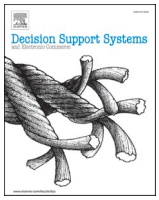
Decision Support Systems 209 (2026) 114721

Contents lists available at ScienceDirect

Decision Support Systems

journal homepage: www.elsevier.com/locate/dss



Unveiling the hidden layer: Ambivalent cognitive effects of XAI in augmented decision-making under uncertainty

Tizian Matschak [a], Ilja Nastjuk [b], Simon Trang [c], Jose Benitez [d,*]

[a] Department of Information Systems, University of Goettingen, Goettingen, Germany
[b] Department of Information Systems, International College of Management, Sydney, Australia
[c] Department of Information Systems, University of Paderborn, Paderborn, Germany
[d] Department of Information Systems and Business Analytics, Ambassador Crawford College of Business and Entrepreneurship, Kent State University, Kent, USA

ARTICLE INFO

Keywords:
Explainable AI
Human-AI decision-making performance
Competence-based trust
Responsibility attribution
Automation bias

ABSTRACT

This study investigates whether explainable AI (XAI), although intended to improve calibrated reliance in augmented decision-making (ADM), may also increase negative consultations. In a between-subjects online experiment ($n = 200$), participants completed a price classification task supported by AI, while we manipulated the presence of local feature-based explanations. Drawing on dual-process theory as an interpretive lens, we conceptualize explainability as a cognitively ambivalent intervention that reshapes reliance behavior. Our results show that XAI-induced transparency exerts competing effects: while it directly reduces negative consultations, it simultaneously increases competence-based trust and shifts responsibility attribution toward the AI. These mechanisms foster overreliance and significantly increase negative consultations, resulting in a competitive mediation pattern. Notably, comparable effects do not emerge for positive consultations, suggesting that transparency more readily amplifies reliance than it promotes corrective belief updating. By empirically demonstrating this paradox, our study develops IS and Decision Sciences research on automation bias in transparent ADM settings and identifies key psychological mechanisms through which explainability influences reliance. These findings inform the design of XAI systems that prioritize calibrated reliance rather than merely increasing perceived transparency.

1. Introduction

Organizations increasingly deploy AI-enabled decision support systems (DSS) in augmented decision-making (ADM) settings [1], where human decision-makers must evaluate and act upon AI recommendations. Because such recommendations are inherently uncertain, decision performance depends on how decision-makers integrate them into their judgments. This is commonly conceptualized as calibrated reliance behavior [2–4], that is, the ability to follow correct AI recommendations (i.e., positive consultations), while avoiding following incorrect recommendations (i.e., negative consultations).

Explainable AI (XAI) has emerged as a key design response to this calibration challenge. From a stimulus-organism-response (S-O-R, [5,6]) perspective, explanations can be understood as a design stimulus that shapes users' cognitive evaluations, which in turn influence how they act upon AI recommendations. In this view, explanations are intended to support decision-makers in evaluating AI recommendations by improving their understanding of the underlying reasoning, thereby enabling more informed reliance decisions [7]. Importantly, XAI does not directly determine human-AI decision performance but influences it indirectly through its effects on cognition and subsequent reliance behavior.

To better understand when XAI supports effective decision-making, existing IS and Decision Sciences research has started to examine this interplay, primarily focusing on situations in which XAI fails to produce these intended cognitive effects. In this perspective, XAI ineffectiveness has, e.g., been attributed to situations in which decision-makers lack the motivation or ability to engage with explanations (e.g., insufficient motivation [8] or AI literacy [9,10]), explanations may be poorly designed (e.g., overly complex or misleading [11]), or a combination of both (e.g., cognitive overload [12]). While this perspective provides important insights into when explanations fail to support decision-

* Corresponding author.
E-mail addresses: tizian.matschak@uni-goettingen.de (T. Matschak), inastjuk@icms.edu.au (I. Nastjuk), simon.trang@uni-paderborn.de (S. Trang), jbenite1@kent.edu (J. Benitez).

https://doi.org/10.1016/j.dss.2026.114721
Received 24 October 2025; Received in revised form 29 June 2026; Accepted 29 June 2026
Available online 30 June 2026

making, it implicitly assumes that explainability operates through its intended cognitive pathway.

However, consistent with S-O-R research [6], a given stimulus can elicit multiple cognitive responses, including unintended effects that extend beyond its intended function [13]. Such effects can emerge even when explanations are processed as intended and may also shape how users rely on AI recommendations. This perspective becomes particularly relevant in light of recent empirical findings reporting performance degradation effects of XAI [9,12,14]. Such effects are difficult to reconcile with research that focuses on absent or insufficient cognitive processing.

Building on this prior work, our research aims to examine how explainability may trigger unintended cognitive effects and how these effects shape calibrated reliance behavior in ADM. We conceptualize ADM performance through positive and negative consultations, which capture whether AI interaction improves or deteriorates decisions and thus serve as direct antecedents of overall decision performance. Drawing on dual-process theory [15], we argue that explainability in ADM is cognitively ambivalent. While explanations can enhance transparency and support critical evaluation, they may also increase competence-based trust in the system and shift perceived responsibility toward it. Because AI recommendations are inherently fallible, these competing mechanisms help explain how explainability can ultimately degrade decision performance.

To test our research model, we examined XAI-advised price estimation in a between-subjects online experiment with 200 participants. The results show that XAI-induced transparency exerts competing effects: while it is directly associated with fewer adoptions of false AI recommendations (i.e., negative consultations), it simultaneously increases competence-based trust and shifts responsibility attribution toward the AI. This, in turn, is significantly associated with higher negative consultations. Notably, comparable effects do not emerge for situations in which the AI recommendation is correct (i.e., positive consultations).

With its focus on ADM under high uncertainty, our study differs from the substantial body of IS research on AI aversion [16–18] which implicitly assumes that increased reliance on AI improves human-AI performance and thus focuses on positive consultations (i.e., when an AI recommendation may correct an erroneous human decision). Instead, we contribute to the emerging body of literature that examines decision contexts with fallible AI and emphasizes calibrated reliance [2,19]. In line with recent discussions on novelty in IS research [20], our study identifies a previously underexplored mediation mechanism through which explainability induces unintended cognitive effects that reshape reliance behavior and can lead to automation bias. In doing so, our study advances emerging IS and Decision Sciences research on calibrated reliance in ADM under high uncertainty in three ways. First, we conceptualize explainability as a cognitively ambivalent intervention that gives rise to both intended and unintended cognitive effects. Second, we demonstrate how transparency reshapes reliance behavior through changes in competence-based trust and responsibility attribution, highlighting important cognitions underlying reliance calibration. Third, we advance DSS design theory by challenging the assumption that transparency inherently improves performance and deriving corresponding design safeguards.

2. Background

2.1. Augmented decision-making under high uncertainty

In the context of human-AI decision-making, a common distinction differentiates between ADM and human-in-the-loop (HITL) approaches. In traditional HITL systems, humans supervise a high-performing AI system rather than construct decisions independently. Their role is to monitor system performance, retain override authority, and serve as an accountability mechanism by detecting errors, performance degradation, or malfunctions across decision instances [22].

By contrast, ADM emphasizes decision integration rather than supervision. Here, AI systems provide probabilistic and fallible recommendations that are not merely inspected but incorporated into users' reasoning processes as one cue among others [23,24]. Rather than functioning as a gatekeeper, the human decision-maker combines AI recommendations with contextual knowledge, expertise, and other information sources to construct a final judgment [25].

The role of the human decision-maker is particularly important in this context, as AI recommendations are embedded in decision environments characterized by uncertainty. This uncertainty does not only imply probabilistic and fallible AI, but rather reflects the difficulty of assessing the correctness of a specific recommendation at the moment of decision-making. Even highly capable models may produce outputs that are context-dependent or difficult to verify without additional information. As a result, decision-makers cannot rely on AI recommendations unconditionally but must evaluate their validity on a case-by-case basis. This differentiates ADM from research streams such as algorithm aversion [16–18], which conceptualize reliance on AI as inherently beneficial whenever algorithms significantly outperform human decision-makers in accuracy. In ADM, by contrast, both following and rejecting AI recommendations can be either beneficial or detrimental, depending on the specific instance.

Recent advances in AI further accelerate the relevance of ADM. The emergence of agentic AI systems extends ADM beyond static recommendation provision toward more proactive forms of decision support, thereby shaping not only the content but also the structure of decision-making processes [26]. While this development does not fundamentally alter the core logic of ADM, it expands the scope and complexity of decisions that can be supported by AI.

2.2. Calibrated reliance in augmented decision-making

The probabilistic and fallible nature of ADM implies that human-AI decision performance depends not solely on algorithm appreciation (or reduced algorithm aversion) [16–18], but on calibrated reliance on those recommendations [2–4]. Calibrated reliance is the context-sensitive and proportionate weighting of AI-generated recommendations relative to human expertise, such that reliance corresponds to the system's actual competence and the specific decision environment [2–4,19]. AI may affect calibrated reliance through two types of consultation outcomes (see Table 1).

Positive consultations of an AI occur when decision-makers correct an initially erroneous judgment (i.e., it departs from the ground truth) by following an accurate AI recommendation. In contrast, negative consultations of an AI arise when decision-makers abandon a correct initial judgment (i.e., it equals the ground truth) in favor of an erroneous AI recommendation. All remaining cases represent no consultation, where AI recommendations do not alter the final decision. Importantly, consultations constitute the behavioral building blocks of decision performance in ADM settings [2,23,27]. Each consultation reflects a discrete adjustment relative to the ground truth that either improves or deteriorates the decision outcome. Aggregated across decision instances, these consultation behaviors are thus a direct behavioral determinant of overall decision performance.

Table 1
Calibrated reliance: decomposition of negative and positive AI consultations.

