International Journal of Scientific Research & Engineering Trends
Volume 12, Issue 3, May-Jun-2026, ISSN (Online): 2395-566X

The Explainable AI Paradox: When Transparency Improves Decision Quality and When It Creates Overconfidence

Dr. Harsha Sammangi, Poloju Pravalika

Abstract- Explainable artificial intelligence (XAI) is widely promoted as a remedy for algorithmic opacity, premised on the assumption that revealing a model's reasoning improves human oversight and decision quality. This study investigates the Explainable AI Paradox: the possibility that explanations simultaneously increase trust, reliance, and adoption while also producing overconfidence in flawed models — improving decision quality when models are valid but amplifying errors when models are not. Using a controlled experiment with 921 managers making pricing, lending, or inventory decisions assisted by AI systems of deliberately varied validity, participants were randomly assigned to one of five explanation conditions: no explanation, feature-importance, counterfactual, uncertainty-aware, or a combined feature-importance-plus-uncertainty condition. The study measured decision accuracy, confidence calibration, reliance behavior, override justification quality, and simulated financial outcomes. Results show that feature-importance explanations increased reliance (+9.8 percentage points, $p < .001$) and confidence (+0.61 scale points, $p < .001$) relative to no explanation, but produced the largest overconfidence increase (+0.084, $p < .001$) and the only negative financial outcome effect (–0.14 SD, $p < .01$) among all explanation types — concentrated specifically in the flawed-model conditions, where a significant Explanation × Model-Quality interaction ($\beta = 0.047$ to 0.089 across outcomes, all $p < .001$) confirms that feature-importance explanations' benefits accrue under valid models while their overconfidence costs accrue under flawed models. Uncertainty-aware explanations, by contrast, improved calibration (–0.058, $p < .001$), reduced overconfidence (–0.047, $p < .001$), and produced the only significant positive financial outcome (+0.29 SD, $p < .001$) relative to no explanation. A twelve-stage design intervention pilot demonstrates that combining five calibration-oriented design principles — feature reliability tagging, confidence-first ordering, disagreement prompts, active verification nudges, and explanation-accuracy feedback — reduces the Overconfidence Index by 83% (from 0.084 to 0.014, $p < .001$) relative to unmanaged feature-importance explanations. Thematic analysis of 40 participant interviews identifies six mechanisms underlying these patterns, including a 'plausibility heuristic substitution' through which surface-level explanation coherence substitutes for independent verification. The paper contributes a theory of the Explainable AI Paradox to behavioral information systems research, identifies model-quality and explanation-type interactions as the central moderating mechanism, and provides a five-level maturity roadmap and design decision framework for deploying explainable, uncertainty-aware managerial AI systems.

Keywords- Explainable AI, algorithmic decision-making, overconfidence, trust calibration, automation bias, human-AI collaboration, behavioral information systems, managerial decision-making, uncertainty quantification, reliance.

I. INTRODUCTION

Explainable artificial intelligence (XAI) has emerged as a near-universal prescription for the governance challenges posed by algorithmic decision-making in organizations. Regulatory frameworks increasingly mandate or encourage explanation provisions for high-stakes automated decisions; vendors market explanation features as differentiators; and a substantial academic literature has developed methods for generating model explanations, including feature-importance attributions (Ribeiro et al., 2016), counterfactual explanations (Wachter et al., 2018), and a growing array of uncertainty quantification techniques. The underlying assumption across this literature and its practical applications is that explanations serve an epistemically corrective function: by revealing how a model arrived at a recommendation, explanations enable human

decision-makers to identify when a recommendation should be trusted and when it should be scrutinized or overridden (Doshi-Velez & Kim, 2017; Miller, 2019).

However, a growing body of behavioral research on human-AI interaction suggests this assumption may be incomplete, and in some conditions, may be inverted. Bansal et al. (2021) found that AI explanations did not consistently improve human-AI team performance and in some configurations reduced it. Bucinca et al. (2021) demonstrated that explanations can increase 'overreliance' — reliance on AI recommendations that exceeds what the recommendations' actual accuracy would justify — even as they increase subjective measures of trust and satisfaction. These findings point toward what this study terms the Explainable AI Paradox: the possibility that the same explanatory content that, under valid models, helps decision-makers appropriately calibrate their reliance, may, under flawed models, instead provide a veneer of legitimacy that increases reliance on recommendations that do not warrant it — with the paradox arising precisely because decision-makers typically cannot directly observe whether a given model is valid or flawed, and must instead infer model quality partly from the explanations themselves.

This paradox has acute practical significance because the conditions under which explanations are most needed — novel deployments, evolving environments, models trained on incomplete or biased data — are often precisely the conditions under which models are most likely to be flawed in ways that are not yet known to the organization deploying them (Lebovitz et al., 2021). If explanations systematically increase confidence and reliance regardless of underlying model validity — as several of the mechanisms identified in the behavioral literature (automation bias, Mosier et al., 1998; the Dunning-Kruger-adjacent overconfidence dynamics, Dunning, 2011) would predict — then XAI's widespread deployment as a governance and trust-building tool may be systematically increasing organizational exposure to exactly the algorithmic errors it is intended to help organizations detect and correct. This concern is especially salient as generative AI and large language models are increasingly integrated into security-sensitive and IoT-connected organizational systems, where model opacity and ethical oversight gaps compound the governance risks identified here (Sammangi, Jagatha, & Liu, 2025b).

This study addresses four research questions: (RQ1) Do different types of AI explanations (feature-importance, counterfactual, uncertainty-aware, combined) differentially affect decision accuracy, confidence calibration, and reliance behavior relative to no explanation? (RQ2) Does the effect of explanations on these outcomes depend on whether the underlying AI model is valid or flawed — that is, does an Explanation × Model-Quality interaction exist, and if so, what is its direction and magnitude? (RQ3) Which explanation types are most susceptible to producing the overconfidence pattern central to the Explainable AI Paradox, and which, if any, mitigate it? (RQ4) Can explanation interface design interventions — independent of the underlying explanation method — reduce overconfidence while preserving the trust and adoption benefits that motivate XAI deployment?

Drawing on a controlled experiment with 921 managers across three decision domains (pricing, lending, and inventory management) and five explanation conditions, with AI model validity experimentally manipulated (valid versus deliberately flawed models, the latter containing spurious features, subgroup miscalibration, or brittleness to regime shifts depending on domain), this study makes four contributions. First, it provides among the first experimental demonstrations of a significant Explanation × Model-Quality interaction across multiple explanation types and decision domains, directly testing the Explainable AI Paradox's central proposition. Second, it identifies uncertainty-aware explanations as a distinct category whose effects diverge from feature-importance and counterfactual explanations in both direction and mechanism, with direct implications for XAI method selection. Third, it provides intervention-based evidence — a design pilot spanning five calibration-oriented interface principles — demonstrating that the overconfidence costs of feature-importance explanations can be substantially mitigated through interface design independent of the underlying explanation method. Fourth, it develops a theory of the Explainable AI Paradox grounded in a plausibility-heuristic-substitution mechanism identified through qualitative analysis, along with a practical design decision framework and maturity roadmap for XAI deployment in managerial decision contexts.

II. THEORETICAL BACKGROUND

Explainable AI: Methods, Promises, and Critiques

The XAI literature has developed a diverse methodological toolkit for generating model explanations. Feature-importance methods, including LIME (Ribeiro et al., 2016) and related approaches, identify which input features most influenced a

given prediction, typically presented as a ranked list or weighted contribution chart. Counterfactual explanations (Wachter et al., 2018) describe how an input would need to change to produce a different output, framed as 'what-if' statements that some scholars argue align more closely with how humans naturally reason about causal explanation (Miller, 2019). Uncertainty quantification methods produce calibrated confidence scores or intervals reflecting a model's case-specific certainty, distinct from feature-importance or counterfactual methods in that they characterize the model's epistemic state rather than its reasoning process per se.

Critiques of XAI have emerged along several lines. Rudin (2019) argued that post-hoc explanation methods applied to inherently opaque models may produce explanations that are themselves unreliable approximations of the model's true decision process — an explanation may be plausible without being accurate, a distinction with direct relevance to this study's central paradox. Lipton (2018) similarly cautioned against conflating interpretability with the human-perceived plausibility of an explanation, noting that explanations optimized for human acceptance may diverge from explanations that accurately represent model behavior. Ghassemi et al. (2021), writing in a healthcare context, argued that current XAI approaches may provide 'false hope' by creating an appearance of interpretability without corresponding gains in clinical decision quality — a critique this study's experimental design directly operationalizes by manipulating model validity independently of explanation provision. The stakes of this gap are amplified in clinical AI deployments, where privacy, data security, and predictive reliability are simultaneously at issue: decentralized AI architectures for healthcare IoT must navigate both model transparency and security constraints (Sammangi, Ambati, Liu, & Jagatha, 2025); blockchain-based frameworks for telemedicine offer complementary mechanisms for preserving data integrity while AI recommendations are presented to clinicians (Sammangi, Jagatha, & Liu, 2025c); and deep-learning diagnostic aids — such as those applied to kidney stone forecasting — face the same overreliance risks this study identifies, making calibrated uncertainty communication critical in clinical contexts (Sharma et al., 2025a).

Trust, Reliance, and Algorithm Aversion/Appreciation

The behavioral literature on human reliance on algorithmic recommendations has documented two seemingly contradictory phenomena. Algorithm aversion (Dietvorst et al., 2015) describes a tendency for people to abandon reliance on an algorithm after observing it err, even when the algorithm's average performance exceeds human performance — a pattern Dietvorst et al. (2018) found could be mitigated by giving people even minimal control over the algorithm's outputs. Algorithm appreciation (Logg et al., 2019), conversely, describes a tendency for people to weight algorithmic advice more heavily than equivalent human advice in certain contexts, particularly for objective or quantitative judgments.

XAI sits at the intersection of these phenomena: explanations are frequently proposed as a mechanism for increasing appropriate algorithm appreciation while reducing inappropriate algorithm aversion — helping people trust algorithms more when warranted and less when not. However, Bucinca et al.'s (2021) overreliance findings, and Zhang et al.'s (2020) demonstration that confidence scores and explanations can both independently and jointly affect trust calibration in ways that do not always track actual model accuracy, suggest that explanations may shift the overall level of trust and reliance without necessarily improving — and in some cases while degrading — the calibration of that trust to actual model performance. This distinction between trust level and trust calibration is central to this study's outcome measures (Section 3.3): the Overconfidence Index specifically captures miscalibration (confidence exceeding accuracy) rather than confidence level per se, enabling this study to distinguish 'explanations increase trust' (a level effect, well-documented) from 'explanations improve or worsen trust calibration' (a calibration effect, the focus of the Explainable AI Paradox).

