Information Systems Frontiers
https://doi.org/10.1007/s10796-026-10716-4

Effects of Explanations in Human-AI Interaction: A Systematic Review and Framework for Future Research on Explainable AI

Philipp Reinhard[1] · **Mahei Manhai Li**[1,2] · **Christoph Peters**[3] · **Jan Marco Leimeister**[1,2]

Received: 29 August 2024 / Accepted: 2 March 2026

Abstract
Advancements in artificial intelligence (AI), particularly in generative AI and agentic AI, have intensified challenges related to transparency and explainability. While explainable AI (XAI) research has evolved to facilitate human interaction with such complex, black-box systems, research and practice lack clarity on the causal chain linking explanations, user perceptions, and real-world outcomes, a relationship that remains conceptually fragmented. To address this gap, we conducted a systematic literature review of 107 experimental user studies on XAI. We developed a conceptual framework guided by the stimulus-organism-response-consequences (S-O-R-C) model to systematize current human-XAI research and examine how users respond to explanations. Our study contributes to the literature by clarifying how explanations shape user interactions and downstream effects in real-world settings. We propose five research directions to help navigate the challenges of emerging AI systems (e.g., LLMs, AI agents) and evolving human-AI delegation.

Keywords Explainable AI (XAI) · Artificial intelligence · Generative artificial intelligence · Human-AI delegation · Systematic literature review · Research agenda · Stimulus-organism-response-consequences (S-O-R-C)

1 Introduction

Artificial intelligence (AI) is transforming the way humans perform tasks and make decisions across professional and everyday contexts. With the advent of generative artificial intelligence (GenAI) such as ChatGPT and DALL-E, the delegation to AI-based systems has now reached mainstream appeal and is thus steadily increasing the number and intensity of human-AI interactions (Dell'Acqua et al., 2023; Li et al., 2023; Brynjolfsson et al., 2025). Enterprises expect to benefit from GenAI, especially large language models (LLMs), by delegating simple tasks to AI agents and improving employee performance through augmentation. For example, in customer service, GenAI has enabled workers to resolve 15% more issues on average (Brynjolfsson et al., 2025).

However, these systems still come with inherent challenges such as biases, hallucinations, or lack of transparency (Hannigan et al., 2024), which can undermine their reliability, trustworthiness, and responsible use (Meske et al., 2022) and hamper human-AI performance. For example, state-of-the-art LLMs produce highly persuasive responses (Vaswani et al., 2017), yet introduce the challenge of non-factual outputs (i.e., hallucinations) (Huang et al., 2025), which goes beyond conventional AI error types (Poursabzi-Sangdeh et al., 2021). The lack of transparency and explainability of GenAI's underlying decision rationales (Lee & Ram, 2023) makes it difficult for users to detect erroneous outputs, thereby limiting their ability to rely on the system appropriately.

To address these challenges, research has increasingly turned to explainable AI (XAI) as a potential solution, offering methods to make black-box systems more transparent and, in turn, more trustworthy (Brasse et al., 2023). Even before the rise of GenAI and LLMs, XAI had been explored as a way to enhance human–AI interaction by equipping

✉ Philipp Reinhard
philipp.reinhard@uni-kassel.de

[1] Research Center for IS Design (ITeG), University of Kassel, Pfannkuchstr. 1, Kassel DE-34121, Germany

[2] Institute of Information Systems and Digital Business, University of St.Gallen, Müller-Friedberg-Strasse 8, St.Gallen CH-9000, Switzerland

[3] esp. Digital Process Management, University of the Bundeswehr Munich, Werner-Heisenberg-Weg 39, Neubiberg DE-85579, Germany

Published online: 05 May 2026

users with meaningful explanations (e.g., Silva et al., 2023; Vössing et al., 2022). According to Arrieta et al. (2020, p. 85), XAI *"is one that produces details or reasons to make its functioning clear or easy to understand"*. Consequently, XAI acts as a stimulus for building trust and reliance in AI, thereby promoting greater utilization and acceptance of AI systems (Vössing et al., 2022). For example, in clinical decision support systems, XAI can visually highlight the most crucial places in CT images to explain its advised diagnosis (Herm, 2023).