Ground Truth	Initial Human Judgment	AI Recommendation	Final Human Decision	Reliance Behavior
1	1	0	0	Negative Consultations
0	0	1	1	
1	0	1	1	Positive Consultations
0	1	0	0	
Others				No Consultations

Calibrated reliance requires decision-makers to critically assess whether an AI's recommendation is warranted in the specific case and how strongly it should influence the final judgment [8]. However, this evaluative process is cognitively demanding and requires information on the inner working of the AI. In practice, decision-makers may not always engage in such careful scrutiny or have the necessary information available. Instead, reliance may be guided by intuitive impressions and heuristic cues when determining the extent to which to rely on AI recommendations [24,28].

This tendency is reinforced by the increasingly agentic nature of contemporary AI systems. Following Maruping and Baird's [26] notion of agentic IS artifacts, AI systems are increasingly perceived as entities capable of assuming decision-making responsibilities under uncertainty rather than as passive tools. This perceptual shift fundamentally changes how users evaluate and interact with AI systems. Instead of calibrating reliance on a case-by-case basis, users increasingly make delegation decisions at the process level, determining whether and to what extent an agentic AI system should act on their behalf.

2.3. The role of XAI in augmented decision-making

Transparency mechanisms, particularly XAI, are proposed as design features to support the human decision-maker's evaluative process and facilitate calibrated reliance [21,29–31]. XAI has emerged to render complex AI models (e.g., deep neural networks) that sacrifice transparency and interpretability (black-box models) for higher predictive performance [29] more understandable to human users. Therefore, XAI typically generates interpretable approximations of black-box predictions to help users understand why a specific output was produced [21], when the AI recommendation is erroneous, and whether the recommendation should be translated into action [7]. Accordingly, explanations are expected to enhance users' understanding of AI recommendations in ADM and promote analytical scrutiny to enable superior joint human-AI performance compared to either human or AI acting alone [30]. However, because such explanations approximate the behavior of complex models rather than fully reveal their internal logic, they may be incomplete, imprecise, or even misleading.

In practice, XAI applies a range of explanatory techniques that make individual predictions interpretable to users [21,31]. Among these techniques, local post-hoc explanations have received particular attention in research [32,33]. These explanations aim to approximate the relationship between input features and model outputs for a specific prediction [34,35]. We refer specifically to local post-hoc explanations when discussing explainability throughout the remainder of the paper. While XAI is intended to enhance deliberate evaluation by increasing algorithmic transparency, its effects may extend beyond this intended function. Given that reliance is also shaped by heuristic cues and perceptions of AI as an agent [26], explanations may simultaneously influence multiple cognitive processes, with potentially unintended consequences for reliance behavior.

3. Research model

This research sets out to examine how explainability induces unintended cognitive effects, and how these effects shape reliance behavior in ways that may lead to human-AI performance degradation. To do so, we draw from S-O-R theory [5,6] as the lens to structure our research model. In this view, explanations of an AI output represent the stimulus that shapes users' internal cognitive evaluations, which in turn drive their behavioral responses to AI recommendations. Specifically, we conceptualize cognitive evaluations as the organism component and reliance behavior as the response. To theorize the organism component, we draw on dual-process theory [15] as an interpretive framework. An overview of the research model is depicted in Fig. 1.

3.1. Dual-process theory as interpretive framework

Consciously or unconsciously, decision-makers may accept AI recommendations that would not be optimal under full rationality. Therefore, research refers to this phenomenon as automation bias [36]. The theoretical foundation of automation bias is rooted in the dual-process theory of System 1 and System 2 [15]. This theory proposes two fundamentally different cognitive systems involved in human information processing and decision-making. System 1 is described as fast, automatic, effortless, associative, and difficult to control or modify. While it is supposed to improve efficiency in complex decision contexts, reliance on heuristics may also give rise to systematic cognitive biases and distortions [37]. In contrast, the cognitive processes of System 2 are characterized as slower, serial, effortful, and deliberately controlled.

The dual-process framework has been widely adopted across disciplines and has proven informative in prior research on automation bias [38]. It is particularly well suited as an interpretive framework for our research question because it explains why identical cognitive tasks may produce divergent outcomes within the same usage context. Specifically, variation in the activation and interplay of System 1 and System 2 provides a useful lens for understanding whether individuals critically evaluate or uncritically rely on automated recommendations. Building on this perspective, we argue that the presentation of XAI information may unintentionally increase the salience of heuristic cues rather than consistently activating reflective System 2 processing.

Prior research [e.g., 39,40] highlighted that the decision context influences the relevance of stimuli that may activate either mode. In ADM, users often face repeated similar decisions. Each decision involves considerable uncertainty yet carries only limited individual risk. These characteristics shape how users process AI recommendations. Because the correctness of a recommendation is often difficult to verify and immediate performance feedback is limited, users may not be able or willing to evaluate each recommendation in detail. At the same time, the repetitive nature of ADM tasks creates incentives for efficiency-oriented processing, as users seek to conserve cognitive resources across many similar decisions [36,41,42]. These contextual conditions make

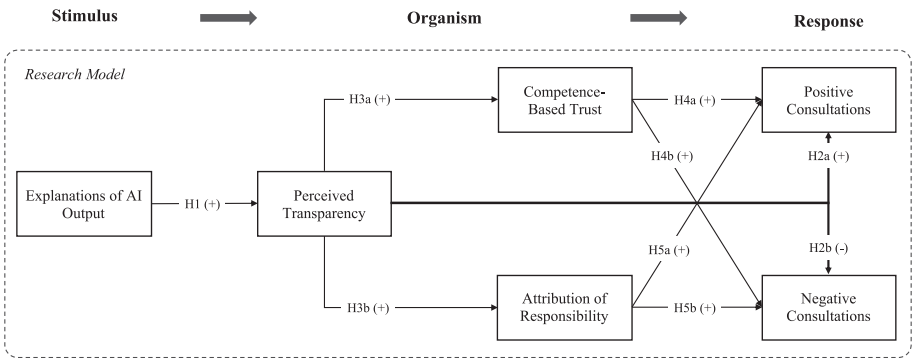


Fig. 1. Research model.

cognitive effects particularly relevant that reduce case-specific scrutiny and increase reliance on the AI system. Building on perceived transparency, we therefore consider competence-based trust and attribution of responsibility to be particularly relevant cognitive effects that may adversely shape reliance behavior (see Table 2 for construct definitions).

While repetition shapes users' beliefs about a technology, prior research shows that repeated interaction with technology fosters trust [46]. Trustworthiness, in turn, is a well-established antecedent of overreliance [3] and is conceptually closely linked to automation bias. Moreover, compared to HITL contexts (which are typically characterized by high-risk expert decision-making), responsibility in more automated settings is less clearly delineated between users and the AI system. During interaction, perceived responsibility often shifts away from the user toward the system [47]. When individual decisions have limited personal consequences, users have weaker incentives to invest significant cognitive effort to identify or process relevant information [41]. This becomes particularly relevant in light of growing research conceptualizing AI as possessing a dual identity, functioning both as a tool and as an autonomous actor [48].

Importantly, these constructs are related but not interchangeable. A user may perceive an AI system as transparent without judging it to be highly competent. Conversely, a user may trust an AI's competence without understanding how it arrived at its recommendation. Likewise, users may attribute responsibility to an AI system even when they neither fully understand its reasoning nor consider it fully competent.

3.2. Hypotheses development

In AI-enabled decision support contexts, XAI provides explainability to individual predictions with the aim of enhancing decision transparency [34,35]. However, the degree of transparency perceived by the user may vary. Our research focuses on the perceived transparency of the AI system as the central explanatory mechanism through which XAI exerts its influence on our dependent variables. Similarly, Shin [49] argues that transparency from a user perspective is not merely a technical property of the system, but is strongly shaped by how understandable and interpretable the system appears to its users. This decision aligns with contemporary studies on XAI, such as Hamm et al. [50], which identify perceived transparency as a significant antecedent of cognitive variables such as trust. In addition, Sullivan and Weger [51] find that the presence of explanatory features significantly increases perceived transparency and ease of understanding in AI-enabled decision-making contexts. Thus, XAI-based explanations function as a cognitive bridge between algorithmic processes and users' subjective perceptions. By emphasizing perceived transparency as the mediating construct, we capture users' subjective interpretations rather than merely the presence of explanatory information, thereby accounting for the previously reported variability of explanation effectiveness [14]. Building on this, we argue that integrating XAI components into AI-enabled DSS will increase users' perceived transparency of the system's recommendations:

Table 2
Definition of key constructs.

Key Construct	Definition	Ref.
Perceived Transparency	Perceived transparency refers to the extent to which a user perceives the basis of an AI recommendation as clear and intelligible. It captures a subjective judgment of process clarity regarding how the recommendation was generated. While an AI model's output may be fully explainable from a technical standpoint, transparency is a subjective perception that can vary significantly across users of the same AI system. Accordingly, perceived transparency is distinct from judgments of the AI's competence, the correctness of its recommendation, and responsibility for the decision outcome.	[43]
Competence-based Trust	Competence-based trust refers to the extent to which a user believes that the AI has the ability to perform a focal task reliably and effectively. It captures an evaluative belief about the system's task-specific capability. Accordingly, competence-based trust is distinct from judgments about how transparent the recommendation process is and from judgments about who bears responsibility for the outcome.	[44]
Attribution of Responsibility	Attribution of responsibility refers to the extent to which a user assigns causal responsibility for a decision outcome to the AI system rather than to oneself or other actors involved in the decision process. It captures perceived outcome ownership within human-AI collaboration. Accordingly, attribution of responsibility is distinct from judgments of process clarity and from beliefs about the AI's capability.	[41,45]

Hypothesis 1. *(H1): Explanations of AI output increase users' perceived transparency of the AI system.*

A growing stream of literature positions transparency as a critical enabler for enhancing human judgment in ADM [29,30]. Qualitative and experimental evidence further indicates that explanations prompt users to slow down and reassess their judgments, particularly in contexts characterized by uncertainty or cognitive conflict [52]. From a dual-process perspective, explanations can, under some conditions, interrupt default cue-based responding and encourage more reflective consideration, especially when users encounter signals that their initial judgment may be incomplete, uncertain, or potentially incorrect. Prior research suggests two mechanisms through which explanations can trigger more reflective processing.