Automation Bias and the Plausibility Heuristic

Automation bias (Mosier et al., 1998) — the tendency to favor automated system outputs and to use them as a heuristic replacement for vigilant information-seeking and processing — provides a foundational mechanism for understanding how explanations might increase reliance without improving calibration. If an explanation increases the perceived legitimacy or sophistication of an AI recommendation, it may increase automation bias by providing additional surface-level justification for deferring to the recommendation, even if the explanation itself does not provide information that would allow a careful decision-maker to assess the recommendation's actual reliability in a given case.

This study proposes the plausibility heuristic as the specific cognitive mechanism through which this dynamic operates:

when an explanation's content is consistent with a decision-maker's existing mental model of the decision domain — for instance, a feature-importance explanation for a pricing recommendation that lists 'demand history' and 'seasonality' as top factors, both of which a manager would expect to be relevant — the explanation's surface plausibility may be processed as evidence of the recommendation's validity, substituting for the more effortful process of independently verifying whether the recommendation is well-founded in the specific case at hand. Crucially, surface plausibility and actual validity are not the same thing: a flawed model could produce an explanation listing plausible-sounding features (because the model's spurious features happen to co-occur with, or be labeled similarly to, genuinely relevant features) while the underlying recommendation remains systematically biased — precisely the scenario this study's flawed-model conditions are designed to create (Table 2).

Uncertainty Communication and Calibration

A distinct literature on uncertainty communication (Hofman et al., 2020) examines how the presentation of uncertainty information affects human judgment and decision-making, with findings suggesting that explicit uncertainty communication can improve calibration but may also, in some presentation formats, be misread or ignored. This study's inclusion of an uncertainty-aware explanation condition (E3) — distinct from feature-importance (E1) and counterfactual (E2) explanations in that it characterizes the model's confidence rather than its reasoning — allows this study to test whether uncertainty communication's calibration benefits, documented primarily in contexts outside AI-assisted managerial decision-making, extend to this study's experimental contexts and whether they interact with model validity in the same way feature-importance and counterfactual explanations do.

The Explainable AI Paradox: A Conceptual Model

Synthesizing the preceding theoretical perspectives, this study proposes the conceptual model presented in Figure 1, which characterizes the Explainable AI Paradox as arising from the interaction between explanation presence (and type), the plausibility heuristic mechanism, and model quality as a moderator. The model proposes that explanations activate a plausibility assessment process whose outcome — high or low perceived plausibility — shapes subsequent reliance and verification behavior largely independent of actual model validity (since plausibility is assessed against the decision-maker's prior mental model, not against the AI model's true decision logic, which remains unobserved). When model quality and explanation plausibility are aligned (valid models producing plausible explanations, or invalid models producing implausible explanations that appropriately trigger skepticism), explanations support an improvement pathway. When they are misaligned (flawed models producing plausible-sounding explanations — the scenario this study's experimental design specifically constructs), explanations support a paradox pathway in which increased trust and reduced verification compound rather than correct underlying model errors.

Figure 1. The Explainable AI Paradox: A Conceptual Model of Explanation Type, Plausibility Heuristic, and Model-Quality Moderation

Explanation Presence	Cognitive Mechanism	Moderator: Model Quality	Dual Outcome Pathways
Explanation Types: • None (E0) • Feature-Importance (E1) • Counterfactual (E2) • Uncertainty-Aware (E3) • Combined (E4) Each explanation type makes specific aspects of model reasoning visible, but visibility does not guarantee verifiability	**Plausibility Heuristic:** • Explanation 'looks like' sound reasoning • Surface coherence substitutes for verification • Reduces perceived need for independent judgment **Calibration Pathway:** • Uncertainty signals enable case-specific reliance	**Valid Model:** Explanations reflect genuine model logic; plausibility heuristic and accuracy are aligned **Flawed Model:** Explanations reflect spurious or miscalibrated logic; plausibility heuristic and accuracy diverge sharply	**Improvement Pathway:** • Higher accuracy • Better-calibrated confidence • Appropriate selective reliance **Paradox (Overconfidence) Pathway:** • Increased trust and reliance • Increased confidence

	• Counterfactuals enable mental-model alignment checks		• Decreased verification • Errors persist or amplify under flawed models

Note. The model proposes that explanation type shapes which aspects of model reasoning are made visible, the plausibility heuristic mechanism processes this visibility largely independent of underlying model validity, and model quality (valid vs. flawed, experimentally manipulated in this study, Table 2) moderates whether the resulting reliance and confidence patterns track or diverge from actual decision quality.

Hypothesized Relationships

Based on the conceptual model and literature review, this study formulates the following hypotheses. H1: Explanation presence (any of E1–E4, relative to E0) increases reliance on AI recommendations and confidence in decisions made with AI assistance. H2: The effect of explanation presence on the Overconfidence Index (confidence minus accuracy) is moderated by model quality (valid vs. flawed), such that explanations increase overconfidence specifically under flawed models (Explanation × Flawed Model interaction, positive). H3: Uncertainty-aware explanations (E3) produce smaller overconfidence increases — or overconfidence decreases — relative to feature-importance (E1) and counterfactual (E2) explanations, reflecting uncertainty-aware explanations' distinct calibration-oriented mechanism (Section 2.4).

H4: Combining feature-importance and uncertainty-aware explanations (E4) does not simply average their individual effects but produces outcomes reflecting which explanatory element receives attentional priority (tested via the qualitative analysis, Section 5). H5: Calibration-oriented interface design interventions — independent of explanation method — reduce the Overconfidence Index associated with feature-importance explanations under flawed models.

III. RESEARCH METHODOLOGY

Experimental Design Overview

This study employs a 5 (explanation condition) × 2 (model quality: valid vs. flawed) × 3 (decision domain: pricing, lending, inventory) between-subjects factorial experiment, with model quality manipulated within each decision domain through deliberately introduced model flaws specific to that domain (Table 2). Participants were randomly assigned to one explanation condition (Table 1) and one decision domain (Table 2), and within their assigned domain, completed a sequence of 20 decisions, with model quality (valid vs. flawed) randomized at the decision level within-subjects — that is, each participant encountered both valid-model and flawed-model recommendations across their 20 decisions, in randomized order, without being informed which decisions involved which model version. This within-subjects model-quality manipulation, combined with between-subjects explanation condition assignment, enables this study to estimate the Explanation × Model-Quality interaction (central to H2) with high statistical power while limiting the total number of between-subjects cells required.

Table 1. Experimental Design: Explanation Conditions

Condition	Explanation Type	AI Recommendation Format	n (Participants)	Theoretical Expectation
E0 (Control)	No Explanation	Recommendation presented as a single output value with confidence not disclosed	186	Baseline reliance and accuracy; algorithm-aversion susceptible
E1	Feature-Importance	Recommendation accompanied by ranked list of top contributing input features and their relative weights	184	Increased trust and reliance; risk of misplaced confidence in plausible-sounding features
E2	Counterfactual	Recommendation accompanied by a 'what would need to change' statement (e.g., 'if income were $4,200 higher, approval likely')	182	Improved actionable understanding; may improve calibration by

Condition	Explanation Type	AI Recommendation Format	n (Participants)	Theoretical Expectation
				revealing decision boundaries
E3	Uncertainty-Aware	Recommendation accompanied by a calibrated confidence interval and model-confidence score reflecting case-specific uncertainty	188	Best calibration outcomes; may reduce reliance on low-confidence recommendations appropriately
E4	Feature-Importance + Uncertainty (Combined)	Recommendation accompanied by both feature-importance ranking and calibrated uncertainty score	181	Tests whether combining explanation types compounds benefits or compounds overconfidence risk

Note. N totals sum to 921 across the five conditions. Explanation format examples are illustrative; full stimulus materials, including exact wording and visual layout for each condition and decision domain, are available from the corresponding author upon request. All explanation content was generated from the same underlying model outputs; explanation conditions varied only in which aspects of model output were surfaced to participants, not in the underlying recommendations themselves (except where model quality, Table 2, was experimentally varied).

Table 2 details the three decision domains and the specific model flaws introduced in each domain's 'flawed' condition. Flaws were designed to be realistic analogues of documented real-world AI failure modes: spurious feature reliance (pricing), subgroup miscalibration (lending), and brittleness to distributional shift (inventory) — each representing a distinct class of model validity failure with potentially different implications for how explanations might or might not help decision-makers detect the flaw.

Table 2. Decision Domain Task Characteristics and Model Flaw Manipulations

Decision Domain	Task Description	Ground-Truth Availability	AI Model Type (Underlying)	n (Participants)
Pricing	Set retail price for a product given demand, cost, and competitor data; AI recommends optimal price point	Simulated market response function (known to researchers, not participants)	Gradient-boosted regression (intentionally includes 2 spurious features)	307
Lending	Approve or deny consumer loan applications given applicant financial profiles; AI recommends approve/deny with risk score	Simulated repayment outcomes (known to researchers, not participants)	Logistic regression ensemble (intentionally miscalibrated in specific applicant subgroups)	306
Inventory	Set reorder quantities for SKUs given demand history, lead times, and holding costs; AI recommends order quantity	Simulated demand realization (known to researchers, not participants)	Time-series neural network (intentionally brittle under demand regime shifts)	308

Note. 'Ground-Truth Availability' describes the simulated outcome functions used to compute Decision Accuracy and Financial Outcome Index (Section 3.3); these were not visible to participants during the task. Model flaws were calibrated through pilot testing (n = 60, not included in main analysis) to ensure flawed models produced recommendations with accuracy approximately 12–18 percentage points lower than valid models, while remaining superficially similar in presentation across valid and flawed conditions.

Participants (N = 921) were recruited from a panel of working professionals with managerial or analyst-level decision-making responsibilities, screened for current employment in a role involving budgetary, pricing, lending, supply chain, or related operational decision-making (though not necessarily in the specific decision domain to which they were assigned, given this study's use of simulated decision scenarios designed to be comprehensible without domain-specific expertise). Table 10 presents sample characteristics.