Although the literature provides a good foundation regarding available XAI methods (what) (Chromik & Butz, 2021; Schwalbe & Finzel, 2023), relevant stakeholders (who) (Langer et al., 2021; Ferreira & Monteiro, 2020), and their intended goals (why) (Laato et al., 2022; Brasse et al., 2023), the relationships among explanations, users' perceptions and behavior, and real-world consequences remain conceptually fragmented. This lack of conceptual integration hinders the development of causal models that account for how explanations shape decision-making and influence outcomes in applied settings. Against this backdrop, we challenge prevailing perspectives that conceptualize human–XAI interaction either at the level of the XAI system as a monolithic entity (Haque et al., 2023) or through isolated features (Chromik & Butz, 2021), typically characterized by their type, representational form, or modality. Understanding the underlying mechanisms requires connecting individual explanation features to the broader AI system context, enabling a more integrated view of how design decisions influence user perceptions and system explainability through key theoretical constructs. Finally, we reconsider the traditional emphasis on high-level stakeholder explanation goals that predominantly guide system design and intended outcomes. Instead, we advocate for theory-driven XAI research that links explanations to real-world outcomes, aiming to foster more responsible and effective AI use concerning the growing societal impact of human-AI delegation and the rapid advances in technologies like LLM and agentic AI (Dennehy et al., 2023). We aim to address these challenges by developing a conceptual framework for human-XAI interactions that identifies higher-order theoretical structures (Paré et al., 2015) explaining how explanations shape user behavior. Drawing on insights from experimental studies, our results guide future research as AI technologies continue to evolve. Accordingly, the focus of our study leads to the following research questions:

RQ1 How can existing experimental research on XAI be synthesized to reveal the mechanisms through which explanations influence user perceptions, behaviors, and outcomes?

RQ2 What directions should guide XAI research in response to emerging AI capabilities (e.g., LLMs, AI agents, and human-AI delegation)?

To answer our research questions, we analyzed 107 peer-reviewed papers on XAI from the fields of Information Systems (IS) and Human-Computer Interaction (HCI), drawn from an initial screening of 853 papers, identified through a targeted search of 20 high-quality outlets and supplemented by forward and backward search. We aimed to aggregate prior experimental research on explanations in human-AI interaction to summarize findings by performing a theoretical review of the resulting sample (Paré et al., 2015). Our analysis focused on quantitative, experimental user studies, allowing us to identify frequently studied constructs and discover recurring relationships between independent and dependent variables. Our rationale for focusing on user-centric experimental studies lies in their ability to generate controlled, theory-relevant evidence on how explanations influence user perceptions, behaviors, and outcomes, including boundary conditions and moderating factors. This approach complements existing conceptual, technical, and qualitative reviews (e.g., Brasse et al., 2023; Haque et al., 2023) by offering deeper insights into causal mechanisms and supporting the development of theory-driven XAI research.

Our study synthesizes empirical findings and offers a structured analysis of how explanations influence cognitive and behavioral outcomes in human–XAI interaction. To advance theoretical integration in the field, we introduced the stimulus-organism-response-consequences (S-O-R-C) model (Mehrabian & Russell, 1974; Goldfried & Sprafkin, 1976) as a lens for exploring the psychological mechanisms underlying human–XAI engagement. Our synthesis examines how different explanation types and modalities (stimuli) influence user perceptions (organism), such as understanding, trust, transparency, and cognitive load, which in turn shape user responses, including reliance and behavioral intention, as well as task-related consequences like performance, engagement, and error detection. This structured perspective is especially valuable in a field where prior reviews lack a clear theoretical basis and do not clarify how explanation design influences outcomes. We adopt the concepts of explanans (i.e., the elements providing explanations) and explanandum (i.e., the elements being explained) from the long-standing philosophical tradition on reasoning (Hempel & Oppenheim, 1948). Our systematization

particularly emphasizes the connection between the explanans and the explanandum. This relationship is reflected in the organism component, where we show that both the perception of individual explanation features (e.g., types, combinations, modalities) and the perception of the broader XAI system (e.g., the model, decision, reasoning traces) call for deeper, theory-driven investigation and integration. Finally, the resulting conceptual framework consolidates prior findings and introduces five key research directions to advance experimental XAI: (1) grounding explanation design in theory, (2) developing novel explanation types for GenAI and agentic AI systems, (3) clarifying the link between explanans and explanandum, (4) tailoring explanations to diverse user needs, and (5) examining engagement and long-term effects of XAI on human–AI delegation. These directions respond to the rising complexity of AI technologies and evolving human–AI delegation.

2 Prior Literature Reviews on XAI

The scope of this systematic literature review is related to prior reviews focusing on the interaction between humans and XAI – human-XAI interaction – with regard to explanations, irrespective of the underlying system and domain. An overview of the related prior reviews of human-XAI interaction is provided in Table 1.

Despite numerous existing reviews (e.g., Brasse et al., 2023; Laato et al., 2022; Haque et al., 2023), there is no comprehensive overview of experimental XAI studies across AI systems, neither in IS nor HCI research, that considers the underlying mechanisms determining explanations' impacts on user behavior and actual outcomes. Prior reviews have either focused on explanation design (Schwalbe & Finzel, 2023; Chromik & Butz, 2021), emphasized stakeholder goals and objectives (Brasse et al., 2023), or overlooked the intermediate processes linking XAI to the achievement of those goals (Haque et al., 2023; Ferreira & Monteiro, 2020). In addition, dominant perspectives either treat XAI as a system-level construct (Haque et al., 2023) or reduce explanations to isolated design features (e.g., type, object, locality) (Chromik & Butz, 2021), lacking integration across levels of analysis. This lack of conceptual integration limits our understanding of how explanations function as part of a causal chain, from explanation input to user cognition and perception to behavior.