On the one hand, exposing elements of the AI's reasoning introduces diagnostic information that increases cognitive friction. The provision of additional information requires users to allocate cognitive resources to interpretation, comparison, and integration, thereby slowing decision-making and fostering more reflective evaluation rather than immediate acceptance or rejection of the recommendation [53].

On the other hand, XAI explanations often surface discrepancies between users' intuitive judgments and the AI's reasoning, thereby creating cognitive conflict. When users observe that the AI relies on cues they had overlooked, weighted differently, or not considered at all, this mismatch may act as a conflict signal that interrupts heuristic processing. Dual-process research suggests that such conflict detection is a key trigger for System 2 engagement, prompting users to re-examine their initial beliefs and reassess the validity of both their own reasoning and the AI's recommendation [54,55]. Against this background, we posit that decision transparency in AI systems enhances users' ability to make accurate judgments regarding when to follow algorithmic recommendations and when to override them [30,56,57]. As a result, we expect that users who initially make an incorrect decision will be more likely to recognize their error when the AI provides a correct recommendation and, consequently, follow the AI's recommendation to revise their initial judgment (positive consultations; see Table 1). At the same time, the decision transparency provided by AI allows users to recognize and avoid situations where the AI would make an incorrect recommendation that could lead them away from an initially correct decision (negative consultations). Accordingly, we hypothesize:

Hypothesis 2a. *(H2a): Perceived transparency of the AI system is positively related to positive consultations.*

Hypothesis 2b. *(H2b): Perceived transparency of the AI system is negatively related to negative consultations.*

While these direct effects are theoretically plausible and represent the preferred outcomes of XAI from a practitioner's perspective, prior research [e.g., 9] provides mixed evidence regarding the consequences of XAI for decision performance in ADM, suggesting the existence of a

potential 'hidden layer' in this causal relationship. This hidden layer refers to the subsequent cognitive effects triggered by increased transparency of AI decisions, which may interfere with the intended positive impact of XAI on decision-making. We elaborate on this in the following hypotheses.

Among multidimensional conceptualizations of trust in technology, competence-based trust, defined as the belief that a system possesses the ability and reliability required to perform effectively within a given domain [44], is particularly critical in AI-enabled decision-making contexts, where users must often rely on system outputs under conditions of limited observability. Because users typically lack direct access to the algorithmic processes underlying AI recommendations, assessing system competence becomes inherently uncertain.

Perceived transparency helps mitigate this uncertainty by providing users with diagnostic insight into the system's reasoning processes [43]. Explanations can be understood as a form of knowledge transfer from the system to the user, enabling individuals to infer how and why specific decisions are generated [58]. When users are able to trace the logic underlying AI recommendations, they are better positioned to evaluate whether the system's performance reflects system expertise rather than randomness. Such evaluability is foundational to competence-based trust, as perceptions of ability emerge when outcomes can be attributed to skillful operation.

Moreover, transparency facilitates the development of more accurate mental models of the AI system. These mental representations allow users to approximate the boundaries of the system's capabilities and anticipate potential failure modes, thereby enhancing perceived predictability and functional reliability [9,33]. By supporting users' ability to mentally simulate the system's decision logic, transparency reduces cognitive uncertainty and strengthens beliefs that the AI is operating correctly [58,59]. This mechanism is particularly relevant in ambiguous situations in which objective performance verification is difficult or costly, prompting users to rely on process-based cues when forming competence judgments [60–62].

In sum, this concept is supported by multiple empirical studies in XAI that have identified perceived transparency as a key antecedent of trust [32,63]. Accordingly, we hypothesize:

Hypothesis 3a. *(H3a): Perceived transparency is positively related to competence-based trust in the AI system.*

As individuals increasingly collaborate with intelligent technologies, they tend to treat computer-based systems as quasi-team members, exhibiting social responses toward them that resemble those directed at human collaborators [64,65]. Within such collaborative arrangements, responsibility is often distributed across actors, giving rise to phenomena such as responsibility diffusion and social loafing [36]. Rather than being perceived solely as technical tools, AI systems are progressively construed as task-performing actors. This shift is driven by individuals' propensity to attribute human-like qualities to intelligent technologies (homunculus fallacy) [58] and the fact that AI artifacts are no longer passive tools and can assume responsibility for tasks [26]. Reflecting this evolution, scholars increasingly conceptualize AI as possessing a dual identity: simultaneously a computational artifact and an autonomous agent capable of shaping decision processes [48]. Consequently, users engaged in human-AI collaboration are predisposed to assign at least partial responsibility for decision outcomes to the system.

However, responsibility attribution requires more than mere participation in a task. It depends on whether the AI is perceived as a causally efficacious decision-maker rather than an opaque instrument. Here, perceived transparency plays a pivotal role as it provides insight into the system's reasoning and thereby clarifies how specific outputs are generated. As Josephson and Josephson [66] note, "[…] explanation is an assignment of causal responsibility" because it identifies the factors that produced a given outcome. By rendering the decision process intelligible, transparency reduces epistemic opacity (the condition in which users cannot access the system's internal reasoning or epistemic states) [67] and instead enables informed causal inferences about the origin of a recommendation.

Making the AI's epistemic state visible fulfills what philosophical accounts describe as the epistemic criterion for responsibility attribution: actors are more likely to be held responsible when their beliefs, reasoning structures, or decision rules can be inferred [68,69]. Baird and Maruping [26] emphasize that human motivation to delegate responsibility to an agentic IS artifact is contingent upon the transparency of the decision model, which encapsulates how options are ranked and choices are made. Understanding how an AI evaluates information and arrives at judgments strengthens perceptions that the AI actively shapes the decision rather than merely executing predefined commands. This shift promotes perceptions of functional agency, which is foundational for attributing responsibility. Therefore, we hypothesize the following:

Hypothesis 3b. *(H3b): Perceived transparency is positively related to users' attribution of responsibility to the AI system.*

While H3a and H3b establish competence-based trust and responsibility attribution as cognitive consequences of perceived transparency, the next step is to understand how these perceptions shape users' subsequent decision behavior. Prior research based on dual-process theory suggests that individuals allocate cognitive effort selectively, investing greater analytic processing when information is perceived as credible and diagnostically valuable [70]. Consequently, when an algorithm is trusted as competent (H3a), its recommendations are more likely to receive cognitive consideration rather than heuristic dismissal or disregard [71]. This increases the likelihood that users treat AI recommendations as meaningful input to their decision-making.

In line with this reasoning, trust is widely recognized as a critical determinant of whether users accept, rely upon, or disregard AI-generated recommendations [8,72]. Furthermore, when trust evolves into blind trust, users may become less inclined to rigorously evaluate potential alternatives and instead defer uncritically to the system's judgment [30,73]. Building on this reasoning, we propose that competence-based trust shapes whether users treat AI recommendations as valuable, thereby increasing the likelihood that they act on the recommendations, either by accepting correct recommendations or following incorrect ones:

Hypothesis 4a. *(H4a): Competence-based trust is positively related to positive consultations.*

Hypothesis 4b. *(H4b): Competence-based trust is positively related to negative consultations.*

In their review, Lerner and Tetlock [41] argue that responsibility is a concept of considerable logical complexity, shaped by the interaction between characteristics of the decision-maker and the surrounding task environment, giving rise to diverse behavioral consequences. Prior research consistently shows that individuals who perceive themselves as accountable engage in more vigilant information processing [74], produce higher-quality judgments [75], and are less likely to exhibit cognitive biases [41]. These effects are commonly attributed to the motivational demands of accountability: perceiving oneself as responsible for outcomes increases the perceived need for justification, thereby encouraging more deliberate cognitive processing [76].

Attributing responsibility to an AI system reshapes how users cognitively approach a decision by shifting perceived ownership of the outcome. *Accountability theory* [41] suggests that individuals process information more carefully when they expect to be held responsible for their judgments, largely because responsibility creates a perceived need for justification. Anticipating that one may have to defend a decision motivates more deliberate and systematic evaluation of available information. This aligns with the dual-process perspective, according to which perceived accountability increases intrinsic motivation to process information more analytically [70,77].

When responsibility is attributed to the AI rather than the user, this

justificatory pressure is reduced. Consequently, the shift can induce a social-loafing-like reduction in effort, because the user perceives the burden of careful evaluation as shared with the AI [78]. At the same time, it creates a moral-hazard-like incentive structure in which the perceived personal cost of insufficient scrutiny decreases. As a result, responsibility attribution does not merely alter role perception; it weakens the user's motivation to engage in effortful verification and increases reliance on the AI's recommendation as a peripheral cue. Consequently, we expect that this enhances users' acceptance of AI recommendations, leading to both increased negative and positive consultations. Accordingly, we posit:

Hypothesis 5a. *(H5a): Attribution of responsibility is positively related to positive consultations.*

Hypothesis 5b. *(H5b): Attribution of responsibility is positively related to negative consultations.*

4. Research design

4.1. Experimental setup

To test the proposed research model, we conducted an online experiment with a between-subjects design, comparing two scenarios: a group receiving local feature-based XAI explanations and a control group without explanations. A between-subjects design was chosen to avoid carryover, learning, and contrast effects that would have occurred if participants had been exposed to both explained and non-explained AI recommendations. The experimental task required participants to classify car prices, deciding whether each car was above or below a market value threshold ($15,000). The scenario was selected because it requires no specific domain knowledge, represents an uncertain real-world decision problem where AI support adds value, and can be repeated across multiple iterations.