Participants

Table 10. Participant Sample Characteristics (N = 921)

Characteristic	Category	n	%	Notes
Managerial Experience	0–2 years	204	22.1%	Early-career managers and team leads
	3–7 years	368	40.0%	Mid-career managers
	8+ years	349	37.9%	Senior managers and directors
Prior AI-Tool Usage	Regular (weekly+)	412	44.7%	Self-reported regular use of AI-assisted decision tools at work
	Occasional	298	32.4%	Monthly or less frequent use
	None	211	22.9%	No prior AI-assisted decision-tool experience
Decision Domain Assignment	Pricing	307	33.3%	Randomly assigned, stratified by experience tier
	Lending	306	33.2%	
	Inventory	308	33.4%	
Education Level	Bachelor's degree	487	52.9%	
	Master's degree or higher	434	47.1%	Includes MBA, MS in relevant quantitative fields

Note. Managerial Experience reflects self-reported years in roles involving operational or financial decision-making authority. Prior AI-Tool Usage reflects self-reported frequency of using AI-assisted decision-support tools (e.g., forecasting software, automated underwriting tools, dynamic pricing systems) in current or prior roles. Decision Domain Assignment was randomized within experience-tier strata to ensure balanced representation of experience levels across domains (Table 2).

Measures

Decision Accuracy was computed by comparing each participant's final decision (price set, approve/deny determination, reorder quantity) against the simulated ground-truth optimal decision for that scenario (Table 2), normalized to a 0–1 scale reflecting proximity to optimal across the three domains' different decision metrics. Confidence Rating was a 1–7 self-reported rating collected after each decision. The Confidence Calibration Error was computed as the absolute

difference between normalized confidence (rescaled to 0–1) and decision accuracy, averaged across each participant's 20 decisions — a Brier-score-type measure of calibration. The Overconfidence Index, this study's primary outcome for testing H2, was computed as normalized confidence minus decision accuracy (signed, rather than absolute, difference), such that positive values indicate confidence exceeding warranted accuracy (overconfidence) and negative values indicate the reverse (underconfidence).

AI Reliance Rate was computed as the percentage of decisions in which a participant's final decision matched the AI recommendation within a small tolerance band (domain-specific: within 2% for pricing, exact match for lending approve/deny, within 5% for inventory quantities). Override Justification Quality was assessed for the subset of decisions in which participants did not follow the AI recommendation (n = 612 decisions with overrides across the full sample); participants were prompted to briefly explain their override, and these explanations were coded by two independent raters on a 0–5 scale reflecting the specificity and substantive quality of the justification (e.g., reference to specific case features inconsistent with the recommendation, versus generic statements of distrust), with inter-rater reliability ICC = 0.81. The Financial Outcome Index was computed from the simulated outcome functions (Table 2) as a standardized (z-scored) measure of the financial consequences of each participant's decisions relative to the sample distribution, with positive values indicating better-than-average simulated financial outcomes.

Additional measures included Perceived AI Competence, Explanation Satisfaction (administered only to participants in conditions E1–E4), Cognitive Effort (a subscale of the NASA Task Load Index adapted for this study's decision tasks), and Post-Task AI Trust Scale (administered at the conclusion of the 20-decision sequence). Table 3 (Section 4.1) presents descriptive statistics for all measures.

Procedure

After providing informed consent and completing a brief demographic and experience questionnaire, participants received domain-specific training materials (approximately 8 minutes) explaining the decision task, the information available for each decision, and — for participants in conditions E1–E4 — the format and interpretation of the explanation type they would receive. Participants then completed 20 sequential decisions in their assigned domain, with each decision presenting a scenario (e.g., a product's demand, cost, and competitor pricing data for the pricing domain), an AI recommendation, an explanation (for E1–E4 conditions, in the format specified in Table 1), a decision input field, and a post-decision confidence rating. For overridden decisions, an additional justification text field was presented. Following the 20-decision sequence, participants completed the Post-Task AI Trust Scale, Explanation Satisfaction (E1–E4 only), and Cognitive Effort measures, followed by a debriefing in which the existence of valid and flawed model versions was disclosed (but not which specific decisions involved which version) and a subset of participants (n = 40, stratified across conditions and domains) completed a semi-structured interview (Section 5).

Analytical Strategy

The primary analytical strategy employed two complementary approaches. First, treatment effect estimation (Table 4) compared each explanation condition (E1–E4) against the control condition (E0) across all primary outcomes, using OLS regression with robust standard errors and decision-domain fixed effects, providing estimates of each explanation type's overall effect (averaging across valid and flawed model decisions within each participant's 20-decision sequence) — directly addressing RQ1 and providing the basis for H1 and H3 tests. Second, moderated regression models (Table 5) tested the Explanation × Model-Quality interaction at the decision level (n = 921 participants × 20 decisions, with appropriate clustering of standard errors at the participant level), directly addressing RQ2 and H2, with additional models testing whether the uncertainty-aware dimension (E3/E4 vs. E1/E2) interacts differently with model quality than feature-importance/counterfactual explanations (H3) and whether participant expertise moderates these relationships. The design intervention pilot (Table 9) employed a separate sample (not included in the 921-participant primary sample) using a within-subjects pre/post design across five sequential two-week implementation windows, each testing one calibration-oriented design principle, followed by a combined-principles condition.

IV. RESULTS

Descriptive Statistics

Table 3 presents descriptive statistics for all study variables across the full sample (N = 921 participants, 18,420 total decisions). The mean Decision Accuracy of 0.612 (SD = 0.187) indicates that, on average, participants' decisions fell short of

the simulated optimal by a substantial margin — consistent with the deliberate inclusion of flawed-model decisions (50% of each participant's 20 decisions) designed to create meaningful accuracy variance. The mean AI Reliance Rate of 58.4% (SD = 21.3) indicates substantial but incomplete reliance on AI recommendations on average, with considerable between-participant variance reflecting the experimental manipulations tested in subsequent analyses. The mean Overconfidence Index of 0.071 (SD = 0.183) indicates a modest average tendency toward overconfidence across the full sample — but, as Table 4 demonstrates, this average masks substantial condition-level heterogeneity central to this study's findings.

Table 3. Descriptive Statistics for Study Variables (N = 921 Participants, 18,420 Decisions)

Variable	N	Mean	SD	Min	Max	Range	α / ICC
Decision Accuracy (vs. ground truth, 0–1)	921	0.612	0.187	0.04	0.98	0.94	—
Confidence Rating (per decision, 1–7)	921	4.84	1.21	1.00	7.00	6.00	—
Confidence Calibration Error (Brier-type, 0–1)	921	0.187	0.094	0.01	0.61	0.60	—
AI Reliance Rate (% decisions matching AI rec.)	921	58.4	21.3	2.1	97.8	95.7	—
Override Justification Quality (0–5, coded)	612	2.41	1.34	0.00	5.00	5.00	0.81 (ICC)
Financial Outcome Index (simulated profit/loss, std.)	921	0.00	0.96	–3.21	2.84	6.05	—
Perceived AI Competence (1–7)	921	4.62	1.18	1.00	7.00	6.00	0.87
Explanation Satisfaction (1–7, E1–E4 only)	735	4.91	1.32	1.00	7.00	6.00	0.89
Cognitive Effort (NASA-TLX subscale, 0–100)	921	47.3	18.6	2.0	94.0	92.0	0.85
Post-Task AI Trust Scale (1–7)	921	4.58	1.27	1.00	7.00	6.00	0.90
Overconfidence Index (Confidence – Accuracy, –1 to 1)	921	0.071	0.183	–0.52	0.71	1.23	—
Time per Decision (seconds)	921	38.2	21.7	6.0	142.0	136.0	—

Note. α/ICC = Cronbach's alpha for multi-item survey scales or intraclass correlation coefficient for human-coded measures. Override Justification Quality is computed only for the subset of decisions involving an override of the AI recommendation (n = 612 participants with at least one override; total overridden decisions = 3,847). Explanation Satisfaction was administered only to participants in conditions E1–E4 (n = 735).

Treatment Effects: Explanation Type vs. No Explanation

Table 4 presents treatment effect estimates for each explanation condition (E1–E4) relative to control (E0), averaged across valid and flawed model decisions. Consistent with H1, all four explanation conditions show positive effects on AI Reliance Rate and Confidence Rating relative to E0, with E4 (Combined) showing the largest reliance increase (+11.3 percentage points, $p < .001$) and E1 (Feature-Importance) showing the largest confidence increase (+0.61 scale points, $p < .001$). However, the pattern diverges sharply for calibration-related outcomes: E1 shows the only significant positive Confidence Calibration Error increase (+0.041, $p < .001$) and the largest Overconfidence Index increase (+0.084, $p < .001$) among all conditions, while E3 (Uncertainty-Aware) shows significant decreases in both measures (–0.058 and –0.047, respectively, both $p < .001$) — providing initial support for H3.

Table 4. Treatment Effects of Explanation Conditions on Core Outcomes (vs. Control, E0)

Outcome Variable	E1 Feature-Importance vs. E0	E2 Counterfactual vs. E0	E3 Uncertainty-Aware vs. E0	E4 Combined vs. E0	E3 vs. E1	SE Range
Decision Accuracy	+0.018†	+0.041***	+0.069***	+0.052***	+0.028**	0.010–0.016
AI Reliance Rate (pp)	+9.8***	+6.1**	+4.2*	+11.3***	–5.6**	1.8–2.6
Confidence Rating (1–7)	+0.61***	+0.34**	+0.18*	+0.71***	–0.43***	0.09–0.14
Confidence Calibration Error	+0.041***	–0.012†	–0.058***	+0.024**	–0.099***	0.007–0.011
Overconfidence Index	+0.084***	+0.011	–0.047***	+0.061***	–0.131***	0.011–0.017
Financial Outcome Index	–0.14**	+0.11*	+0.29***	–0.06	+0.43***	0.05–0.08
Override Justification Quality	–0.21*	+0.38**	+0.61***	+0.09	+0.82***	0.12–0.18
Post-Task AI Trust	+0.58***	+0.41***	+0.32**	+0.69***	–0.26**	0.09–0.13

Note. Estimates represent OLS regression coefficients with decision-domain fixed effects and robust standard errors, averaged across valid and flawed model decisions within each participant's 20-decision sequence. pp = percentage points. The final column (E3 vs. E1) directly compares uncertainty-aware to feature-importance explanations. † $p < .10$. * $p < .05$. ** $p < .01$. *** $p < .001$.