To tackle these challenges, our review centers on experimental XAI studies that uncover causal links between explanation features, user perceptions, and real-world outcomes. This focus is motivated by the lack of theory-driven experimental research, especially given the rise of generative and agentic AI, and the unique value of user-centric experiments in isolating the effects of explanations under controlled conditions. By complementing existing conceptual, technical, and qualitative reviews, our approach offers granular insights into the mechanisms of human–XAI interaction and supports a theory-driven consolidation of causal effects.

As shown in Table 1, our review distinguishes itself by: (1) covering both IS and HCI domains, (2) systematically analyzing experimental studies that provide quantifiable

Table 1 Overview of related prior reviews of human-XAI interaction

Author(s)	Period of review	Sample size	Scope	Review focus
Haque et al. (2023)	Up to 2022	58	Various domains	Analyzing XAI representations, effects, explored domains, and future research avenues, including user-centric empirical studies with a focus on the overall XAI system.
Laato et al. (2022)	2008–2020	25	HCI	Emphasizing the communication for users with high-level goals, such as understandability or trustworthiness and deriving design recommendations for user-centric XAI.
Ferreira and Monteiro (2020)	Up to 2020	52	HCI	Deriving a taxonomy of explainability and outlining the main stakeholders and reasons and motivations for providing explanations.
Schwalbe and Finzel (2023)	2010–2021	70	Various domains	The paper provides a comprehensive overview of crucial aspects or properties of XAI methods and assists in locating detailed surveys on the subject.
Rong et al. (2023)	2019–2023	97	HCI and more AI	The paper categorizes human-based XAI evaluations by trust and usability, highlights XAI's varied adoption across fields, and offers practical guidelines for designing XAI user studies.
Chromik and Butz (2021)	Up to 2020	91	HCI	Reviews research on user explanation interfaces for ML systems, highlighting key concepts and design principles to guide human-centric interface creation for varied audiences.
Brasse et al. (2023)	Up to 2023	180	IS and electronic markets	This review offers insights into the most receptive publications, tracks the evolution of academic discourse, and identifies key concepts (e.g., explanation goals, explanation scope) and methodologies of XAI with a focus on the recipients of the explanations (e.g., lay users, developers, etc.)
Langer et al. (2021)	Up to 2021	>100	Various domains	Presenting five primary classes of stakeholders who demand XAI and revealing their requirements.
This review	*2000–2025*	*107*	*HCI and IS*	*Unlike previous reviews, we focus on experimental studies that reveal the effects of different explanation configurations. This review contributes to user-centered XAI research by developing a conceptual framework and outlining future research directions for HCI and IS.*

evidence on the effects of explanations, (3) examining the full range of sub-processes and mechanisms in human–XAI interaction, and (4) introducing a conceptual framework based on higher-order theoretical constructs to explain how explanation features shape user perceptions, decisions, and outcomes. In doing so, we offer the first comprehensive, theory-informed synthesis of experimental research on XAI, paving the way for more responsible and effective AI use.

3 Theoretical Background

To address the somewhat fragmented and often atheoretical nature of prior human–XAI studies, we draw on the S-O-R-C model (Goldfried & Sprafkin, 1976; Mehrabian & Russell, 1974). This model allows us to trace the full causal chain: from the stimuli to internal organism states, their responses, and finally to downstream consequences. This perspective is particularly useful for unpacking the psychological mechanisms through which explanation features influence user behavior in AI-supported decision-making. Broader reviews in the field often remain descriptive and lack the theoretical grounding necessary to uncover the underlying mechanisms behind user responses (Brasse et al., 2023; Laato et al., 2022). Our use of S-O-R-C offers a response to this problem by providing a systematic human–XAI interaction framework to connect explanation design with user perceptions, reactions, and outcomes.

The stimulus-organism-response framework (S-O-R), first presented by Mehrabian and Russell (1974), states that external factors (stimuli) can influence an individual's perceptions and assessments (organism), which result in certain behavioral intentions (response). These factors shape the behavioral outcome of an event (Arora, 1982). Stimuli are defined as the causal factors that affect the mental state of individuals. The response refers to the final behavioral state and reaction that has been caused or at least been influenced by the stimulus. While stimuli and response are observable elements of the S-O-R model, the organism is referred to as the unobservable phenomenon. It comprises the cognitive and affective processes that are triggered and performed during decision-making. Prior human–XAI research has adapted the theory to explain and understand the drivers of the adoption of XAI-based decision support systems (Dalvi-Esfahani et al., 2023) as well as the deceptive nature of XAI (Schneider et al., 2023). We argue that this theoretical foundation is suitable for analyzing human–XAI comprehensively because explanations are designed to purposefully impact the user's perceptions of AI and result in desirable responses (Langer et al., 2021; Brasse et al., 2023). The S-O-R model was, however, further extended by the *consequences* that the behavioral intentions imply. Thus, the S-O-R-C model emerged (Goldfried & Sprafkin, 1976). Consequences represent the real-world effects of XAI, such as changes in work efficiency, decision quality, or fulfillment of user needs.