We trained a machine learning (a subset of AI) algorithm to perform the given task. This comes with the advantage of using the original AI output in our experiment. For this purpose, we used data from a leading US online car sale website. Based on that, we implemented a data pre-processing pipeline (e.g., categorization of car brands to reduce complexity) and trained a Random Forest Classifier by using 5-fold cross-validation. By doing so, we obtained an accuracy of 98% on our test sample of 1000 data points. With an Area Under the Curve (AUC) of 0.98, our model represents a strong classifier. Given that our algorithm generates very few misclassifications, which are difficult to detect, we leveraged findings from adversarial attack research and deliberately manipulated a portion of the input data to provoke sufficiently clear and identifiable classification errors. These induced errors served as the foundation for our experimental setup and allowed for a controlled investigation of erroneous AI recommendations. Since the Random Forest Classifier belongs to the category of black-box AI, we used SHapley Additive exPlanations (SHAP) [79]. By adjusting the feature values of each data sample and observing the resulting changes in output, SHAP provides local interpretability by explaining predictions with respect to each feature. Local explanations are particularly valuable for end users who must interpret and act on individual predictions [50]. The generated classification and XAI results were then used as inputs for the experimental tasks.

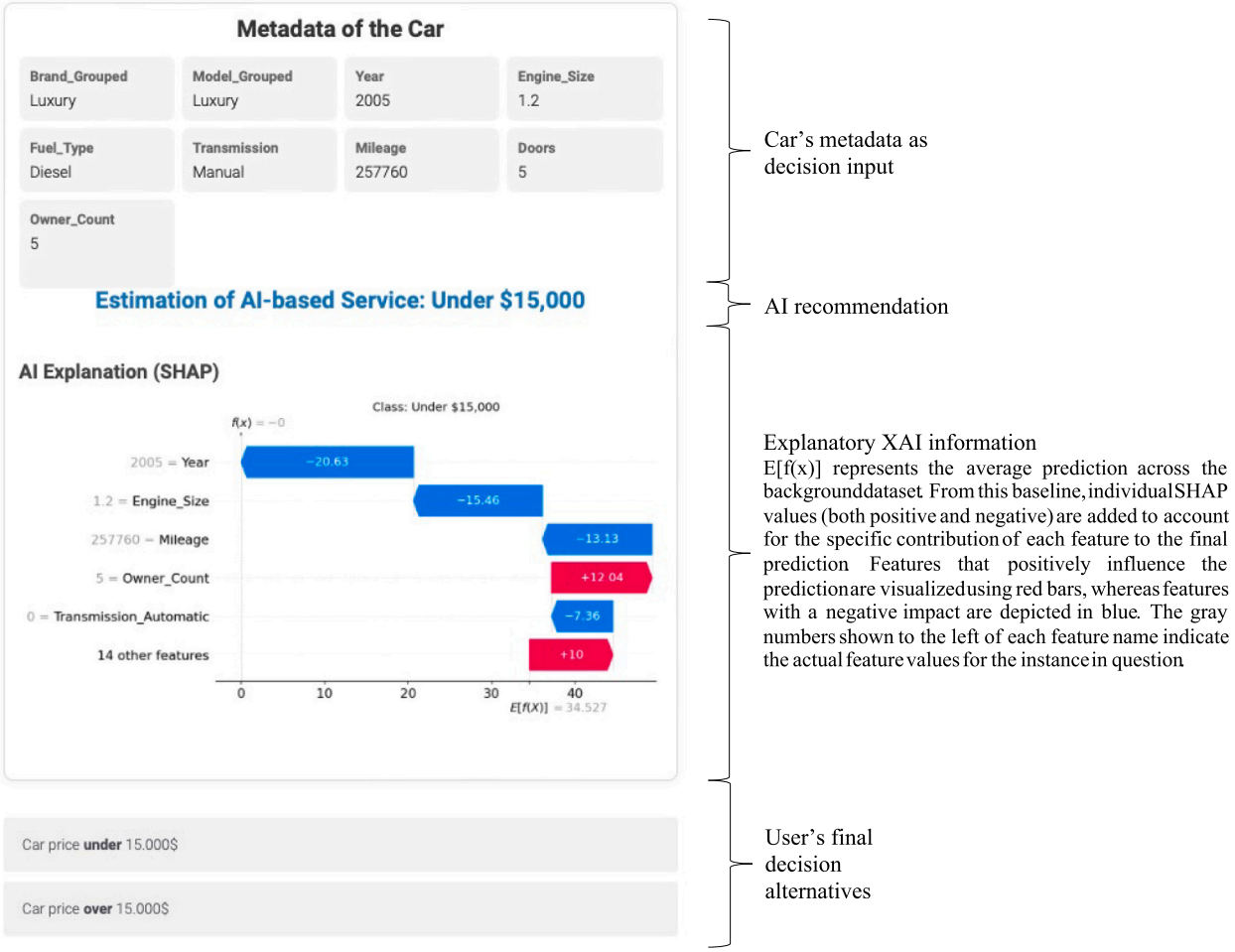


Fig. 2. Exemplary user interface with XAI.

The core experiment for each participant consisted of a brief introduction, the completion of twelve tasks, and a subsequent survey. Among other considerations, we deliberately limited the number of features presented in the explanations (see Fig. 2). Constraining the volume of explanatory information serves to mitigate cognitive overload and has been shown to enhance decision-making performance [35]. At no time were participants given information about their performance in order to limit additional methodological complexity (e.g., anchoring bias) that complicates the measurement and interpretation of explanation-specific effects on users' cognition [9].

To assess the effectiveness of the experimental manipulation of explanation provision (XAI vs. no XAI), we conducted a one-way ANOVA on a four-item scale measuring perceived explanatory information (e.g., "I had plenty of explanatory information regarding the AI-enabled service's estimation"; 1 = strongly disagree, 7 = strongly agree; $\alpha = 0.73$). Participants in the XAI condition reported significantly higher scores (M = 4.24, SD = 1.17) compared to participants in the no XAI condition (M = 2.86, SD = 1.31). The difference was statistically significant, $F(1, 198) = 61.78$, $p < .001$, indicating that our manipulations worked as intended.

4.2. Measures

All latent variables in our research model were measured reflectively using seven-point Likert scales ranging from 1 ("strongly disagree") to 7 ("strongly agree"). *Perceived transparency* was measured using four adapted items based on the scale developed by Wang and Benbasat [43] and later refined by Hamm et al. [50]. *Competence-based trust* was assessed using four contextualized items adapted from McKnight et al. [44]. *Attribution of responsibility* was captured using three items adapted from Yue and Li [80] and complemented by one additional item developed for the experimental context. Because participants received no outcome feedback during the experiment, we did not distinguish between pre- and post-decisional responsibility.

Our dependent variables, *positive consultations* and *negative consultations*, were operationalized as behavioral measures based on participants' decision changes across tasks, following established approaches in automation bias research [27]. As control variables, we included AI knowledge, AI usage experience, gender, and age to account for potential individual differences in familiarity with AI systems and decision-making behavior. All measurement items are reported in Appendix A.

4.3. Data collection

Data collection was carried out in March 2025 via the online crowdsourcing platform Prolific [81]. Within the survey, manipulation and attention checks were implemented to confirm participants' ability to engage with the presented scenario.

We estimated the minimum sample size using a power analysis for structural equation models [82]. We anticipate a medium effect size (0.3) to be appropriate for our analysis. Assuming a statistical power of 0.9 and a significance level of 0.05, the analysis indicated that a minimum sample size of 153 observations is required to detect effects in the path model estimation. After data cleaning (i.e., attention checks and non-zero variance across survey items), we received a final sample of 200 responses. Accordingly, we consider our statistical analysis to be sufficiently powered.

Our sample is composed of 102 responses from the scenario with XAI and 98 responses from the control group. On average, the participants needed 13.65 min (SD = 7.06) to complete the twelve price classification tasks. The average age of the participants is 37 years (SD = 14.75). 100 (50.0%) respondents identified themselves as female and 100 (50.0%) respondents identified themselves as male. In addition, we asked about the level of knowledge of AI technology (M = 3.65; SD = 1.37) and AI use experience (M = 4.12; SD = 1.52).

To ensure that the experimental task is sufficiently challenging to justify the need for ADM support, we asked participants to rate task difficulty on a 7-point Likert scale: "I found this task difficult."With an average of 5.78, the task difficulty is in the upper middle range of the 7-point Likert scale.

5. Results

5.1. Measurement validation

A confirmatory factor analysis of the measurement model suggested an acceptable fit ($\chi^2 = 164.73$, df = 78, $\chi^2/df = 2.112$, RMSEA = 0.075, CFI = 0.972, TLI = 0.962, SRMR = 0.037 [83]). For the multi-item constructs, internal consistency (CR) and convergent validity (AVE) were evaluated, and standardized factor loadings were inspected as an indicator reliability check. Discriminant validity was confirmed using HTMT (max HTMT = 0.789, which is below 0.85 [83]) and Fornell-Larcker criterion (i.e., the square root of the average variance extracted exceeds the correlations with any other construct in the model, [84]). As perceived transparency, attribution of responsibility, and competence-based trust are highly correlated, we further assessed their bivariate discriminant validity and compared constrained and unconstrained models [85]. Specifically, the chi-square difference tests revealed significant differences for perceived transparency and attribution of responsibility ($\Delta\chi^2(1) = 503.96$, $p < .001$), perceived transparency and competence-based trust ($\Delta\chi^2(1) = 376.77$, $p < .001$), and attribution of responsibility and competence-based trust ($\Delta\chi^2(1) = 488.56$, $p < .001$). The results provide further evidence that the constructs are empirically distinct. Descriptives, reliabilities, and correlations are reported in Table 3.

5.2. Hypotheses testing

In experimental research designs with latent variables, structural equation modeling (SEM) can be preferable to other methods because it is advantageous for correcting measurement errors [86]. Thus, we use SEM to analyze our data. As negative and positive consultations represent count variables, we utilize maximum likelihood estimation with robust (Huber-White) standard errors (implemented as MLR in lavaan 0.6–21 [87]) to account for violations of non-normality. The model structure displayed an acceptable fit ($\chi^2 = 237.651$, df = 119, $\chi^2/df = 1.997$, RMSEA = 0.071, CFI = 0.962, TLI = 0.948, SRMR = 0.051). The results of the path estimation are shown in Fig. 3.