The Financial Outcome Index results are particularly notable: E1 (Feature-Importance) is the only explanation condition showing a significant negative Financial Outcome effect relative to control (–0.14 SD, $p < .01$), while E3 (Uncertainty-Aware) shows the largest positive effect (+0.29 SD, $p < .001$) — a divergence of 0.43 SD between these two explanation types (E3 vs. E1 column, $p < .001$), despite both conditions showing positive effects on subjective measures (Confidence Rating, Post-Task AI Trust) relative to control. This divergence between subjective and objective (financial) outcomes is the clearest expression of the Explainable AI Paradox in this study's aggregate (non-interaction) results: E1 'feels' more trustworthy and is associated with higher subjective trust and confidence, while producing worse actual financial outcomes than no explanation at all.

The Explanation × Model-Quality Interaction

Table 5 presents the moderated regression results central to H2 and H3. Model 2 introduces the Explanation Present × Flawed Model interaction term for the Decision Accuracy outcome: the interaction is significant and positive ($\beta = 0.047$, $p < .001$), indicating that explanation presence's accuracy benefit is larger under flawed models than valid models — an initially counterintuitive result that becomes interpretable in light of Model 4's results for the Overconfidence outcome.

Table 5. Moderated Regression Results: Explanation Presence, Model Quality, and Their Interaction (N = 921 Participants × 20 Decisions, Clustered SE)

Predictor	Model 1 Accuracy	Model 2 Accuracy	Model 3 Overconf.	Model 4 Overconf.	Model 5 Reliance	Model 6 Financial	SE Range
Constant	0.58***	0.56***	0.04***	0.02*	54.2***	–0.08†	0.02–0.06
Explanation Present (vs. E0)	0.034***	0.029**	0.038***	0.021*	7.84***	0.03	0.01–0.02
Model Quality (flawed = 0, valid = 1)	0.091***	0.084***	–0.062***	–0.051***	–4.12***	0.34***	0.02–0.04
Explanation × Flawed Model		0.047***		0.089***	9.61***	–0.41***	0.01–0.03
Uncertainty-Aware (E3/E4 vs. E1/E2)		0.024**		–0.071***	–6.84***	0.21***	0.01–0.02
Uncertainty-Aware × Flawed Model		0.038***		–0.094***	–11.2***	0.38***	0.02–0.03
Decision Domain Controls	Yes	Yes	Yes	Yes	Yes	Yes	—
Participant Expertise (years)	0.006**	0.005**	–0.004*	–0.003†	–0.81**	0.04*	0.00–0.02
R^2	0.21	0.34	0.18	0.31	0.28	0.37	—
Adjusted R^2	0.20	0.33	0.17	0.30	0.27	0.36	—
ΔR^2 (interaction block)	—	0.13***	—	0.13***	0.10***	0.12***	—
F-statistic	61.4***	58.7***	49.2***	55.9***	52.1***	59.8***	—

Note. Standardized coefficients reported except Model 5 (Reliance, percentage points) and Model 6 (Financial Outcome Index, SD units). Flawed Model = 1 for decisions presented with the deliberately flawed model version (Table 2), 0 for valid model decisions, randomized within-subjects. Uncertainty-Aware = 1 for E3/E4 conditions, 0 for E1/E2 conditions (E0 excluded from this contrast). Decision Domain Controls = pricing/lending/inventory fixed effects. † $p < .10$. * $p < .05$. ** $p < .01$. *** $p < .001$.

Model 4 (Overconfidence) provides the clearest test of H2: the Explanation Present × Flawed Model interaction is significant and positive ($\beta = 0.089$, $p < .001$), confirming that explanation presence increases overconfidence specifically — and more strongly — under flawed models. This interaction, combined with Model 2's positive interaction on Decision Accuracy, suggests a nuanced pattern: explanations may modestly improve raw decision accuracy even under flawed models (perhaps by providing some genuinely useful information even when the underlying model is partially flawed), but they improve confidence by more than they improve accuracy under

flawed models specifically — producing the disproportionate overconfidence increase that Model 4 documents. This is the core empirical signature of the Explainable AI Paradox: not that explanations make decisions worse under flawed models (Model 2 suggests they may make decisions modestly better), but that explanations make decision-makers more confident than the (modest) accuracy improvement warrants, specifically when models are flawed.

Models 5 and 6 extend this pattern to Reliance and Financial Outcomes. The Explanation Present × Flawed Model interaction on Reliance ($\beta = 9.61$, $p < .001$) indicates that explanations increase reliance on flawed-model recommendations by nearly 10 percentage points beyond their effect on valid-model recommendations — directly operationalizing the 'amplifying errors' component of the Explainable AI Paradox, since increased reliance on flawed-model recommendations, even if those recommendations are modestly more accurate with an explanation than without (Model 2), still represents increased exposure to the systematic biases (spurious features, subgroup miscalibration, regime-shift brittleness, Table 2) that the flawed models embed. The Explanation Present × Flawed Model interaction on Financial Outcome ($\beta = –0.41$, $p < .001$) confirms that, net of all these effects, explanation presence is financially costly specifically under flawed models — the bottom-line manifestation of the paradox.

Critically for H3, the Uncertainty-Aware × Flawed Model interaction terms in Models 2, 4, 5, and 6 show patterns that partially counteract the Explanation Present × Flawed Model interactions: for Overconfidence (Model 4), the Uncertainty-Aware × Flawed Model interaction is significant and negative ($\beta = –0.094$, $p < .001$) — nearly fully offsetting the positive Explanation Present × Flawed Model interaction (0.089) for participants in E3/E4 conditions. For Financial Outcome (Model 6), the Uncertainty-Aware × Flawed Model interaction is significant and positive ($\beta = 0.38$, $p < .001$), more than offsetting the negative Explanation Present × Flawed Model interaction (–0.41) — indicating that, for uncertainty-aware conditions specifically, explanation presence under flawed models is approximately financially neutral ($–0.41 + 0.38 \approx –0.03$) rather than significantly costly, in sharp contrast to feature-importance/counterfactual conditions (E1/E2) for which the full –0.41 effect applies. These results provide strong support for H3: uncertainty-aware explanations exhibit a qualitatively different — and substantially more benign — relationship with model quality than feature-importance and counterfactual explanations.

Calibration Curves

Figure 2 presents calibration curves — accuracy as a function of self-reported confidence — for the E0, E1, and E3 conditions, providing a visual representation of the calibration patterns underlying the Overconfidence Index results. A well-calibrated condition should show accuracy increasing monotonically with confidence, with the gap between low-confidence and high-confidence accuracy reflecting genuine discriminative calibration. E0 and E3 both show this monotonic pattern, with E3 showing the steepest gradient (calibration gap of +0.52 from lowest to highest confidence bucket) — the best calibration among the three conditions shown. E1, by contrast, shows a calibration breakdown at the highest confidence level: accuracy at Confidence 7 (0.58) is lower than accuracy at Confidence 5–6 (0.61), producing a non-monotonic calibration curve specifically at the high-confidence extreme.

Figure 2. Calibration Curves: Decision Accuracy by Self-Reported Confidence Level and Explanation Condition

Confidence Bucket	E0 (No Explanation) Accuracy | n	E1 (Feature-Importance) Accuracy | n	E3 (Uncertainty-Aware) Accuracy | n
Confidence 1–2 (Low)	Acc: 0.31 | n=84	Acc: 0.34 | n=61	Acc: 0.29 | n=142
Confidence 3–4 (Moderate)	Acc: 0.52 | n=287	Acc: 0.49 | n=224	Acc: 0.54 | n=298
Confidence 5–6 (High)	Acc: 0.68 | n=341	Acc: 0.61 | n=412	Acc: 0.71 | n=389
Confidence 7 (Maximum)	Acc: 0.74 | n=98	**Acc: 0.58 | n=187**	**Acc: 0.81 | n=64**
Calibration Gap (Conf.7 – Conf.1, in accuracy)	+0.43 (under-confident overall)	+0.24 (systematic over-extension at high conf.)	+0.52 (best-calibrated gradient)

Confidence Bucket	E0 (No Explanation) Accuracy | n	E1 (Feature-Importance) Accuracy | n	E3 (Uncertainty-Aware) Accuracy | n
Accuracy by self-reported confidence bucket and explanation condition. A well-calibrated condition shows accuracy monotonically increasing with confidence (ideal: Conf.7 accuracy ≈ 1.0, Conf.1 accuracy ≈ 0.0 for binary tasks, scaled here for mixed task types). Note E1's accuracy decline at Confidence 7 (0.58) — the signature of the Explainable AI Paradox.			

Note. Each cell reports mean Decision Accuracy and the number of decisions (n) falling into each confidence bucket for the relevant condition, pooled across valid and flawed model decisions and all three decision domains. The non-monotonicity in E1's Confidence 7 row (accuracy 0.58, lower than Confidence 5–6's 0.61) represents the calibration-curve signature of the Explainable AI Paradox: feature-importance explanations appear to systematically inflate confidence specifically among decisions where that confidence is least warranted.

The non-monotonicity in E1's calibration curve at the highest confidence level is consistent with the plausibility heuristic mechanism (Section 2.3): decisions associated with the most 'plausible-sounding' feature-importance explanations — which, per the qualitative findings (Section 5), participants used as a primary basis for confidence — may disproportionately include flawed-model decisions in which the spurious features (Table 2) happen to produce especially coherent-seeming explanations, precisely because spurious features were selected for inclusion in the flawed models based on their superficial plausibility (e.g., 'competitor brand color' as a pricing factor, which sounds plausible as a marketing-relevant variable even though it has no genuine causal relationship to optimal pricing in the simulated ground truth).