In contrast to widely used frameworks and models such as the Technology Acceptance Model (TAM) or the Theory of Planned Behavior (TPB), which primarily emphasize system-level adoption and intention (Haque et al., 2023), the S-O-R-C model allows for a more granular and process-oriented analysis of human–XAI interaction. While TAM and TPB often reduce mediators to perceived usefulness and ease of use, typically culminating in satisfaction or intention, the S-O-R-C framework captures a full causal chain, including intermediate perceptual mechanisms and downstream behavioral outcomes. This is especially important in the context of XAI, where individual features like explanations function as distinct stimuli that evoke specific cognitive and affective responses (organism), which in turn drive behavioral reactions (response) and influence task-related or real-world consequences. Unlike traditional acceptance models, which assume monolithic system evaluations, S-O-R-C allows for analyzing the impact of specific explanation types, formats, or modalities (Xiao & Benbasat, 2011). This makes the framework particularly well-suited for understanding emerging AI systems, where user interaction is modular, dynamic, and often emotionally or cognitively demanding (He et al., 2025). By disentangling stimuli (e.g., explanation features) from organism-level processes (e.g., trust, understanding), S-O-R-C provides the analytical flexibility needed to capture the fragmented, adaptive nature of human–AI interaction in these evolving contexts. Finally, understanding the real-world outcomes of XAI is essential for HCI and IS research, given the close alignment of these systems with practical applications and the pursuit of measurable, real-world impact (Dennehy et al., 2023; Abedin et al., 2022). By using S-O-R-C, we provide a theoretically grounded lens to unpack the complexity of human–XAI interaction and advance both explanation design and evaluation in IS and HCI research.

Our choice aligns with recent calls for higher-order theory building in IS (Paré et al., 2015), especially in domains where technological artifacts influence individual cognition and behavior in dynamic and modular ways. As such, the S-O-R-C framework serves as a conceptual lens to systematically structure and synthesize experimental findings, expose underexplored relationships, and guide future research at the intersection of explanation design, human factors, and system performance.

4 Research Approach

We performed a systematic literature review approach following the procedures of vom Brocke et al. (2009) and Webster and Watson (2002). A systematic literature review provides us with a trustworthy and rigorous methodology (Keele, 2007) to evaluate and interpret available research relevant to our particular research questions.

4.1 Source Selection

At first, we defined the objective and the scope of our research and classified our review according to Cooper (1998). We aim for exhaustive coverage to provide generalizable insights into empirical evidence of explanations in human–XAI interactions. The target group is specialized scholars in the field of IS and HCI who develop, design, evaluate, and examine XAI systems. To answer our research questions, we focus on research outcomes of experimental studies, while simultaneously considering the applied theories that motivate the research. As a second step of our SLR, we prepared keywords and keyword combinations as well as inclusion and exclusion criteria for our database search process.

This process included selecting relevant outlets and databases as well as defining the search string. To ensure quality and relevance, we followed an outlet-driven search strategy, identifying the top 20 IS and HCI journals and conference proceedings based on established rankings (e.g., AIS Senior Scholars' Basket) and prior reviews (Elshan et al., 2022; Rzepka & Berger, 2018). Importantly, we first defined the outlets of interest and then selected the appropriate databases (e.g., EBSCO, ProQuest, AISel, ACM DL, Scopus) that index these sources. Databases not covering our targeted outlets (e.g., IEEE Xplore, Web of Science) were excluded to maintain a focused and consistent search scope aligned with our selected publication outlets. To ensure a comprehensive selection of representative literature covering experimental research perspectives on user interaction with XAI, our search encompassed both journals and conference proceedings. This approach was necessary because journal publication timelines are notably longer than those of conferences, and XAI research is rapidly growing, making recent conference findings, such as those from the Conference on Human Factors in Computing Systems (CHI), highly relevant. Conference proceedings are particularly suited to capturing novel research in this rapidly evolving field. As our analysis of publication timelines reveals (Sect. 5), XAI is a contemporary challenge, necessitating attention to outlets, such as conferences, that operate on shorter publication cycles.

Given our research questions and focus on the effects of explanations in human–XAI interaction, we targeted top journals and conferences in the fields of Information Systems (IS) and Human–Computer Interaction (HCI). Specifically, we selected outlets from the AIS Senior Scholars' Basket and other leading venues known for their established history of publishing experimental or design-oriented empirical research, aligning closely with the methodological scope of our study. This approach is consistent with prior design-related literature reviews on human-AI interaction (Elshan et al., 2022; Rzepka & Berger, 2018; Zierau et al., 2020), which also employed outlet-driven strategies to ensure quality. We cross-validated our outlet list with input from a senior researcher to ensure comprehensiveness and rigor. This selective, theory-informed approach provided a high-quality foundation for systematically identifying relevant studies across the chosen journals and conferences.