The findings provide partial support for our theorized double-edged sword of XAI in ADM. Consistent with H1, the presentation of AI explanations significantly increases perceived transparency ($b = 0.192$, $p = .006$). With regard to the direct behavioral consequences of perceived transparency, the results reveal a differentiated pattern of effects. Perceived transparency does not significantly increase positive consultations ($b = 0.029$, $p = .825$), providing no support for H2a. In contrast, perceived transparency exhibits a statistically significant negative relationship with negative consultations ($b = -0.255$, $p = .045$), yielding support for H2b.

Turning to the proposed inner model mechanisms, perceived transparency substantially enhances competence-based trust ($b = 0.796$, $p < .001$), providing support for H3a. Likewise, consistent with H3b, perceived transparency significantly increases attribution of responsibility ($b = 0.664$, $p < .001$). The effects of the mediators on consultation behavior reveal significant support for negative consultations but no support for positive consultations. Neither competence-based trust ($b = 0.078$, $p = .434$) nor attribution of responsibility ($b = 0.125$, $p = .126$) significantly predicts positive consultations, providing no support for H4a and H5a. In contrast, competence-based trust significantly increases negative consultations ($b = 0.265$, $p = .019$), supporting H4b, and likewise attribution of responsibility shows a positive effect on negative consultations ($b = 0.327$, $p < .001$),

Table 3
Descriptives, reliabilities, and correlations.

Construct	Scale	Mean (SD)	FL	CR	AVE	1	2	3	4	5	6
1. Perceived Transparency	1–7	3.504 (1.761)	0.844–0.954	0.948	0.820	0.906					
2. Attribution of Responsibility	1–7	3.336 (1.621)	0.837–0.954	0.941	0.800	0.625	0.894				
3. Competence-based Trust	1–7	3.688 (1.623)	0.932–0.955	0.971	0.894	0.774	0.645	0.945			
4. Positive Consultations	0–12	0.600 (0.977)	NA	NA	NA	0.156	0.188	0.174	NA		
5. Negative Consultations	0–12	0.995 (1.380)	NA	NA	NA	0.139	0.316	0.263	0.360	NA	
6. Explanations of AI Output	Binary	0.510 (0.501)	NA	NA	NA	0.230	0.100	0.097	0.039	0.105	NA

Notes. SD, standard deviation; FL, factor loading (range); CR, composite reliability; AVE, average variance extracted; bolded numbers indicate the square root of the AVE.

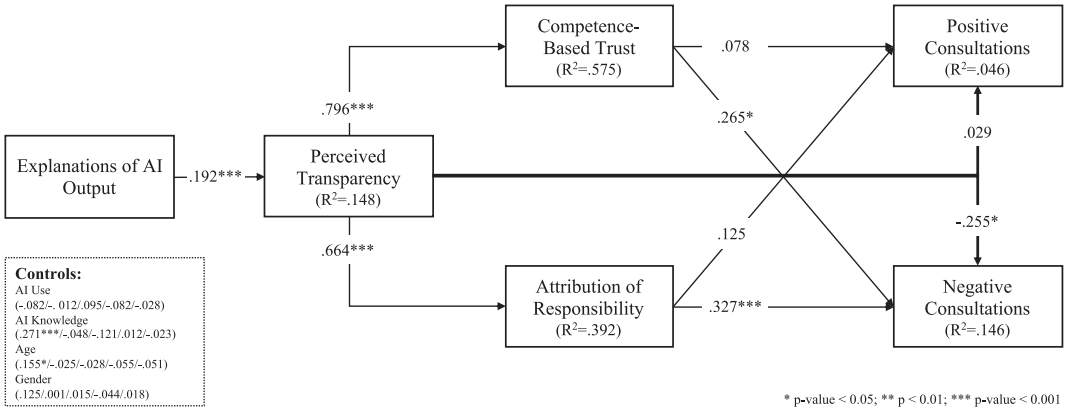


Fig. 3. Results.

supporting H5b.

To better understand the proposed competitive mediation between perceived transparency and consultation behavior, we calculated total effects as well as indirect effects. Considering the total effects first, perceived transparency exhibits a positive overall association with both consultation outcomes. Specifically, perceived transparency shows a significant total effect on positive consultations (b = 0.174, p = .037) as well as on negative consultations (b = 0.173, p = .015). Next, we examined the indirect pathways through the proposed mediators. For positive consultations, neither of the specific indirect effects reaches statistical significance. In contrast, the effects on negative consultations are mediated through both heuristic pathways. Perceived transparency shows a significant positive indirect effect via attribution of responsibility (b = 0.217, p = .001) and a significant positive indirect effect via competence-based trust (b = 0.211, p = .018). Together, these mechanisms yield a substantial and statistically significant total indirect effect on negative consultations (b = 0.428, p < .001). Overall, the results provide support for the proposed competitive mediation on negative consultation behavior where the intended benefits of increased transparency are partially offset by unintended cognitive and behavioral effects.

5.3. Robustness checks

Our results may be sensitive to the distributional specification of consultation behavior and alternative causal model structures. First, the count nature and potential zero inflation of the consultation outcomes warrant additional scrutiny (positive consultations: Skewness = 2.51, Kurtosis = 9.93, 62% zero counts; negative consultations: Skewness = 1.49, Kurtosis = 1.63, 53% zero counts). We therefore re-estimated the outcome models using generalized linear models that explicitly account for overdispersion and zero inflation. Poisson models revealed overdispersion (positive consultations: dispersion = 1.44, z = 1.84, p = .033; negative consultations: dispersion = 1.71, z = 3.13, p = .001), leading us to estimate negative binomial regressions. Results were consistent with the SEM findings: no focal predictors significantly related to positive consultations, whereas perceived transparency negatively predicted negative consultations (b = −0.264, p = .014), and attribution of responsibility (b = 0.371, p < .001) as well as competence-based trust (b = 0.257, p = .023) showed positive associations.

Second, we examined alternative causal models, including interaction and reciprocal structures between competence-based trust and attribution of responsibility. Both models showed inferior fit compared to the proposed research model. Specifically, the interaction model exhibited substantially worse fit (ΔBIC = 923.115), providing strong evidence in favor of the proposed model, while the bidirectional model also showed poorer fit (ΔBIC = 1.298), indicating no empirical support for reciprocal effects between the constructs.

6. Discussion and implications

6.1. Summary of findings

This research aims to investigate how explainability can trigger unintended cognitive effects, and how these effects influence reliance behavior in ways that may ultimately reduce performance. Building on S-O-R [5,6] and on a dual-process perspective [15], we theorized that explanation-induced transparency can exert a double-edged effect (i.e., enhancing or degrading ADM performance) through ambivalent cognitive pathways. Overall, our results support this claim.

Considering the total effects, perceived transparency shows a positive and significant overall association with both positive and negative consultations. In other words, transparency in our experimental setting alters reliance calibration in both directions: some decisions improve through increased reliance on correct AI recommendations, while others deteriorate through increased reliance on incorrect recommendations. At the same time, a closer examination of the mediation pathways reveals a double-edged effect that unfolds for negative consultations. Specifically, transparency influences consultation behavior through two competing cognitive pathways (i.e., competence-based trust and

attribution of responsibility), such that the intended benefits of increased transparency coexist with unintended behavioral consequences. These findings support the core logic of our model and demonstrate that explainability reshapes decision behavior rather than uniformly improving it. This, in turn, can explain the degraded decision performance associated with XAI in ADM.

Notably, the indirect effects via competence-based trust and attribution of responsibility are estimated to have a similar magnitude. While we do not interpret this as evidence that both mechanisms exert identical influence, the comparable effect sizes suggest that epistemic evaluations of AI competence and shifts in perceived responsibility may represent similarly relevant pathways through which transparency affects reliance behavior.

Finally, the absence of significant effects on positive consultations, despite clear effects on negative consultations, points to a central XAI paradox. Explanations appear to increase users' reliance on AI recommendations without correspondingly improving their ability to identify when such reliance is warranted. In this sense, XAI may function less as a corrective aid than as a persuasive cue that lowers the threshold for following AI recommendations. One plausible explanation lies in the asymmetry of decision correction processes. Adopting AI recommendations when holding an initially incorrect judgment (i.e., 'acknowledging the AI is right, while I am wrong') rather than holding an initially correct judgment (i.e., 'correctly rejecting the AI recommendation, because the AI's flaw can be identified') requires effortful belief updating, which has been shown to rely on significant motivational engagement [88] and which transparency alone may not sufficiently trigger. In contrast, transparency-induced trust- and responsibility-related cues can lower the threshold for deferring to AI recommendations more generally. This reliance-enhancing tendency does not systematically promote correct consultations, but it increases the likelihood of changing decisions in situations where users already experienced uncertainty.

6.2. Contributions to IS and decision sciences research

With our focus on calibrated reliance, our study makes three contributions to IS and Decision Sciences research on ADM under high uncertainty. In doing so, we set our study apart from research on algorithmic aversion, which conceptualizes reliance on AI recommendations as inherently beneficial [e.g., 16–18].

First, we reconceptualize explainability in ADM contexts with fallible (i.e., probabilistic and therefore inherently subject to errors) AI recommendations as a cognitively ambivalent intervention which may trigger not only intended but also unintended cognitive effects. Existing literature has treated transparency as an inherently beneficial property that provides the capability to improve human judgment (i.e., intended cognitive effect) which leads to higher decision performance [29,56,57]. Our central proposition and findings challenge this assumption by proposing and showing that, when AI recommendations are fallible, explainability can simultaneously generate beneficial and detrimental reliance outcomes. This dual effect helps to explain the inconsistent findings regarding performance degradation in the existing IS and Decision Sciences literature. Importantly, this reconceptualization (e.g., ambivalent consequences) extends the current theoretical understanding of automation bias. Prior work has largely associated automation bias with opaque or highly automated systems [1,36,78]. Our results demonstrate that even in transparent ADM settings with explainable but fallible AI recommendations, explainability itself can become a source of biased reliance by amplifying heuristic decision processes.