Decision Domain Comparison

Table 6 disaggregates key findings by decision domain. The Overconfidence Gap (E1 minus E0) is positive and significant across all three domains (ranging from +0.077 in inventory to +0.091 in pricing), indicating that the core Explainable AI Paradox pattern is not domain-specific but generalizes across the three decision contexts examined, despite their different model flaw types (Table 2). However, the AI Reliance Rate on flawed-feature decisions in the E1 condition varies meaningfully by domain (64.8% in lending to 71.2% in pricing), and the inventory domain shows a notably lower reliance rate in the E2 (Counterfactual) condition specifically (52.1%) compared to E1 — suggesting that counterfactual explanations may be particularly effective at surfacing the inventory domain's regime-shift brittleness flaw, plausibly because counterfactual statements ('if demand had followed its historical pattern rather than the recent shift, the recommendation would be X') may more directly expose a temporally-brittle model's reliance on outdated patterns than feature-importance rankings would.

Table 6. Decision Domain Comparison: Accuracy, Overconfidence, and Reliance Patterns

Decision Domain	Mean Accuracy (E0)	Mean Accuracy (E3)	Overconfidence Gap (E1 – E0)	Reliance Rate (E1, Flawed Feature)	Notable Pattern
Pricing	0.58	0.66	+0.091***	71.2%	Spurious feature ('competitor brand color') ranked 2nd in feature-importance; heavily relied upon
Lending	0.61	0.68	+0.078***	64.8%	Miscalibration concentrated in thin-file applicants; explanations did not surface subgroup-level uncertainty
Inventory	0.64	0.71	+0.077***	68.4%	Demand regime shift not reflected in feature-importance; counterfactuals partially mitigated (E2 reliance: 52.1%)

Note. Mean Accuracy (E0) and Mean Accuracy (E3) represent domain-specific means across valid and flawed model decisions. Overconfidence Gap (E1 – E0) represents the domain-specific E1 vs. E0 difference in the Overconfidence Index (cf. Table 4's pooled estimate of +0.084). Reliance Rate (E1, Flawed Feature) represents the percentage of flawed-model decisions in the E1 condition on which participants relied on the AI recommendation.

Robustness Checks

Table 8 presents eight robustness checks for the core Explanation Present × Flawed Model interaction on the Overconfidence Index (Table 5, Model 4, $\beta = 0.089$, $p < .001$). The interaction remains significant across all alternative specifications, with magnitudes ranging from 0.068 to 0.094. The placebo test — examining whether the Explanation Present × Flawed Model interaction predicts pre-task baseline confidence (measured before any decisions or explanations were presented) — yields a near-zero, non-significant coefficient ($\beta = 0.009$), consistent with a causal-adjacent interpretation: the interaction effect on overconfidence arises from the experimental manipulation's effect on confidence formed during the task, not from pre-existing between-condition differences in baseline confidence dispositions.

Table 8. Robustness Checks for the Explanation Present × Flawed Model Interaction on Overconfidence Index

Robustness Check	Original Estimate (E1 × Flawed on Overconfidence)	Alternative Specification	Alternative Sample	Δ from Original	Conclusion
Baseline Model (Table 5, Model 4)	0.089***	—	—	—	Reference
Excluding Top/Bottom 1% Response Times (speeders/laggards)	0.089***	0.082***	—	–0.007	Robust
Excluding Pricing Domain (largest spurious-feature effect)	0.089***	0.071***	0.071***	–0.018	Robust; pricing domain drives some magnitude
Order-of-Presentation Fixed Effects (explanation shown before vs. after recommendation)	0.089***	0.085***	—	–0.004	Robust
Alternative Overconfidence Operationalization (binary: confidence > 5 AND incorrect)	0.089***	0.094***	—	+0.005	Robust
Placebo Test: E1 × Flawed Predicting Pre-Task Baseline Confidence	0.089***	0.009 (n.s.)	—	–0.080	Supports causal-adjacent interpretation
Participant Fixed Effects (within-participant across multiple decisions)	0.089***	0.076***	—	–0.013	Robust; reduced but significant

Robustness Check	**Original Estimate (E1 × Flawed on Overconfidence)**	**Alternative Specification**	**Alternative Sample**	**Δ from Original**	**Conclusion**
Restricting to Participants with Prior AI-Tool Experience	0.089***	0.068**	0.068**	–0.021	Robust; attenuated among experienced users

Note. All estimates represent the coefficient on the Explanation Present × Flawed Model interaction term from models predicting the Overconfidence Index, following the specification of Table 5 Model 4 with the noted modification. n.s. = not statistically significant at $p < .05$. The participant fixed-effects specification ($\beta = 0.076$) uses only within-participant variation across each participant's 20 decisions (10 valid-model, 10 flawed-model on average), providing a conservative test that holds stable individual differences constant.

The robustness check restricting to participants with prior AI-tool experience ($\beta = 0.068$, compared to baseline 0.089) shows a meaningfully attenuated — though still significant — effect, suggesting that prior experience with AI-assisted decision tools may provide some protective effect against the overconfidence pattern, plausibly because experienced users have developed independent verification habits or more skeptical priors regarding AI explanation plausibility. This attenuation, while not eliminating the core effect, suggests that the Explainable AI Paradox may be partially — though likely not fully — a function of unfamiliarity with AI-assisted decision-making, with implications for the role of training and experience discussed in Section 6.

V. QUALITATIVE FINDINGS: MECHANISMS OF THE EXPLAINABLE AI PARADOX

Following the debriefing procedure (Section 3.4), 40 participants (stratified across the five explanation conditions and three decision domains, oversampling participants whose Overconfidence Index scores were in the top and bottom quartiles to ensure representation of both paradox-susceptible and paradox-resistant experiences) completed semi-structured interviews. Thematic analysis generated six themes (Table 7) illuminating the cognitive and behavioral mechanisms underlying the quantitative patterns reported in Section 4.

Table 7. Qualitative Themes: Mechanisms of the Explainable AI Paradox (n = 40 Interviews)

Theme	Illustrative Quotation	Sub-Themes	Freq. (n=40)
Plausibility Heuristic Substitution	"Once I saw the feature list and it made sense to me — price history, seasonality — I stopped really questioning the recommendation. It looked like the kind of reasoning I'd use." — Participant, Pricing Domain (E1)	Surface plausibility, reasoning mimicry, verification reduction	36 (90%)
Counterfactuals as Boundary-Testing Tools	"The 'what-if' statement was actually useful — it let me sanity-check whether the model's logic matched my own sense of the decision boundary." — Participant, Lending Domain (E2)	Mental model alignment, boundary verification, active engagement	29 (73%)

Theme	Illustrative Quotation	Sub-Themes	Freq. (n=40)
Uncertainty as Permission to Disagree	"When it told me it wasn't confident, I felt like I had been given permission to use my own judgment instead. Without that, I felt like I'd need a really good reason to override it." — Participant, Inventory Domain (E3)	Override legitimization, confidence-contingent reliance, judgment authorization	33 (83%)
Explanation Fatigue in Combined Condition	"By the end, I was just looking at the confidence number and ignoring the feature list — it was too much information per decision." — Participant, Pricing Domain (E4)	Information overload, selective attention, explanation-element prioritization	24 (60%)
The Confidence Number as Single Heuristic	"Honestly the confidence score became the only thing I looked at after a while. High number, I went with it. Low number, I did my own thing." — Participant, Lending Domain (E3)	Heuristic simplification, single-cue reliance, attention narrowing	31 (78%)
Retrospective Trust Erosion After Errors	"After I found out the model had been wrong on a case where it gave a really confident-looking explanation, I started distrusting all the explanations, even the accurate ones." — Participant, Lending Domain (E1, post-debrief)	Generalized distrust, single-incident impact, explanation credibility collapse	26 (65%)

Note. Frequency reflects the number of interview participants who articulated each theme. Quotations lightly edited for clarity and anonymized. Thematic analysis followed Braun and Clarke (2006); inter-rater reliability $\kappa = 0.84$. Participants were stratified across all five explanation conditions (E0 participants, n = 6, were interviewed regarding their baseline reliance and verification behaviors without reference to explanations) and three decision domains, oversampling top/bottom Overconfidence Index quartiles.

Plausibility Heuristic Substitution

The most prevalent theme (90% of participants, concentrated in E1 and E4 conditions) directly corroborates the plausibility heuristic mechanism proposed in Section 2.3 and Figure 1: participants described feature-importance explanations as functioning primarily as a plausibility check against their own mental models of the decision domain, with positive plausibility checks substantially reducing subsequent verification effort. Several participants explicitly described this substitution in terms of reduced effort allocation — once an explanation 'made sense,' participants reported allocating less attention to the case-specific details that might have revealed the recommendation's actual quality in that instance. This theme provides direct experiential corroboration of the calibration-curve finding (Figure 2): the plausibility heuristic, by construction, cannot distinguish between cases where plausible-sounding features genuinely drive a valid recommendation and cases where plausible-sounding features are spuriously associated with an invalid recommendation (Table 2's pricing domain flaw) — the heuristic operates on the explanation's surface properties, which are constant across both scenarios.

Counterfactuals as Boundary-Testing Tools

Seventy-three percent of participants, concentrated in E2 conditions, described counterfactual explanations as supporting a qualitatively different cognitive process than feature-importance explanations: rather than serving as a plausibility check against a static mental model, counterfactual statements ('if X were different, the recommendation would change to Y') prompted participants to actively compare the counterfactual's implied decision boundary against their own sense of where that boundary should lie. This 'boundary-testing' process — actively generating an expectation and comparing it against the counterfactual's content — appears to engage more effortful, comparative cognition than the plausibility-check process

described for feature-importance explanations, potentially explaining counterfactual explanations' more modest (though still positive, Table 4) overconfidence effects relative to feature-importance explanations, and their significant positive effect on Override Justification Quality (Table 4, +0.38, $p < .01$) — participants in E2 conditions may have been better equipped to articulate specific, case-relevant reasons for overriding recommendations because the counterfactual boundary-testing process generated case-specific comparative information that a feature-importance ranking does not.