4.2 Search Strategy and Results

The following search string was customized and streamlined to match the specific features of each database, incorporating wildcard characters wherever feasible: (Artificial Intelligence OR AI OR Algorithm) AND (Explainab* OR XAI OR Explanation). The query-based search resulted in 853 hits in total (see Fig. 1). Next, inclusion and exclusion criteria were defined. For the period, we applied 2000–2025 as a hard-coded filter to also consider foundational work from the 2000 s as shown by prior surveys (Brasse et al., 2023). Furthermore, the literature must be peer-reviewed. We excluded articles that solely addressed the technical aspects of XAI features, such as highly technical articles that prioritize algorithmic implementation over aspects related to the effects of XAI. The inclusion and exclusion criteria can be found in Table 2. We included XAI research on knowledge-based systems and recommender systems, similar to other literature reviews (Haque et al., 2023; Brasse et al., 2023). We performed the search at the end of May 2025. After a subsequent title, keywords, and abstract screening, 155 papers were transferred to the reading stage, after which, 74 papers were finally included based on the presented inclusion and exclusion criteria. A forward and backward search on Google Scholar added 33 papers not found in the initial database search. In total, 107 papers were analyzed qualitatively.

4.3 Paper Analysis

To ensure reliability and consistency, we conducted an iterative coding process with two researchers, independently coding relevant segments. The primary coder developed the initial codes for several articles, and afterward, both coders

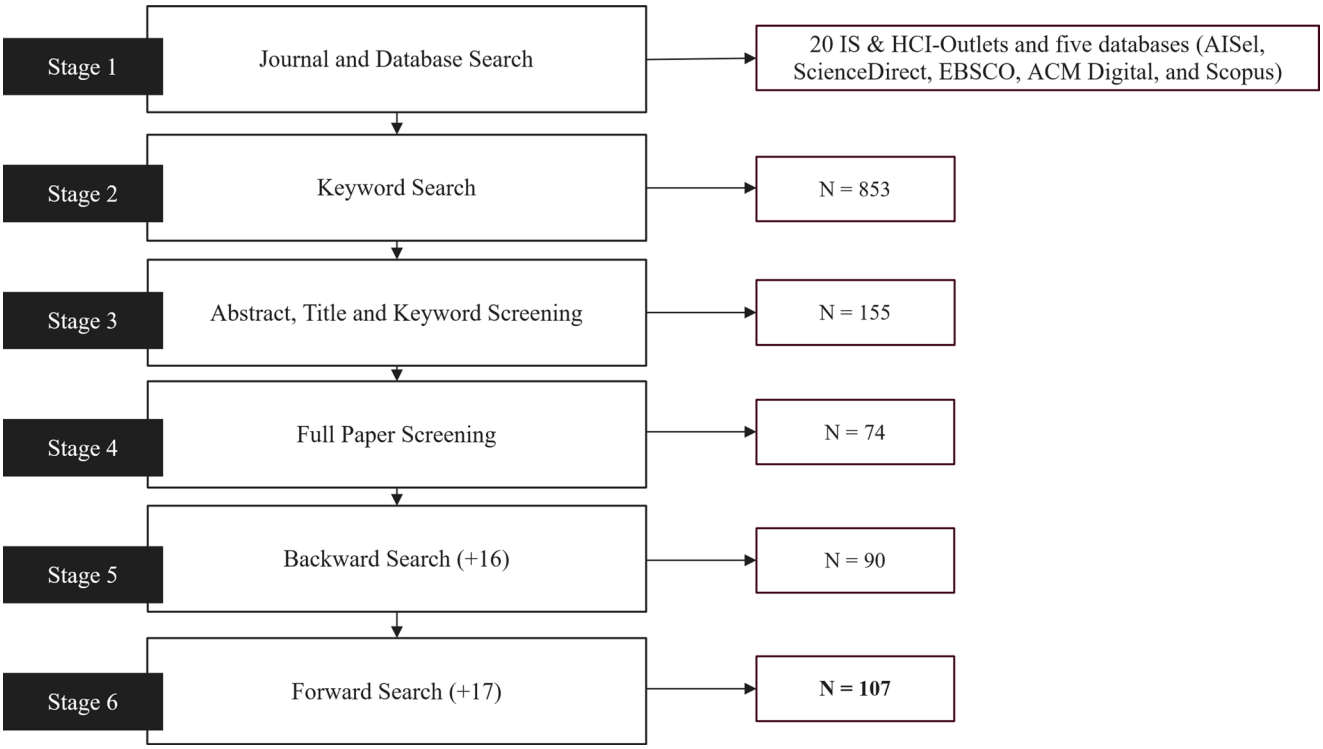

Fig. 1 Literature search process according to vom Brocke et al. (2009)

Table 2 Inclusion and exclusion criteria

Inclusion criteria	Exclusion criteria
1. The research includes an experiment with XAI end users 2. The study examines the effects of explanations on users' behaviors, perceptions, etc. 3. The research specifically addresses the issue from an HCI and IS perspective.	1. Research focusing on technological advancements, explanation methods, and algorithms 2. Empirical studies other than experimental research 3. Editorials, research notes, or other non-peer-reviewed work 4. Studies in languages other than English

examined and reached a consensus on the coding. After a batch of 15 papers, the inductively generated codes and categories were dropped, decomposed, or merged and complemented by new codes (Grodal et al., 2021). The coding process included an iterative approach of analyzing the set of papers randomly and deriving the higher-order constructs based on the underlying codes. The overall coding scheme can be found in Appendix B.

