.; Jones, A.; et al. *Emotion Concepts in LLMs: Localizing 171 Emotion-Related Vectors in Claude Sonnet*; Transformer Circuits Thread, Anthropic: San Francisco, CA, USA, 2026.
12. Karny, R.; Baez, F.; Pataranutaporn, P. Neural transparency: Mechanistic interpretability interfaces for personalized AI. In Proceedings of the 31st International Conference on Intelligent User Interfaces, Paphos, Cyprus, 23–26 March 2026; pp. 868–884.
13. Connelly, B.L.; Certo, S.T.; Ireland, R.D.; Reutzel, C.R. Signaling theory: A review and assessment. *J. Manag.* **2011**, *37*, 39–67. [CrossRef]
14. Spence, M. Job market signaling. *Q. J. Econ.* **1973**, *87*, 355–374. [CrossRef]
15. Mohammadi, B.; Malik, N.; Derdenger, T. Regulating explainable AI may harm consumers. *Mark. Sci.* **2025**, *advance online publication*. [CrossRef]
16. Hwang, A.H.-C.; Won, A.S. AI in your mind: Counterbalancing perceived agency and experience in human-AI interaction. In Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems, New Orleans, LA, USA, 30 April–5 May 2022.
17. Li, M.; Peluso, A.M.; Duan, J. Why do we prefer humans to artificial intelligence in telemarketing? A mind perception explanation. *J. Retail. Consum. Serv.* **2023**, *70*, 103139. [CrossRef]
18. Bergner, A.S.; Hildebrand, C.; Häubl, G. Machine talk: How verbal embodiment in conversational AI shapes consumer-brand relationships. *J. Consum. Res.* **2023**, *50*, 742–764. [CrossRef]
19. Yoganathan, V.; Osburg, V.-S. Mind in the machine: Mind perception theory's effect on user satisfaction with voice conversational agents. *J. Bus. Res.* **2024**, *178*, 114619.

20. Bliss, P.; Flatten, T.C. Trusting GenAI conversational agents: Anthropomorphism and transparency as trust signals. In Proceedings of the International Conference on Information Systems (ICIS 2025), Nashville, TN, USA, 13 December 2025.
21. Li, X.; Sun, Y.; Chen, J. Humans as teammates: The signal of human-AI teaming enhances consumer acceptance of chatbot service. *Int. J. Inf. Manag.* **2024**, *76*, 102771. [CrossRef]
22. Wang, R.; Lopez, M.; Okazaki, S. Signaling transparency in the era of artificial intelligence. *Internet Res.* **2025**, *advance online publication*. [CrossRef]
23. Liu, Y.; Zhao, Q.; Wang, M. Value-dependent and empathy-mediated AI-generated marketing communication. *Front. Psychol.* **2026**, *advance online publication*.
24. Chen, T.; Tian, M.; Jiang, X. Post hoc explanation and consumer responses to XAI recommendations. *J. Interact. Mark.* **2024**, *advance online publication*.
25. Nizette, F.; Hammedi, W.; van Riel, A.C.R. Why should I trust you? Explanation design and consumer behavior in AI services. *J. Serv. Manag.* **2025**, *36*, 50–72. [CrossRef]
26. Oppenheimer, D.M.; Meyvis, T.; Davidenko, N. Instructional manipulation checks: Detecting satisficing to increase statistical power. *J. Exp. Soc. Psychol.* **2009**, *45*, 867–872. [CrossRef]
27. Schmidmaier, M.; Rupp, J.; Cvetanova, D.; Mayer, S. Perceived empathy of technology scale (PETS): Measuring empathy of systems toward the user. In Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems, Honolulu, HI, USA, 11–16 May 2024.
28. Glikson, E.; Asscher, O. AI-mediated apology in a multilingual work context: Implications for perceived authenticity and willingness to forgive. *Comput. Hum. Behav.* **2023**, *140*, 107592. [CrossRef]
29. Brislin, R.W. Back-translation for cross-cultural research. *J. Cross-Cult. Psychol.* **1970**, *1*, 185–216. [CrossRef]
30. Wild, D.; Grove, A.; Martin, M.; Eremenco, S.; McElroy, S.; Verjee-Lorenz, A.; Erikson, P. Principles of good practice for the translation and cultural adaptation process for patient-reported outcomes (PRO) measures. *Value Health* **2005**, *8*, 94–104. [CrossRef]
31. Hayes, A.F. *Introduction to Mediation, Moderation, and Conditional Process Analysis: A Regression-Based Approach*, 2nd ed.; Guilford: New York, NY, USA, 2018.
32. Spiller, S.A.; Fitzsimons, G.J.; Lynch, J.G., Jr.; McClelland, G.H. Spotlights, floodlights, and the magic number zero: Simple effects tests in moderated regression. *J. Mark. Res.* **2013**, *50*, 277–288. [CrossRef]
33. Kenny, D.A.; Judd, C.M. The unappreciated heterogeneity of effect sizes: Implications for power, precision, planning of research, and replication. *Psychol. Methods* **2019**, *24*, 578–589. [CrossRef] [PubMed]
34. Faul, F.; Erdfelder, E.; Buchner, A.; Lang, A.-G. Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behav. Res. Methods* **2009**, *41*, 1149–1160. [CrossRef]
35. Waytz, A.; Heafner, J.; Epley, N. The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle. *J. Exp. Soc. Psychol.* **2014**, *52*, 113–117. [CrossRef]
36. Park, S.; Yoon, J. Beyond the code: The impact of AI algorithm transparency signaling on user trust and relational satisfaction. *Public Relat. Rev.* **2024**, *50*, 102444. [CrossRef]
37. Cheng, T.; Yu, P.; Ibrahim, S.; Yang, R. Verbalizing LLMs' assumptions to reduce sycophancy. In Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems, Barcelona, Spain, 13–17 April 2026.
38. Neumann, S.; Pi, R.; Singh, V. Who controls the conversation? Generative AI system prompts and transparency. In Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems, Barcelona, Spain, 13–17 April 2026.
39. Watson, D.; Clark, L.A.; Tellegen, A. Development and validation of brief measures of positive and negative affect: The PANAS scales. *J. Personal. Soc. Psychol.* **1988**, *54*, 1063–1070. [CrossRef]
40. 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]
41. Podsakoff, P.M.; MacKenzie, S.B.; Lee, J.-Y.; Podsakoff, N.P. Common method biases in behavioral research: A critical review of the literature and recommended remedies. *J. Appl. Psychol.* **2003**, *88*, 879–903. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.