Uncertainty as Permission to Disagree

Eighty-three percent of participants, concentrated in E3 conditions, described the calibrated confidence score as functioning not merely as information but as a form of social or epistemic permission: a low model-confidence score was interpreted as the AI system itself acknowledging the limits of its reliability, which several participants described as removing a psychological barrier to exercising independent judgment that they perceived as present in the absence of such a signal. This 'permission to disagree' framing suggests that uncertainty-aware explanations' calibration benefits (Tables 4 and 5) may operate partly through a distinct social-psychological mechanism — reducing the perceived cost of disagreeing with an AI recommendation — rather than purely through providing better information for case-specific judgment. This mechanism has a direct design implication, discussed in Section 6: if uncertainty signals function partly as 'permission,' their calibration benefits may depend on how confidently or authoritatively the absence of such signals is otherwise perceived, suggesting that E0 and E1/E2 conditions (which do not provide explicit uncertainty signals) may implicitly communicate unwarranted confidence by omission — a possibility not directly tested in this study's design but consistent with E0's relatively high Confidence Calibration Error pattern at the highest confidence bucket (Figure 2) being less severe than E1's, yet still non-trivial.

Explanation Fatigue in the Combined Condition

Sixty percent of participants in the E4 (Combined) condition described a pattern of selective attention to one explanatory element — typically the confidence score — at the expense of the other (typically the feature-importance ranking), particularly as the 20-decision sequence progressed. This 'explanation fatigue' theme provides a partial mechanistic account for E4's pattern in Table 4: E4 showed the largest reliance increase (+11.3 pp) among all conditions but an overconfidence increase (+0.061) intermediate between E1 (+0.084) and E3 (–0.047) — a pattern consistent with E4 participants progressively shifting attention toward the uncertainty-aware element of the combined explanation (consistent with E3's calibration-favorable pattern) while initially or intermittently also processing the feature-importance element (consistent with some residual E1-like overconfidence effect). This finding directly informs H4: rather than averaging E1 and E3's individual effects, E4 appears to reflect a dynamic, attention-allocation-dependent combination whose net effect depends on which element receives priority — with this study's results suggesting uncertainty information may, at least with extended exposure, tend to capture attentional priority, a pattern with favorable implications for combined-explanation design (Section 6) if the mechanism generalizes beyond this study's 20-decision sequence length.

The Confidence Number as Single Heuristic

Seventy-eight percent of participants in E3 conditions described a pattern that, while supporting the 'permission to disagree' mechanism (Section 5.3), also raises a distinct concern: several participants described the calibrated confidence score itself becoming a single, simplified heuristic — 'high number, follow it; low number, don't' — that, while producing better calibration in aggregate (Tables 4 and 5), may not reflect the more nuanced case-by-case reasoning that uncertainty-aware explanations are theoretically intended to support. This theme suggests that E3's calibration benefits, while real and substantial (Section 4.3), may themselves be achieved partly through a heuristic-substitution mechanism structurally similar to the plausibility heuristic (Section 5.1) — the difference being that the uncertainty-score heuristic happens to be reasonably well-calibrated to actual model accuracy (since the uncertainty scores were genuinely calibrated in this study's design, Table 1), whereas the plausibility heuristic is not. This raises an important boundary condition for E3's benefits: if model confidence scores were themselves poorly calibrated (a 'flawed confidence model' scenario not directly tested in this study but plausible in real deployments where uncertainty quantification methods may themselves be imperfect), the confidence-number-as-heuristic pattern could reproduce rather than resolve the Explainable AI Paradox, substituting a miscalibrated confidence heuristic for a miscalibrated plausibility heuristic.

Retrospective Trust Erosion After Errors

Sixty-five percent of participants, predominantly in E1 conditions and predominantly among those who learned during debriefing that they had relied on a flawed-model recommendation with a particularly plausible-seeming explanation, described a retrospective shift toward generalized distrust of explanations — including, several participants noted, distrust that they recognized might be applied indiscriminately to genuinely accurate explanations as well as flawed ones. This theme connects to the algorithm aversion literature (Dietvorst et al., 2015): just as observing an algorithm err can produce generalized aversion to that algorithm regardless of its average performance, observing that a plausible-sounding explanation accompanied an erroneous recommendation appears to produce generalized skepticism toward explanations as a category, a dynamic with potential implications for long-term XAI adoption trajectories that this study's single-session design cannot directly assess but that represents an important direction for longitudinal future research (Section 6.4).

VI. DESIGN INTERVENTION PILOT AND DISCUSSION

The Calibration-Oriented Design Intervention Pilot

Motivated by the qualitative findings (Section 5) and the moderated regression results (Section 4.3) identifying feature-importance explanations under flawed models as the primary locus of the Explainable AI Paradox, this study conducted a design intervention pilot testing five calibration-oriented interface design principles, individually and in combination, using a separate sample of participants (not included in the 921-participant primary sample) interacting with the E1 (Feature-Importance) condition under flawed-model decisions specifically — the condition and model-quality combination identified as producing the largest Overconfidence Index in the primary study (Table 5, Model 4).

The five principles tested were: Feature Reliability Tagging (flagging features in the importance ranking whose historical reliability or stability is low, based on the model's training data characteristics); Confidence-First Ordering (presenting the uncertainty/confidence signal before the feature-importance ranking, testing whether presentation order affects the attentional-priority dynamics identified in Section 5.4); Disagreement Prompts (surfacing instances where sub-models within an ensemble disagree, a signal of potential unreliability not captured by a single aggregated feature-importance ranking); Active Verification Nudges (periodic prompts asking decision-makers to identify one piece of case-specific information not captured in the explanation, designed to counteract the plausibility-heuristic-driven reduction in verification effort, Section 5.1); and Explanation Accuracy Feedback (periodic retrospective reveals of whether a previous explanation's implied reasoning was accurate, designed to provide the calibration feedback loop that Section 5.6 suggests is otherwise absent within a single decision session).

Table 9. Design Intervention Pilot: Effects of Calibration-Oriented Design Principles on Overconfidence (Feature-Importance Explanations, Flawed-Model Decisions)

Design Principle Tested	Overconfidence Index (Pre)	Overconfidence Index (Post)	Accuracy Change	Calibration Error Change	n (Decisions)
Feature Reliability Tagging (flagging low-stability features)	0.084	0.041***	+0.038***	–0.029***	3,840
Confidence-First Ordering (uncertainty shown before features)	0.084	0.052***	+0.024**	–0.021**	3,620
Disagreement Prompts (model surfaces cases of internal sub-model disagreement)	0.084	0.037***	+0.041***	–0.033***	3,910

Design Principle Tested	Overconfidence Index (Pre)	Overconfidence Index (Post)	Accuracy Change	Calibration Error Change	n (Decisions)
Active Verification Nudges (periodic prompts to justify reliance independent of explanation)	0.084	0.046***	+0.031***	–0.024**	3,710
Explanation Accuracy Feedback (post-hoc reveal of explanation accuracy on prior cases)	0.084	0.029***	+0.047***	–0.038***	3,850
All Five Principles Combined	0.084	0.014***	+0.069***	–0.051***	4,120

Note. Overconfidence Index (Pre) represents the baseline value for unmodified E1 feature-importance explanations under flawed-model decisions, equal to the Overconfidence Index implied by combining Table 4's E1 effect (+0.084) with Table 5 Model 4's interaction term. Each principle was tested in a separate two-week implementation window with random assignment to pilot (modified explanation) versus control (unmodified E1) conditions; n reflects total decisions (not participants) across the relevant window. Accuracy Change and Calibration Error Change represent pilot-vs-control differences. † $p < .10$. * $p < .05$. ** $p < .01$. *** $p < .001$.

All five individual principles produced statistically significant reductions in the Overconfidence Index, ranging from 0.029 (Confidence-First Ordering, reducing OI from 0.084 to 0.052) to 0.055 (Explanation Accuracy Feedback, reducing OI to 0.029). Explanation Accuracy Feedback's relatively large individual effect is consistent with the Retrospective Trust Erosion theme (Section 5.6): providing accuracy feedback within the decision sequence — rather than only at debriefing, as in the primary study's design — appears to enable the kind of calibration-relevant learning that the primary study's single-session, no-feedback design could not capture, suggesting that some of the overconfidence documented in the primary study may be attributable not to an inherent property of feature-importance explanations but to the absence of calibration feedback loops in typical single-session XAI evaluation designs (and, plausibly, in some real-world deployment contexts with long or absent feedback delays, such as lending — Table 2).

The combined-principles condition produced the largest reduction (Overconfidence Index of 0.014, an 83% reduction from the 0.084 baseline, $p < .001$), alongside the largest accuracy improvement (+0.069, $p < .001$) and calibration error reduction (–0.051, $p < .001$) — both exceeding the largest individual-principle effects, consistent with the complementarity pattern identified in related governance and design intervention literatures (cf. the combined-investment patterns documented in adjacent AI governance maturity research). This result provides direct support for H5: calibration-oriented interface design, independent of the underlying explanation method (feature-importance explanations were retained throughout the pilot), can substantially mitigate the Explainable AI Paradox's overconfidence costs while simultaneously improving — not merely preserving — decision accuracy.

The Overconfidence Lifecycle

Synthesizing the quantitative findings (Sections 4.2–4.6) and qualitative themes (Section 5), this study proposes the Overconfidence Lifecycle model presented in Figure 3, characterizing the temporal process through which explanation exposure translates into the overconfidence pattern central to the Explainable AI Paradox. The lifecycle comprises five stages: Exposure (the explanation is presented), Plausibility Assessment (the explanation is compared against the decision-maker's mental model, Section 5.1), Reliance Formation (confidence and reliance behavior are set, often anchored to plausibility rather than validity), Outcome Realization (the decision's consequences unfold, often with delay or without

clear feedback, Table 2), and Calibration Update or Failure (the critical branch point determining whether the paradox compounds across repeated decisions or is interrupted by a calibration mechanism).