We analyzed the experiments utilizing a qualitative content analysis, including coding dependent, independent, and structural variables such as moderators, mediators, and control variables, and by structuring the results according to the S-O-R-C framework. The framework provided the foundation for our concept matrix (Webster & Watson, 2002) (refer to the Appendix A) and allowed us to study the relationship between external explanation treatments (stimuli), cognitive and affective processes (organisms), and an individual's behaviors (responses) and the actual outcomes (consequences). The lowest level of codes refers to the specific XAI treatments as well as the utilized subjective and objective measurements to either adjust stimuli or to understand users' perceptions and responses, as well as the task-related outcomes, presented in Tables 3, 4 and 5. Within our analysis, we aggregate the measurements into recurring constructs. Each study was also analyzed by domain (e.g., healthcare, finance), system type, experimental design, study type, participant count, source, and demographics. These meta-level elements supported the descriptive analysis.

5 Descriptive Analysis

The field of XAI truly accelerated in 2018, resulting in a significant increase in experimental studies thereafter (refer to Fig. 2). This trend signals a growing awareness among researchers and practitioners regarding transparency and explainability issues. However, it is important to recognize

Table 3 Perception of explanations

Variable	Definition	Measures
Quality	The quality of an explanation reflects how well it meets user needs - whether it's clear, convincing, and useful, and whether it makes users feel confident and informed.	*Set of subjective measures*:perceived value of explanation • satisfaction • persuasiveness • effectiveness • efficiency • preferences
Understanding of Explanations	Understanding explanations is about how easily a user can comprehend an explanation. It's often measured through user surveys or tasks testing comprehension and recall.	*Subjective measures* by surveying understandability and explainability *Objective measures* by performing memory tests
Information sufficiency	Information sufficiency in explanations assesses whether enough relevant and accurate details are provided.	*Set of subjective measures*: information accuracy • information amount • informativeness • novelty of information

Table 4 Perception of the XAI system

Variable	Definition	Measures
Trust	Trust in the context of human–XAI interaction is about how much users feel they can rely on AI systems. It's shaped by the quality and clarity of explanations, and it influences whether users will accept or question AI decisions.	*Set of subjective measures*: credibility • accountability • reliability
Understanding of AI	Understanding of AI measures how well users grasp what an AI system does and why. It's a crucial link to building trust and acceptance, and it includes knowing the system's strengths and limitations.	*Set of subjective self-reported measures*: understandability • explainability • simulatability • interpretability *Objective measures* by conducting knowledge tests or simulation of the model behavior
Utility	Utility refers to how users perceive the overall value and effectiveness of an AI system, encompassing aspects like user experience, ease of use, usefulness, quality of predictions, satisfaction, and usability.	Set of *subjective measures*: user satisfaction • usefulness • ease of use • competence • user experience • prediction quality • predictability • reliability
Transparency	Transparency in AI systems refers to how openly and clearly the system's processes and decisions are presented, influencing user trust.	*Subjectively measured* by perceived transparency, perceived causability, or awareness
Cognitive load	Cognitive load in XAI refers to the mental effort required from users when interacting with XAI systems.	Set of *subjective measures*: perceived information overload • mental effort • information overload • workload *Objective measurement* through eye-tracking and memory performance
Confidence	Human confidence relates to users' confidence in their decision-making when using AI systems as well as the confidence in the AI.	*Subjective measurements* of self-confidence and human confidence in decision-making
Justice and fairness	Justice and fairness in XAI concern how users perceive AI decisions in terms of ethical and equitable treatment, especially in critical areas like finance and law.	*Subjective measures* for fairness and justice

Table 5 Acceptance of AI

Variable	Definition	Measures
Acceptance of AI advice (Reliance)	Acceptance of AI advice measures how users respond to and rely on AI suggestions, focusing on their actual behavior rather than their stated trust or perceptions.	Predominantly *objective measures*: number of accepted advice • weight of advice • weighed discernment for example
Behavioral intention	Behavioral intention reflects users' plans or tendencies to adopt, keep using, or revisit an AI system.	Set of *subjective measures*: intention to use • intention of continued use • intention to return • adoption intention • purchase intention
Error detection	Error detection in XAI is about users identifying inaccuracies in AI advice, influenced by how explanations are presented.	*Objective measurement* by the number of rejected erroneous AI advice.