Second, we contribute to IS and Decision Sciences literature through context-specific theorizing [89], by explaining how reliance mechanisms operate in ADM settings characterized by high uncertainty. Baird and Maruping's [26] notion of agentic IS suggests that decision-makers do not merely interact with AI as a passive informational tool, but perceive it as an entity capable of assuming decision-making responsibility under uncertainty. We extend this perspective to XAI-supported ADM by showing that explanations amplify such perceptions. More specifically, we demonstrate that competence-based trust and responsibility attribution are key drivers of reliance calibration by shaping how users interpret and act upon explainable but fallible AI recommendations. This expands our understanding of the boundary conditions under which automation bias emerges in AI-enabled decision support. From this perspective, automation bias in XAI-supported ADM should not only be understood as an episodic judgment error at the level of individual recommendations. Rather, it may reflect a broader transition from recommendation-level reliance to task-level cognitive delegation. As users repeatedly experience an AI system as transparent, competent, and responsible, they may become less inclined to scrutinize individual outputs and more willing to delegate entire cognitive tasks to the system. This insight is particularly relevant for agentic AI settings, where AI systems increasingly act on behalf of users rather than merely provide decision inputs. In such contexts, XAI may unintentionally facilitate automation bias by making delegation appear justified and controllable, thereby shifting the risk from following erroneous recommendations to delegating tasks without sufficient monitoring.

Third, our findings contribute to Decision Sciences design theory in ADM settings characterized by uncertain and fallible AI recommendations by conceptualizing explainability as a reliance-shaping design mechanism rather than as a uniformly performance-enhancing transparency feature. Prior research has predominantly emphasized the intended benefits of XAI-induced transparency [29] and often treats explanation provision as a uniformly beneficial design intervention. In contrast, our results demonstrate that explainability can produce unintended consequences by amplifying competence-based trust and shifting responsibility, thereby reshaping reliance calibration in ways that may increase susceptibility to erroneous recommendations. In doing so, we challenge one-size-fits-all approaches to explainability [43] and argue that explanation design must explicitly account for the fallibility of AI recommendations and the cognitive processes they can activate. Rather than assuming that more transparency inherently improves decision performance, system design should recognize heterogeneity in how users process explanations. Explanations may either be processed as inputs for deliberative scrutiny or as heuristic cues that signal competence and legitimacy. Consequently, explainability mechanisms and related safeguards should be designed not only to enhance understanding but also to regulate reliance behavior. Importantly, the requirements for such safeguards may vary with the degree of agency exhibited by the AI system. In settings centered on individual recommendations, safeguards primarily need to support the calibrated evaluation of a specific AI output and prevent its uncritical acceptance at a discrete decision point. In proactive, multi-step agentic settings, by contrast, safeguards must extend beyond recommendation-level scrutiny and support effective human control across the entire action trajectory. Users need to remain informed about intermediate actions, accumulating uncertainty, and possible error propagation, while retaining meaningful opportunities to intervene before the system takes consequential actions.

6.3. Practical contributions

Our study also offers important implications for practitioners and legislators. Our findings demonstrate that explainability in ADM should not be treated as a universally beneficial design feature but as a behavioral intervention that reshapes reliance. In particular, our results indicate that XAI-induced transparency may increase users' perceived competence of the AI and foster responsibility diffusion toward the system, both of which can contribute to overreliance and increase the likelihood of negative consultations. By linking dual-process mechanisms to prescriptive design logic, we extend prior recommendations on system design by suggesting that effective XAI design should not merely aim to maximize transparency and system acceptance but should incorporate targeted safeguards at selected intervention timings

throughout the decision process to attenuate these unintended cognitive effects while preserving the deliberative benefits of explanations (see Table 4). In environments characterized by fallible AI recommendations, explainability should therefore be aligned with users' cognitive characteristics and reliance tendencies (e.g., cognitive fit), ensuring that transparency enhances decision performance without inadvertently increasing users' susceptibility to automation bias.

The critical role of responsibility in human-AI collaborative decision-making also presents significant challenges for existing legal frameworks, as current legal systems do not recognize AI as an entity capable of bearing legal responsibility. In addition, policy initiatives that mandate "more explainability" may be insufficient if transparency increases overreliance. Regulatory guidance should therefore emphasize calibrated reliance rather than transparency alone.

6.4. Limitations and opportunities for future research

There are, however, some limitations that offer fruitful avenues for future research. More specifically, our findings should be interpreted within the boundary conditions of the present empirical setting, which involve a single explanation modality, a single task with low-stakes decision context, limited process-related measures and a fixed AI accuracy profile. First, XAI was examined in the form of local explanations provided by SHAP. While this decision was made to demonstrate the general effect of XAI on automation bias, one should leverage the full richness of available XAI approaches (e.g., counterfactual, natural language, rule-based) to test for additional insights into the potential impact of diverse design elements on automation bias.

Second, the findings of the study may be limited to the specific task of price classification, which represents a relatively low-stakes context. Therefore, caution should be exercised when generalizing the results to other settings. To increase the generalizability, the study should also be replicated in other task contexts, where decision-makers perceive an actual risk that could influence their tendency toward automation bias, such as healthcare or financial auditing. Building on this, future research should explicitly examine how these mechanisms operate in more advanced forms of human-AI interaction. In addition, we acknowledge that longitudinal or repeated-interaction designs may reveal recursive dynamics that shape human-AI interaction over time. Longitudinal work could examine whether repeated exposure to XAI gradually shifts users from active scrutiny to cognitive delegation. In this sense, automation bias may not only reflect isolated decision errors but also a transitional cognitive phenomenon in which users adapt to increasingly capable and autonomous systems. Future research should investigate how this transition unfolds in agentic AI environments and identify the conditions under which it improves or undermines decision quality.

Third, we do not directly measure System 1 or System 2 processing. Instead, we rely on dual-process theory as an interpretive framework to examine reliance-relevant perceptions and decision changes. Future research should therefore incorporate more direct process-related measures, such as eye-tracking, mouse-tracking, or neurophysiological indicators. Such process-related measures would help distinguish more precisely between different cognitive pathways underlying the same observable reliance behavior by capturing users' attention allocation, explanation inspection, advice comparison, and signs of conflict or effortful reconsideration. Finally, we deliberately set the AI's accuracy to 50% by applying adversarial perturbations to ensure a balanced distribution across the confusion matrix. This design choice may have made the AI's fallibility more salient, potentially lowering perceived competence, trust, and reliance. However, our results provide no clear evidence for this concern; participants still overrode their initial judgments in favor of the AI's recommendations. Nevertheless, future studies should shed light on different or even dynamic AI decision capabilities to reveal additional insights into the calibration of human-AI collaboration.

Table 4
Linking of unintended cueing effects and design-safeguards.

Cueing Effect	Design-Safeguards	Timing of Intervention
Perceived competence-based trust	Explicitly communicate limitations and known failure boundaries of explanations (i.e., explanations as approximations)	Across-task interaction
	Visually separate explanations from endorsement cues to avoid persuasive framing	Across-task interaction
	Include epistemic warning signals (e.g., "The explanation may be imprecise in cases like this.")	AI-exposure stage
	Cognitive forcing functions that make users form and record an independent initial judgment before seeing the AI explanation	Pre-AI exposure
Attribution of responsibility	Adaptive reliance-aware explainable user interface that tailors explanations to observed reliance patterns (e.g., by using contrastive explanations or displaying prominent uncertainty markers)	AI-exposure stage
	Explicit responsibility framing (e.g., "Final decision authority remains with you.")	Post-AI exposure
	Adaptive user interface that requires users to explicitly justify why they accept XAI explanations and thereby change their initial judgment	Post-AI exposure

CRediT authorship contribution statement

Tizian Matschak: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Ilja Nastjuk:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Investigation, Formal analysis, Data curation, Conceptualization. **Simon Trang:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Investigation, Funding acquisition. **Jose Benitez:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Resources, Project administration, Investigation, Funding acquisition.

Declaration of competing interest

None of the authors has any financial or personal relationships that influenced the development of this work.

Acknowledgments

We thank the sponsorship received from the Regional Government of Andalusia, the Government of Spain, and the European Regional Development Fund (European Union) (Research Projects C-SEJ-148-UGR23, PID2021-124725NB-I00, and PID2021-124396NB-I00).

Appendix A. Construct measurement

Perceived transparency (adapted from [43,50]): "The AI clearly presented its reasoning process to me", "It was easy for me to see how this AI generates recommendations", "It was easy for me to understand the inner workings of this AI", and "I could understand why and how this AI makes the recommendations". Competence-based trust (adapted from [44]): "The AI is competent in providing car price estimations", "The AI performs its role of giving car price estimations very well", "Overall, the AI is capable and proficient in giving car price estimations", and "In general, the AI is very knowledgeable about giving car price estimations". Attribution of responsibility (adapted from [80],

item 4 self-developed): “I believe that the task outcome is the responsibility (contribution) of the AI”, “I feel that this outcome is largely due to the AI's estimation”, “I feel that this outcome is strongly related to the AI”, and “The AI should get credit for most of what was accomplished on this task”. AI knowledge: “My knowledge on AI is …”. AI use: “I have used AI …”. Negative and positive consultations were calculated based on actual user decisions (see Table 1).

Data availability

Data will be made available on request.

References

[1] E. Jussupow, K. Spohrer, A. Heinzl, J. Gawlitza, Augmenting medical diagnosis decisions? An investigation into physicians' decision-making process with artificial intelligence, Inf. Syst. Res. 32 (2021) 713–735, https://doi.org/10.1287/isre.2020.0980.

[2] A. Klingbeil, C. Grützner, P. Schreck, Trust and reliance on AI—an experimental study on the extent and costs of overreliance on AI, Comput. Hum. Behav. 160 (2024) 108352.

[3] J.D. Lee, K.A. See, Trust in automation: designing for appropriate reliance, Hum. Factors 46 (2004) 50–80.

[4] J. Schoeffer, J. Jakubik, M. Vössing, N. Kühl, G. Satzger, AI reliance and decision quality: fundamentals, interdependence, and the effects of interventions, J. Artif. Intell. Res. 82 (2025) 471–501.