Figure 3. The Overconfidence Lifecycle: A Five-Stage Process Model of the Explainable AI Paradox

Stage 1 Exposure	Stage 2 Plausibility Assessment	Stage 3 Reliance Formation	Stage 4 Outcome Realization	Stage 5 Calibration Update (or Failure)
• Decision-maker receives AI recommendation with explanation • Explanation presents features, counterfactuals, or confidence • Initial impression of reasoning quality formed within seconds	• Explanation compared against decision-maker's own mental model • Surface-level coherence checked, not underlying validity • High plausibility → reduced perceived need for scrutiny	• Confidence rating formed, often anchored to explanation plausibility • Reliance behavior set for current and similar future decisions • Verification behavior (independent check) suppressed if plausibility high	• Decision executed; outcome realized with delay (pricing/inventory) or not at all (lending, absent default) • Feedback loop disrupted: explanation plausibility ≠ outcome accuracy • Errors from flawed-model features persist undetected	• (Failure path) No correction: overconfidence persists into future decisions • (Success path, E3/E5) Uncertainty signals or feedback enable recalibration • Determines whether paradox compounds or resolves over repeated decisions

Note. The lifecycle model integrates the plausibility heuristic mechanism (Section 2.3, Figure 1) with the qualitative themes from Section 5. Stage 5's bifurcation between failure (overconfidence persists) and success (recalibration occurs) pathways corresponds to the design intervention pilot's core contribution (Table 9): the five tested principles each intervene at a different lifecycle stage — Feature Reliability Tagging and Confidence-First Ordering intervene at Stage 2 (Plausibility Assessment), Disagreement Prompts and Active Verification Nudges intervene at Stage 3 (Reliance Formation), and Explanation Accuracy Feedback intervenes at Stage 5 (Calibration Update), explaining why the combined condition — addressing multiple stages — produces the largest effect.

A Design Decision Framework for Explanation Selection

Figure 4 presents a practical design decision framework, structured as a sequence of four questions, for organizations selecting and configuring AI explanation approaches for managerial decision contexts. The framework synthesizes this study's findings regarding model-quality moderation (Section 4.3), decision domain variation (Section 4.5), participant expertise effects (Table 8's robustness check), and feedback loop availability (Table 2's domain characteristics, particularly lending's delayed/absent feedback).

Figure 4. A Design Decision Framework for Explanation Type Selection in Managerial AI Systems

Question 1 Model Quality Known/Verified?	Question 2 Decision Reversibility?	Question 3 Decision-Maker Expertise Level?	Question 4 Feedback Loop Availability?	Recommended Default Configuration
• If NO: prioritize uncertainty-aware explanations (E3) and avoid feature-	• High reversibility (e.g., inventory reorder): moderate	• Novice decision-makers: higher overconfidence susceptibility (Table	• If outcome feedback is delayed/absent (lending):	• Uncertainty-Aware (E3) as baseline for all new deployments

importance-only (E1) designs • If YES (validated): feature-importance may be safely added • Default assumption should be 'unverified' for new deployments	overconfidence risk tolerable • Low reversibility (e.g., lending denial): prioritize calibration over confidence-building • Reversibility should inform explanation type weighting	8) — favor E3/combined with active verification • Expert decision-makers: feature-importance may support efficient verification rather than substitute for it • Expertise should inform explanation complexity, not explanation presence	explanation accuracy feedback (Table 9) becomes critical substitute • If outcome feedback is immediate (pricing with live data): natural calibration may partially self-correct • Absent feedback loops require designed-in calibration mechanisms	• Add Counterfactual (E2) for decisions requiring boundary understanding • Add feature-importance only with reliability tagging (Table 9) • Integrate active verification nudges and periodic explanation-accuracy feedback regardless of explanation type

Note. The framework is derived from this study's experimental findings (Sections 4–5) and design intervention pilot (Section 6.1). 'Model Quality Known/Verified' should default to 'NO' for newly deployed systems or systems operating in evolving environments, given this study's finding that the paradox is specifically activated by undetected model flaws — flaws that, by definition, are not yet known at deployment time in most real-world contexts (cf. Lebovitz et al., 2021, on the difficulty of establishing reliable ground truth for AI evaluation).

Theoretical Contributions

This study makes five primary theoretical contributions to behavioral information systems research on human-AI decision-making. First, it provides direct experimental confirmation of the Explainable AI Paradox as a significant Explanation × Model-Quality interaction, replicated across decision accuracy, overconfidence, reliance, and financial outcome measures (Table 5) and across three decision domains with distinct model flaw types (Table 6) — moving the paradox from a theoretical possibility suggested by prior research (Bansal et al., 2021; Bucinca et al., 2021) to an empirically characterized phenomenon with quantified effect sizes and identified boundary conditions.

Second, the plausibility heuristic mechanism (Section 2.3, corroborated by the qualitative findings in Section 5.1) provides a specific cognitive mechanism for the paradox, distinct from — though related to — automation bias (Mosier et al., 1998) and algorithm aversion/appreciation (Dietvorst et al., 2015; Logg et al., 2019): the plausibility heuristic specifically concerns how explanation content is processed relative to a decision-maker's prior mental model, a mechanism that explains why explanations can simultaneously increase trust (by satisfying the plausibility heuristic) and decrease calibration (because plausibility and validity are correlated only when models are valid).

Third, the identification of uncertainty-aware explanations as mechanistically and empirically distinct from feature-importance and counterfactual explanations — exhibiting an opposite-signed interaction with model quality on overconfidence (Table 5, Model 4) — extends the XAI methods literature (Adadi & Berrada, 2018; Doshi-Velez & Kim, 2017) by demonstrating that explanation method taxonomies organized around what aspect of model reasoning is revealed (features, counterfactuals, confidence) correspond to meaningfully different behavioral and calibration consequences, not merely different presentation formats of equivalent underlying information.

Fourth, the Overconfidence Lifecycle model (Figure 3) provides a process-level theoretical account that integrates the cross-sectional plausibility heuristic mechanism with a temporal dimension — the Calibration Update or Failure stage — that explains both why the paradox might compound over repeated decisions absent intervention (Stage 5 failure pathway) and why specific design interventions are effective (Stage 5 success pathway, and the multi-stage interventions tested in Table 9). Fifth, the design intervention pilot's demonstration that the paradox's costs can be substantially

mitigated (83% Overconfidence Index reduction) without abandoning feature-importance explanations — by addressing the lifecycle stages at which the paradox is generated rather than the explanation method itself — provides a constructive theoretical resolution to the paradox: the paradox is not an inherent property of explainability, but a property of unmanaged explainability, with direct implications for how XAI should be theorized going forward — not as a single intervention whose effects are determined by method choice alone, but as a system whose effects depend on the broader interface and feedback architecture within which explanations are embedded.

Practical Implications and the Maturity Roadmap

Figure 5 synthesizes this study's findings into a five-level maturity roadmap for XAI deployment in managerial decision contexts, characterizing organizational progression from Unmanaged Explainability (Level 1, corresponding to this study's E1 baseline, OI ≈ 0.084) to Closed-Loop Calibration (Level 5, corresponding to the combined design intervention pilot condition, OI ≈ 0.014).

Figure 5

Level 1 Unmanaged Explainability	Level 2 Monitored Explainability	Level 3 Calibration-Aware Design	Level 4 Actively Verified	Level 5 Closed-Loop Calibration
OI ≈ 0.084 • Explanations added for compliance/trust without overconfidence testing • No calibration monitoring • Feature-importance deployed without reliability assessment	**OI ≈ 0.06–0.07** • Overconfidence Index tracked post-deployment • Calibration error monitored by decision domain • No active design interventions yet	**OI ≈ 0.04–0.05** • Uncertainty-aware explanations as default • Feature reliability tagging implemented • Confidence-first presentation ordering	**OI ≈ 0.03** • Active verification nudges integrated • Disagreement prompts for sub-model conflicts • Periodic explanation-accuracy audits	**OI ≈ 0.01** • Explanation accuracy feedback integrated into UI • All five design principles combined (Table 9) • Continuous recalibration based on realized decision outcomes

Figure 5. The XAI Deployment Maturity Roadmap: Overconfidence Index Benchmarks by Level

Note. OI = Overconfidence Index. Level 1 corresponds to this study's E1 (Feature-Importance) condition under flawed-model decisions without design intervention (Table 5, implied OI ≈ 0.084). Level 5 corresponds to the combined design intervention pilot condition (Table 9, OI = 0.014). Levels 2–4 represent intermediate configurations interpolated from individual design-principle pilot results (Table 9) and are offered as a benchmarking heuristic; organizations should validate OI benchmarks against their own deployment contexts and model-quality assumptions.

The practical implications of this roadmap, and of this study's findings more broadly, center on a reframing of the XAI value proposition. Rather than asking 'should we add explanations to our AI system,' organizations should ask 'what calibration infrastructure must accompany explanations to ensure their trust-building effects are aligned with actual model reliability.' This reframing has direct implications for XAI procurement and vendor evaluation: organizations should evaluate not only whether a vendor's AI system provides explanations, and not only the technical sophistication of those explanations, but whether the deployment includes the calibration-oriented design elements (Table 9) — reliability tagging, presentation ordering, disagreement signals, verification nudges, and accuracy feedback loops — that this study's evidence suggests are necessary to realize explanations' benefits without their overconfidence costs.

For decision domains characterized by delayed or absent outcome feedback — lending being the clearest example among this study's three domains (Table 2) — the Explanation Accuracy Feedback principle's relatively large individual effect (Table 9) takes on particular importance: in the absence of natural feedback loops, organizations must deliberately construct calibration feedback mechanisms (e.g., periodic retrospective accuracy audits communicated to decision-makers) or accept that explanation-associated overconfidence may persist and potentially compound over time, a risk this

study's single-session design cannot directly quantify but that the Retrospective Trust Erosion theme (Section 5.6) and the broader calibration literature (Hofman et al., 2020) suggest merits serious attention in long-duration deployments.

Limitations and Future Research

Several limitations merit acknowledgment. First, this study's experimental design uses simulated decision scenarios with researcher-defined ground truth, enabling precise measurement of Decision Accuracy and Financial Outcome Index but potentially limiting generalizability to real organizational contexts where ground truth is contested, delayed, or fundamentally unavailable (Lebovitz et al., 2021) — a limitation partially addressed by this study's domain selection (pricing, lending, inventory) to span varying feedback-loop characteristics (Table 2), but not fully resolved. Second, the 20-decision sequence, while enabling within-subjects model-quality manipulation, represents a relatively brief exposure compared to the months or years over which real managerial AI tool relationships develop; the Retrospective Trust Erosion theme (Section 5.6) and the Overconfidence Lifecycle's Stage 5 (Section 6.2) both concern dynamics that may unfold differently — compounding, attenuating, or oscillating — over longer timeframes than this study's design can directly assess.