that the academic roots of this research stream date back to the early 2000s. The first experimental study examining the effects of explanations on the use of knowledge-based systems was published by Mao and Benbasat (2001). Despite this early contribution, the field experienced a prolonged period of limited activity, with only a few notable works exploring explanations, primarily in the context of recommender systems (e.g., Pu & Chen, 2007; Cramer et al., 2008). This nearly two-decade gap was followed by a sharp resurgence of interest starting in 2018, marking the emergence of experimental XAI as a central and dynamic subfield of human–AI interaction research. Most

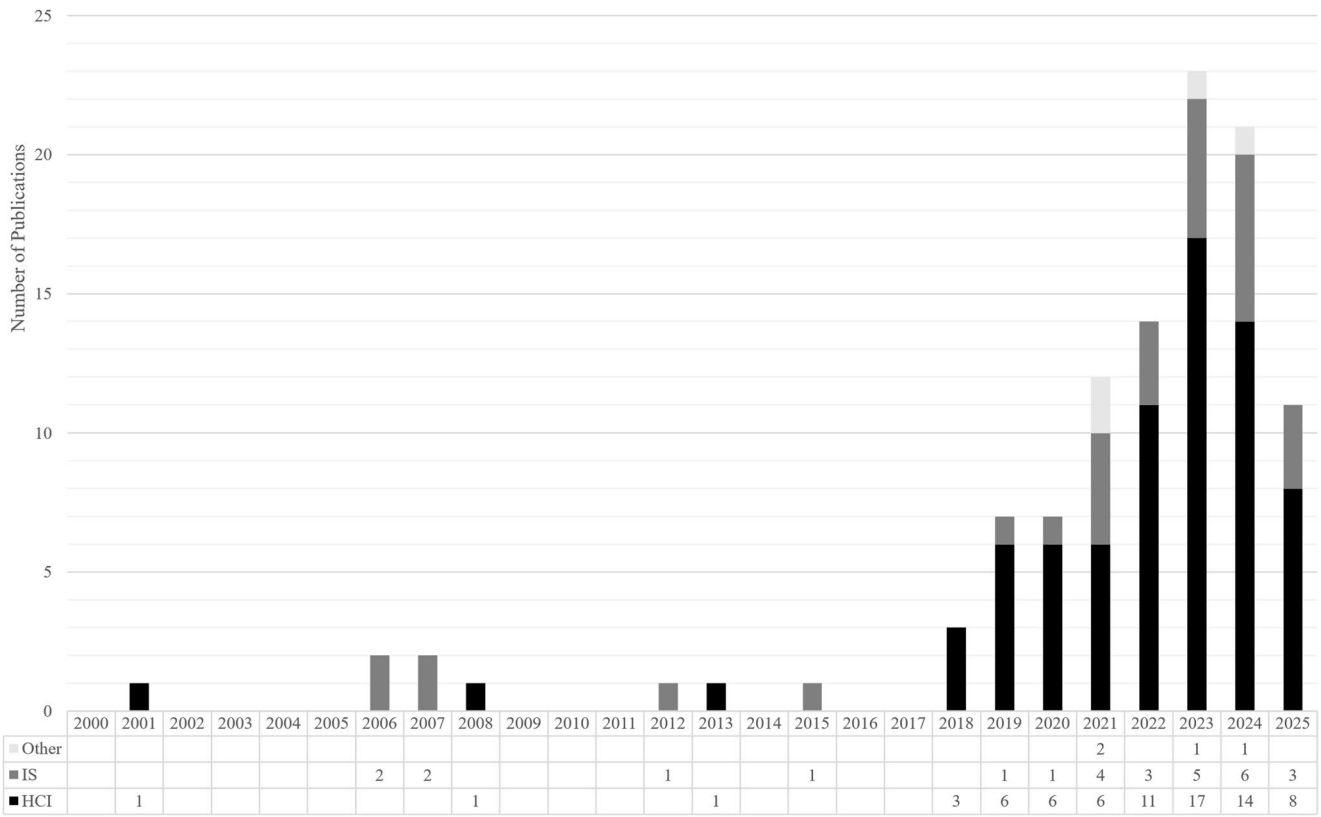

Fig. 2 Distribution of publications according to year

of the studies were published in HCI outlets, highlighting a strong concentration of XAI research in this domain and its close ties to interface design, interaction models, and user experience.

In terms of experimental methodologies, the prevailing approach involved online experiments, with only a minority opting for offline experiments and undertaking field experiments. In this context, XAI researchers demonstrate a consideration of confounding factors and other critical aspects, such as learning effects and algorithm accuracy, within the framework of their controlled experimental setups. Accordingly, a large number of papers do not apply working AI systems, instead opting for vignettes, mockups, or Wizard of Oz systems to control for accuracy and variations among other factors (e.g. Graefe et al., 2023; Matt & Schlusche, 2022; Silva et al., 2023). Only a few papers included working AI systems (e.g., Lee & Chew, 2023; Panigutti et al., 2022). Our sample of literature refers both to between-subjects (e.g., Leichtmann et al., 2023; Westphal et al., 2023; Lima et al., 2023) and within-subjects experiments (e.g., Musto et al., 2021; Naiseh et al., 2023; Panigutti et al., 2022). Although we filtered for experimental user studies, several included papers comprise multiple research methods such as interviews, think-aloud sessions, or quantitative evaluations (e.g., van der Waa et al., 2020).