[5] J. Ochmann, L. Michels, V. Tiefenbeck, C. Maier, S. Laumer, Perceived algorithmic fairness: an empirical study of transparency and anthropomorphism in algorithmic recruiting, Inf. Syst. J. 34 (2024) 384–414.

[6] A. Mehrabian, J.A. Russell, An Approach to Environmental Psychology, MIT Press, 1974.

[7] K. Coussement, M.Z. Abedin, M. Kraus, S. Maldonado, K. Topuz, Explainable AI for enhanced decision-making, Decis. Support. Syst. 184 (2024) 114276.

[8] Z. Buçinca, M.B. Malaya, K.Z. Gajos, To trust or to think: cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making, Proc. ACM Hum. Comput. Interact. 5 (2021) 1–21.

[9] K. Bauer, M. von Zahn, O. Hinz, Expl (AI) ned: the impact of explainable artificial intelligence on users' information processing, Inf. Syst. Res. 34 (2023) 1582–1602.

[10] X. Ning, Y. Lu, W. Li, S. Gupta, How transparency affects algorithmic advice utilization: the mediating roles of trusting beliefs, Decis. Support. Syst. 183 (2024) 114273.

[11] M. Naiseh, D. Al-Thani, N. Jiang, R. Ali, How the different explanation classes impact trust calibration: the case of clinical decision support systems, Int. J. Hum. Comput. Stud. 169 (2023) 102941.

[12] J. Cecil, E. Lermer, M.F. Hudecek, J. Sauer, S. Gaube, Explainability does not mitigate the negative impact of incorrect AI advice in a personnel selection task, Sci. Rep. 14 (2024) 9736.

[13] T. Wolf, S. Trang, W.H. Weiger, M. Trenz, The technology-behavioral compensation effect: unintended consequences of health technology adoption, J. Inf. Technol. 39 (2024) 521–546.

[14] J. van der Waa, E. Nieuwburg, A. Cremers, M. Neerincx, Evaluating XAI: a comparison of rule-based and example-based explanations, Artif. Intell. 291 (2021) 103404.

[15] D. Kahneman, S. Frederick, Representativeness revisited: Attribute substitution in intuitive judgment, in: T. Gilovich, D.W. Griffin, D. Kahneman (Eds.), Heuristics and Biases: The Psychology of Intuitive Judgment, Cambridge University Press, New York, 2002, pp. 49–81.

[16] H. Mahmud, A.N. Islam, X.R. Luo, P. Mikalef, Decoding algorithm appreciation: unveiling the impact of familiarity with algorithms, tasks, and algorithm performance, Decis. Support. Syst. 179 (2024) 114168.

[17] O. Turel, S. Kalhan, Prejudiced against the machine? Implicit associations and the transience of algorithm aversion, MIS Q. 47 (2023) 1369–1394.

[18] K. Kawaguchi, When will workers follow an algorithm? A field experiment with a retail business, Manag. Sci. 67 (2021) 1670–1695.

[19] E.Ö. Dogru, N.C. Krämer, Investigating appropriate reliance on AI-based decision support systems: the role of expertise, trust, and self-confidence, J. Decis. Syst. 34 (2025) 2593251.

[20] H. Sun, W. Wen, J.B. Thatcher, M. Chau, S. Brown, Editor's comments: rebalancing novelty with rigor and relevance in information systems research, MIS Q. 49 (2025) iii–xi.

[21] A. Adadi, M. Berrada, Peeking inside the black-box: a survey on explainable artificial intelligence (XAI), IEEE Access 6 (2018) 52138–52160.

[22] M. Pasupuleti, Human-in-the-loop AI: enhancing transparency and accountability, Int. J. Acad. Ind. Res. Innov. 5 (2025) 574–585.

[23] B.J. Dietvorst, J.P. Simmons, C. Massey, Algorithm aversion: people erroneously avoid algorithms after seeing them err, J. Exp. Psychol. Gen. 144 (2015) 114–126.

[24] J.M. Logg, J.A. Minson, D.A. Moore, Algorithm appreciation: people prefer algorithmic to human judgment, Organ. Behav. Hum. Decis. Process. 151 (2019) 90–103.

[25] E. Kokje, E. Lermer, A.-K. Kleine, S. Gaube, AI-augmented decision-making in face matching: comparing concurrent and non-concurrent advice presentation, Cogn. Res. 11 (2026) 11.

[26] A. Baird, L.M. Maruping, The next generation of research on IS use: a theoretical framework of delegation to and from agentic IS artifacts, MIS Q. 45 (2021) 315–341.

[27] K. Goddard, A. Roudsari, J.C. Wyatt, Automation bias: a systematic review of frequency, effect mediators, and mitigators, J. Am. Med. Inform. Assoc. 19 (2012) 121–127.

[28] J.W. Burton, M. Stein, T.B. Jensen, A systematic review of algorithm aversion in augmented decision making, J. Behav. Decis. Mak. 33 (2020) 220–239.

[29] A. Rai, Explainable AI: from black box to glass box, J. Acad. Mark. Sci. 48 (2020) 137–141, https://doi.org/10.1007/s11747-019-00710-5.

[30] D. Dellermann, P. Ebel, M. Söllner, J.M. Leimeister, Hybrid intelligence, business & information, Syst. Eng. 61 (2019) 637–643.

[31] A.B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, Explainable Artificial Intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI, Inf. Fusion 58 (2020) 82–115.

[32] G. Bansal, T. Wu, J. Zhou, R. Fok, B. Nushi, E. Kamar, M.T. Ribeiro, D. Weld, Does the whole exceed its parts? The effect of AI explanations on complementary team performance, in: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, 2021, pp. 1–16.

[33] M. Schemmer, N. Kühl, C. Benz, G. Satzger, On the influence of explainable AI on automation bias, in: Proceedings of the Thirtieth European Conference on Information Systems, Timişoara, Romania, 2022.

[34] M. Du, N. Liu, X. Hu, Techniques for interpretable machine learning, Commun. ACM 63 (2019) 68–77.

[35] F. Poursabzi-Sangdeh, D.G. Goldstein, J.M. Hofman, J.W. Wortman Vaughan, H. Wallach, Manipulating and measuring model interpretability, in: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, 2021, pp. 1–52.

[36] L.J. Skitka, K.L. Mosier, M. Burdick, Does automation bias decision-making? Int. J. Hum. Comput. Stud. 51 (1999) 991–1006.

[37] A. Tversky, D. Kahneman, Judgment under uncertainty: heuristics and biases, Science 185 (1974) 1124–1131.

[38] T. Kliegr, Š. Bahník, J. Fürnkranz, A review of possible effects of cognitive biases on interpretation of rule-based machine learning models, Artif. Intell. 295 (2021) 103458.

[39] Z. Lei, D. Yin, H. Zhang, Deliberative or automatic: disentangling the dual processes behind the persuasive power of online word-of-mouth, MIS Q. 49 (2025) 331–346.

[40] P.L. Moravec, A. Kim, A.R. Dennis, Appealing to sense and sensibility: system 1 and system 2 interventions for fake news on social media, Inf. Syst. Res. 31 (2020) 987–1006.

[41] J.S. Lerner, P.E. Tetlock, Accounting for the effects of accountability, Psychol. Bull. 125 (1999) 255.

[42] J. Kruger, D. Wirtz, L. Van Boven, T.W. Altermatt, The effort heuristic, J. Exp. Soc. Psychol. 40 (2004) 91–98.

[43] W. Wang, I. Benbasat, Empirical assessment of alternative designs for enhancing different types of trusting beliefs in online recommendation agents, J. Manag. Inf. Syst. 33 (2016) 744–775.

[44] D.H. McKnight, V. Choudhury, C. Kacmar, The impact of initial consumer trust on intentions to transact with a web site: a trust building model, J. Strateg. Inf. Syst. 11 (2002) 297–323, https://doi.org/10.1016/S0963-8687(02)00020-3.

[45] D. Horneber, S. Laumer, Algorithmic accountability, business & information, Syst. Eng. 65 (2023) 723–730, https://doi.org/10.1007/s12599-023-00817-8.

[46] V. Khatri, B.M. Samuel, A.R. Dennis, System 1 and System 2 cognition in the decision to adopt and use a new technology, Inf. Manag. 55 (2018) 709–724.

[47] A. Salatino, A. Prével, E. Caspar, S. Lo Bue, Influence of AI behavior on human moral decisions, agency, and responsibility, Sci. Rep. 15 (2025) 12329.

[48] C. Anthony, B.A. Bechky, A.-L. Fayard, “Collaborating” with AI: taking a system view to explore the future of work, Organ. Sci. 34 (2023) 1672–1694.

[49] D. Shin, User perceptions of algorithmic decisions in the personalized AI system: perceptual evaluation of fairness, accountability, transparency, and explainability, J. Broadcast. Electron. Media 64 (2020) 541–565.

[50] P. Hamm, M. Klesel, P. Coberger, H.F. Wittmann, Explanation matters: an experimental study on explainable AI, Electron. Mark. 33 (2023) 17.

[51] V. Sullivan, K. Weger, Transparency and explainability in AI-assisted decision making: effects on trust, perceived reliability, confidence, and ease of understanding, in: Proceedings of the Human Factors and Ergonomics Society Annual Meeting, SAGE Publications, Los Angeles, CA, 2025, pp. 1827–1833.

[52] V.M. Ngo, The AI transparency dilemma: when more is less for trust and adoption, Inf. Discov. Deliv. (2025), https://doi.org/10.1108/IDD-03-2025-0056.

[53] P. Wang, H. Ding, The rationality of explanation or human capacity?, in: Understanding the Impact of Explainable Artificial Intelligence on Human-AI Trust and Decision Performance, Information Processing & Management vol. 61, 2024, p. 103732.

[54] Y. Litvinova, P. Mikalef, X. Luo, Framework for human–XAI symbiosis: extended self from the dual-process theory perspective, J. Bus. Anal. 7 (2024) 224–255.