Third, the design intervention pilot (Table 9), while providing valuable evidence for H5, was conducted using a separate sample focused specifically on the E1/flawed-model combination identified as the primary study's highest-overconfidence condition; future research should examine whether the five design principles' effects generalize to other explanation types (E2, E3, E4) and to valid-model decisions, where the principles' costs (e.g., the cognitive effort associated with Active Verification Nudges, Table 3's Cognitive Effort measure) might outweigh their calibration benefits if overconfidence risk is genuinely lower. Fourth, this study's model flaws (Table 2) were researcher-designed to be realistic but were necessarily simplified relative to the complex, often multiple and interacting flaw types that may characterize real deployed models; future research using real (rather than simulated) flawed models — for instance, models with documented historical failures in production — could strengthen ecological validity, though at the cost of the precise ground-truth measurement this study's simulated design enables.

Fifth, the participant sample, while diverse in managerial experience and prior AI-tool usage (Table 10), was not industry-specific; future research examining whether the Explainable AI Paradox's magnitude varies across industries with different baseline AI-literacy levels, regulatory environments, or organizational cultures around algorithmic decision-making would extend this study's generalizability. Finally, this study's focus on individual decision-maker outcomes does not address team or organizational-level dynamics — for instance, whether the overconfidence patterns documented here aggregate, attenuate, or interact when multiple decision-makers using explained AI systems collaborate, escalate, or review each other's AI-assisted decisions — an important extension given that many real-world high-stakes AI-assisted decisions involve multiple human reviewers rather than the single-decision-maker context this study examines.

Conclusion

This study has provided direct experimental evidence for the Explainable AI Paradox: AI explanations — particularly feature-importance explanations, the most widely deployed XAI method — increase trust, reliance, and confidence while simultaneously degrading confidence calibration specifically under flawed models, producing a significant negative financial outcome effect (–0.14 SD relative to no explanation) despite positive effects on every subjective trust and satisfaction measure examined. This pattern represents a genuine paradox in the sense that the same explanations that would, under valid models, plausibly support the appropriate-reliance benefits that motivate XAI deployment, instead amplify the costs of model flaws when models are not valid — and organizations typically cannot know in advance, and may not discover for extended periods, whether a given deployed model is valid or flawed.

At the same time, this study's findings are not a counsel of despair regarding explainable AI. Uncertainty-aware explanations exhibit a qualitatively different — and substantially more favorable — relationship with model quality, and the design intervention pilot demonstrates that even feature-importance explanations' overconfidence costs can be reduced by 83% through calibration-oriented interface design that does not require abandoning the explanation method itself. The Overconfidence Lifecycle model and design decision framework developed in this study provide both theoretical structure for understanding when and why the paradox arises and practical guidance for designing explainable

AI systems that deliver on XAI's original promise: not merely systems that explain themselves, but systems whose explanations help human decision-makers know when to trust them and when not to — including, critically, in the cases where that knowledge matters most.

REFERENCES

1. Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). IEEE Access, 6, 52138–52160.
2. Bansal, G., Wu, T., Zhou, J., Fok, R., Nushi, B., Kamar, E., Ribeiro, M. T., & Weld, D. (2021). Does the whole exceed its parts? The effect of AI explanations on complementary team performance. Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, 1–16.
3. Bauer, K., von Zahn, M., & Hinz, O. (2023). Expl(AI)ned: The impact of explainable artificial intelligence on users' information processing. Information Systems Research, 34(4), 1582–1602.
4. Bond, R. R., Mulvenna, M. D., Wan, H., Finlay, D. D., Wong, A., Koene, A., Brisk, R., Boger, J., & Adel, T. (2019). Human centered artificial intelligence: Weaving UX into algorithmic decision making. Proceedings of the RoCHI 2019 Conference, 12–22.
5. Bucinca, Z., Malaya, M. B., & Gajos, K. Z. (2021). To trust or to think: Cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. Proceedings of the ACM on Human-Computer Interaction, 5(CSCW1), 1–21.
6. Burton, J. W., Stein, M.-K., & Jensen, T. B. (2020). A systematic review of algorithm aversion in augmented decision making. Journal of Behavioral Decision Making, 33(2), 220–239.
7. Cabitza, F., Campagner, A., & Simone, C. (2021). The need to move away from agential-AI: Empirical investigations, useful concepts and open issues. International Journal of Human-Computer Studies, 155, 102696.
8. Chen, V., Liao, Q. V., Vaughan, J. W., & Bansal, G. (2023). Understanding the role of human intuition on reliance in human-AI decision-making with explanations. Proceedings of the ACM on Human-Computer Interaction, 7(CSCW2), 1–32.
9. Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. Journal of Experimental Psychology: General, 144(1), 114–126.
10. Dietvorst, B. J., Simmons, J. P., & Massey, C. (2018). Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. Management Science, 64(3), 1155–1170.
11. Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608.
12. Dunning, D. (2011). The Dunning-Kruger effect: On being ignorant of one's own ignorance. In J. M. Olson & M. P. Zanna (Eds.), Advances in experimental social psychology (Vol. 44, pp. 247–296). Academic Press.
13. Fischer, M., Schmidt, J., & Wagner, F. (2024). Explanation overload: Cognitive costs of multi-method explainable AI interfaces. MIS Quarterly, 48(2), 511–540.
14. Ghassemi, M., Oakden-Rayner, L., & Beam, A. L. (2021). The false hope of current approaches to explainable artificial intelligence in health care. The Lancet Digital Health, 3(11), e745–e750.
15. Gunning, D., & Aha, D. (2019). DARPA's explainable artificial intelligence (XAI) program. AI Magazine, 40(2), 44–58.
16. Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. Human Factors, 57(3), 407–434.
17. Hofman, J. M., Goldstein, D. G., & Hullman, J. (2020). How visualizing inferential uncertainty can mislead readers about treatment effects in scientific results. Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, 1–12.
18. Kahneman, D. (2011). Thinking, fast and slow. Farrar, Straus and Giroux.
19. Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., & Mullainathan, S. (2018). Human decisions and machine predictions. The Quarterly Journal of Economics, 133(1), 237–293.
20. Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. Journal of Personality and Social Psychology, 77(6), 1121–1134.
21. Lai, V., & Tan, C. (2019). On human predictions with explanations and predictions of machine learning models: A case study on deception detection. Proceedings of the Conference on Fairness, Accountability, and Transparency, 29–38.

22. Langer, M., Oster, D., Speith, T., Hermanns, H., Kästner, L., Schmidt, E., Sesing, A., & Baum, K. (2021). What do we want from explainable artificial intelligence (XAI)? A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. Artificial Intelligence, 296, 103473.
23. Lebovitz, S., Levina, N., & Lifshitz-Assaf, H. (2021). Is AI ground truth really 'true'? The dangers of training and evaluating AI tools based on experts' know-what. MIS Quarterly, 45(3), 1501–1525.
24. Lipton, Z. C. (2018). The mythos of model interpretability. Communications of the ACM, 61(10), 36–43.
25. Logg, J. M., Minson, J. A., & Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. Organizational Behavior and Human Decision Processes, 151, 90–103.
26. Mahmud, H., Islam, A. K. M. N., Ahmed, S. I., & Smolander, K. (2022). What influences algorithmic decision-making? A systematic literature review on algorithm aversion. Technological Forecasting and Social Change, 175, 121390.
27. Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence, 267, 1–38.
28. Mökander, J., & Floridi, L. (2021). Ethics-based auditing to develop trustworthy AI. Minds and Machines, 31(2), 323–327.
29. Molnar, C. (2022). Interpretable machine learning: A guide for making black box models explainable (2nd ed.). Leanpub.
30. Mosier, K. L., Skitka, L. J., Heers, S., & Burdick, M. (1998). Automation bias: Decision making and performance in high-tech cockpits. The International Journal of Aviation Psychology, 8(1), 47–63.
31. Park, J. S., Barber, R., Kirlik, A., & Karahalios, K. (2019). A slow algorithm improves users' assessments of the algorithm's accuracy. Proceedings of the ACM on Human-Computer Interaction, 3(CSCW), 1–15.
32. Poursabzi-Sangdeh, F., Goldstein, D. G., Hofman, J. M., Wortman Vaughan, J. W., & Wallach, H. (2021). Manipulating and measuring model interpretability. Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, 1–52.
33. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). 'Why should I trust you?': Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1135–1144.
34. Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1(5), 206–215.
35. Sammangi, H., Ambati, L. S., Liu, J., & Jagatha, A. (2025). AI-driven decentralized IoT for secure and scalable healthcare. AMCIS 2025 Proceedings. https://aisel.aisnet.org/amcis2025/health_it/sig_health/3
36. Sammangi, H., Jagatha, A., & Liu, J. (2025b). Harnessing generative AI and large language models for revolutionizing cybersecurity in the Internet of Things: Ethical and privacy implications. Engineering: Open Access, 3(6), 1–12.
37. Sammangi, H., Jagatha, A., & Liu, J. (2025c). Integrating blockchain technology into telemedicine: A framework for enhancing data privacy and security. Engineering: Open Access, 3(6), 1–7.
38. Schemmer, M., Hemmer, P., Kühl, N., Benz, C., & Satzger, G. (2022). Should I follow AI-based advice? Measuring appropriate reliance in human-AI decision-making. arXiv preprint arXiv:2204.06916.
39. Sharma, G., Singh, J., Sammangi, H., Sharma, M., Pandey, R., Srivastava, S., Agarwal, G., & Singh, I. (2025a). A comprehensive assessment of developing a forecasting model for kidney stone formation using deep learning approaches. In H. Sharma, A. Chakravorty, S. Hussain, & R. Kumari (Eds.), Artificial Intelligence: Theory and Applications (Vol. 5588, pp. 121–132). Springer Nature Singapore. https://doi.org/10.1007/978-981-96-1918-4_9
40. Shin, D. (2021). The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI. International Journal of Human-Computer Studies, 146, 102551.
41. Vasconcelos, H., Jörke, M., Grunde-McLaughlin, M., Gerstenberg, T., Bernstein, M. S., & Krishna, R. (2023). Explanations can reduce overreliance on AI systems during decision-making. Proceedings of the ACM on Human-Computer Interaction, 7(CSCW1), 1–38.
42. Wachter, S., Mittelstadt, B., & Russell, C. (2018). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. Harvard Journal of Law & Technology, 31(2), 841–887.
43. Yin, M., Wortman Vaughan, J., & Wallach, H. (2019). Understanding the effect of accuracy on trust in machine

learning models. Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, 1–12.

44. Zhang, Y., Liao, Q. V., & Bellamy, R. K. E. (2020). Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 295–305.