[55] H. Kaur, H. Nori, S. Jenkins, R. Caruana, H. Wallach, J. Wortman Vaughan, Interpreting interpretability: understanding data scientists' use of interpretability tools for machine learning, in: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, 2020, pp. 1–14.

[56] B.R. Kim, K. Srinivasan, S.H. Kong, J.H. Kim, C.S. Shin, S. Ram, ROLEX: a novel method for interpretable machine learning using robust local explanations, MIS Q. 47 (2023) 1303–1332.

[57] L. Pumplun, F. Peters, J.F. Gawlitza, P. Buxmann, Bringing machine learning systems into clinical practice: a design science approach to explainable machine

learning-based clinical decision support systems, J. Assoc. Inf. Syst. 24 (2023) 953–979.
[58] T. Miller, Explanation in artificial intelligence: insights from the social sciences, Artif. Intell. 267 (2019) 1–38.
[59] Y. Zhang, Q.V. Liao, R.K. Bellamy, Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making, in: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 2020, pp. 295–305.
[60] K.A. Hoff, M. Bashir, Trust in automation: integrating empirical evidence on factors that influence trust, Hum. Factors 57 (2015) 407–434.
[61] D.H. McKnight, M. Carter, J.B. Thatcher, P.F. Clay, Trust in a specific technology: an investigation of its components and measures, ACM Trans. Manag. Inf. Syst. 2 (2011) 1–25.
[62] S. Thiebes, S. Lins, A. Sunyaev, Trustworthy artificial intelligence, Electron. Mark. 31 (2021) 447–464.
[63] Z. Buçinca, P. Lin, K.Z. Gajos, E.L. Glassman, Proxy tasks and subjective measures can be misleading in evaluating explainable AI systems, in: Proceedings of the 25th International Conference on Intelligent User Interfaces, 2020, pp. 454–464.
[64] M.K. Lee, Understanding perception of algorithmic decisions: fairness, trust, and emotion in response to algorithmic management, Big Data Soc. 5 (2018), 2053951718756684.
[65] K.T. Wynne, J.B. Lyons, An integrative model of autonomous agent teammate-likeness, Theor. Issues Ergon. Sci. 19 (2018) 353–374.
[66] J.R. Josephson, S.G. Josephson, Abductive Inference: Computation, Philosophy, Technology, Cambridge University Press, 1996.
[67] J.M. Durán, K.R. Jongsma, Who is afraid of black box algorithms? On the epistemological and ethical basis of trust in medical AI, J. Med. Ethics 47 (2021) 329–335.
[68] M. Coeckelbergh, Virtual moral agency, virtual moral responsibility: on the moral significance of the appearance, perception, and performance of artificial agents, AI & Soc. 24 (2009) 181–189.
[69] H. Chockler, J.Y. Halpern, Responsibility and blame: a structural-model approach, J. Artif. Intell. Res. 22 (2004) 93–115.
[70] R.E. Petty, J.T. Cacioppo, The Elaboration Likelihood Model of Persuasion, Springer, 1986.
[71] E. Jussupow, I. Benbasat, A. Heinzl, Why are we averse towards algorithms? A comprehensive literature review on algorithm aversion, in: Proceedings of the 28th European Conference on Information Systems (ECIS), 2020, pp. 1–16.
[72] R. Riedl, Is trust in artificial intelligence systems related to user personality? Review of empirical evidence and future research directions, Electron. Mark. 32 (2022) 2021–2051.
[73] M.T. Dzindolet, L.G. Pierce, H.P. Beck, L.A. Dawe, Misuse and disuse of automated aids, in: Proceedings of the Human Factors and Ergonomics Society Annual Meeting, SAGE Publications, 1999, pp. 339–343.
[74] I. Simonson, Choice based on reasons: the case of attraction and compromise effects, J. Consum. Res. 16 (1989) 158–174.
[75] M.D. Brtek, S.J. Motowidlo, Effects of procedure and outcome accountability on interview validity, J. Appl. Psychol. 87 (2002) 185.
[76] N.P. Mero, R.M. Guidice, S. Werner, A field study of the antecedents and performance consequences of perceived accountability, J. Manag. 40 (2014) 1627–1652.
[77] C.M. Angst, R. Agarwal, Adoption of electronic health records in the presence of privacy concerns: the elaboration likelihood model and individual persuasion, MIS Q. (2009) 339–370.
[78] K.L. Mosier, L.J. Skitka, M. Dunbar, L. McDonnell, Aircrews and automation bias: the advantages of teamwork? Int. J. Aviat. Psychol. 11 (2001) 1–14.
[79] S.M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, Adv. Neural Inf. Proces. Syst. 30 (2017) 4765–4774.
[80] B. Yue, H. Li, The impact of human-AI collaboration types on consumer evaluation and usage intention: a perspective of responsibility attribution, Front. Psychol. 14 (2023) 1277861.
[81] S. Palan, C. Schitter, Prolific.Ac—a subject pool for online experiments, journal of behavioral and experimental, Finance 17 (2018) 22–27.
[82] D.S. Soper, A-Priori Sample Size Calculator for Structural Equation Models [Software]. https://www.danielsoper.com/statcalc, 2025.
[83] R.P. Bagozzi, Y. Yi, On the evaluation of structural equation models, J. Acad. Mark. Sci. 16 (1988) 74–94, https://doi.org/10.1007/BF02723327.
[84] C. Fornell, D.F. Larcker, Structural equation models with unobservable variables and measurement error: algebra and statistics, J. Mark. Res. 18 (1981) 382–388.
[85] J.C. Anderson, D.W. Gerbing, Structural equation modeling in practice: a review and recommended two-step approach, Psychol. Bull. 103 (1988) 411.
[86] R.P. Bagozzi, Y. Yi, On the use of structural equation models in experimental designs, J. Mark. Res. 26 (1989) 271–284.
[87] Y. Rosseel, Lavaan: an R package for structural equation modeling, J. Stat. Softw. 48 (2012) 1–36, https://doi.org/10.18637/jss.v048.i02.
[88] Z. Kunda, The case for motivated reasoning, Psychol. Bull. 108 (1990) 480.
[89] W. Hong, F.K. Chan, J.Y. Thong, L.C. Chasalow, G. Dhillon, A framework and guidelines for context-specific theorizing in information systems research, Inf. Syst. Res. 25 (2014) 111–136.

Tizian Matschak (tizian.matschak@uni-goettingen.de) is a Research Associate at the Department of Information Systems (IS) at the University of Paderborn, Germany, and a Ph.D. candidate within the Research Group on Information Security and Compliance at the University of Goettingen, Germany. He earned his Master's degree in IS. Currently, his research is centered around human-AI interaction, fraud detection, and the information security dimensions of AI. His work has been presented at various international conferences, including the *European Conference on Information Systems, Americas Conference on Information Systems, Pacific Asia Conference on Information Systems, Hawaii International Conference on System Sciences, and International Conference on Design Science Research in Information Systems and Technology*.

Ilja Nastjuk (inastjuk@icms.edu.au) is an Associate Professor at ICMS, Sydney, Australia. He earned a Ph.D. in IS from the University of Goettingen and a Ph.D. in Accounting and Corporate Governance from Macquarie University. His research interests span the influence of technology on stress and human behavior, the management of information security, and the adoption of self-driving cars. His work has been published in numerous peer-reviewed journals and conference proceedings, such as the *European Journal of Information Systems*, *Computers & Security*, *Technological Forecasting and Social Change*, *Electronic Markets*, *Transportation Research Part D: Transport and Environment*, *Transportation Research Part F: Traffic Psychology*, *International Conference on Information Systems*, and *European Conference on Information Systems*.

Simon Trang (simon.trang@uni-paderborn.de) is Professor of IS, esp. Sustainability, at the University of Paderborn. He received his Ph.D. in Management Science, specializing in Management Information Systems, from the University of Goettingen. His work focuses on information security management, privacy, and sustainable IS. His research has been published in outlets such as the *Information Systems Research*, *Journal of the Association for Information Systems*, the *European Journal of Information Systems*, *Information Systems Frontiers*, and others. He has served as an Associate Editor for *Information & Management*.

Jose Benitez is Professor of IS and Bridgestone Endowed Chair in International Business in the Department of Information Systems and Business Analytics at the Ambassador Crawford College of Business and Entrepreneurship, Kent State University, Kent, Ohio, USA. He served as the Inaugural Department Chair of Information Systems and Business Analytics from 2023 to 2025. Before joining Kent State University, he held senior academic appointments in Europe, including positions such as Endowed Chair Professor, Department Chair, and Associate Dean of Faculty and Programs. His research and teaching focus on how emergent digital technologies transform individuals, teams, organizations, and society, and on how individuals and organizations can leverage these technologies responsibly and effectively to create value. His work has appeared in more than 80 articles in premier journals, including *MIS Quarterly, Information Systems Research, Journal of Operations Management, Production and Operations Management, Journal of International Business Studies, Journal of Management Information Systems, Journal of the Association for Information Systems, European Journal of Information Systems, Information Systems Journal, Journal of Strategic Information Systems, Journal of Information Technology, Information & Management, Decision Support Systems, IEEE Transactions on Engineering Management,* and *Decision Sciences*. Professor Benitez is an AIS Distinguished Member Cum Laude and recipient of the AIS Sandra Slaughter Service Award. He serves as Senior Editor of *European Journal of Information Systems, Information & Management,* and *Decision Support Systems*; Associate Editor of the *Journal of the Association for Information Systems*; and member of the Editorial Review Board of *Information Systems Research*. He has also served as Guest Editor of *Decision Sciences* and will serve as Program Co-Chair of the *European Conference on Information Systems 2028*. Professor Benitez was ranked No. 1 worldwide in 2025 for the number of publications in the AIS Senior Scholars' List of 11 Premier Journals. He was also recognized in the 2025 Top 2% Global Scientists list published by Stanford University and Elsevier, where he was ranked No. 50 in the USA in the IS discipline. Furthermore, he was named a 2025 Clarivate Highly Cited Researcher, placing him among the top 1% of researchers in Web of Science citations over the past decade. He can be contacted at jbenite1@kent.edu.