6 Conceptual Framework of Human–XAI Interaction

In the following sections, we elaborate on the analyzed higher-order theoretical constructs and corresponding aspects of human–XAI research, forming the human–XAI interaction framework based on the S-O-R-C model. An overview of the relevant components of human–XAI interaction can be found within the concept matrix in the Appendix as well as within the following Fig. 3.

To synthesize the accumulated knowledge on the effects of XAI and to accommodate the development of new theoretical foundations, we developed a conceptual framework for human–XAI interaction guided by the theory of S-O-R (Mehrabian & Russell, 1974) (Fig. 4). This allows us to understand the higher-order concepts and constructs as well as the effects and relationships in the field of human-centered XAI. Unlike models such as TAM or TPB, which focus primarily on system-level adoption and intention, the S-O-R-C framework captures both the granular impact of individual system features and the broader influence of the XAI system on users' cognitive and affective responses, behaviors, and resulting outcomes. This makes it particularly suitable for evaluating feature-rich AI systems, as supported by its use in explanation-focused studies (e.g., Xiao & Benbasat,

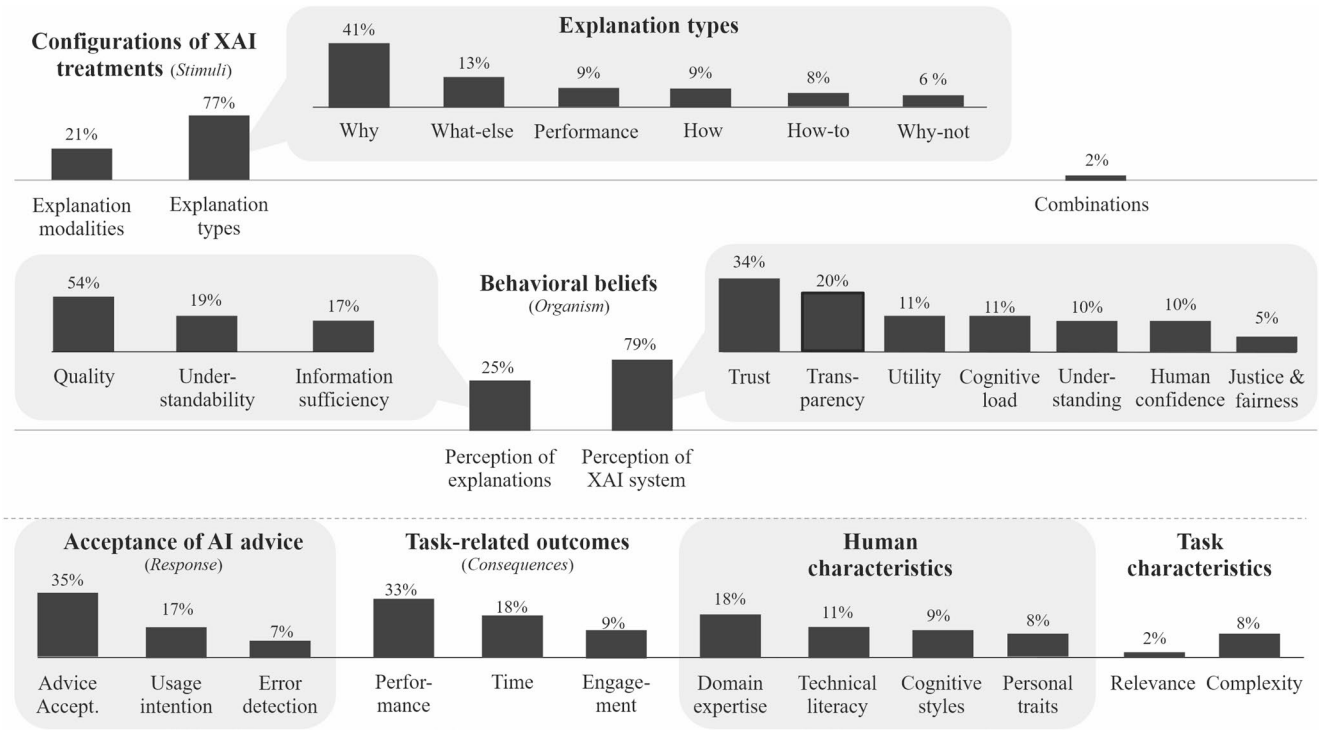

Fig. 3 Characteristics of experimental research on human–XAI interaction

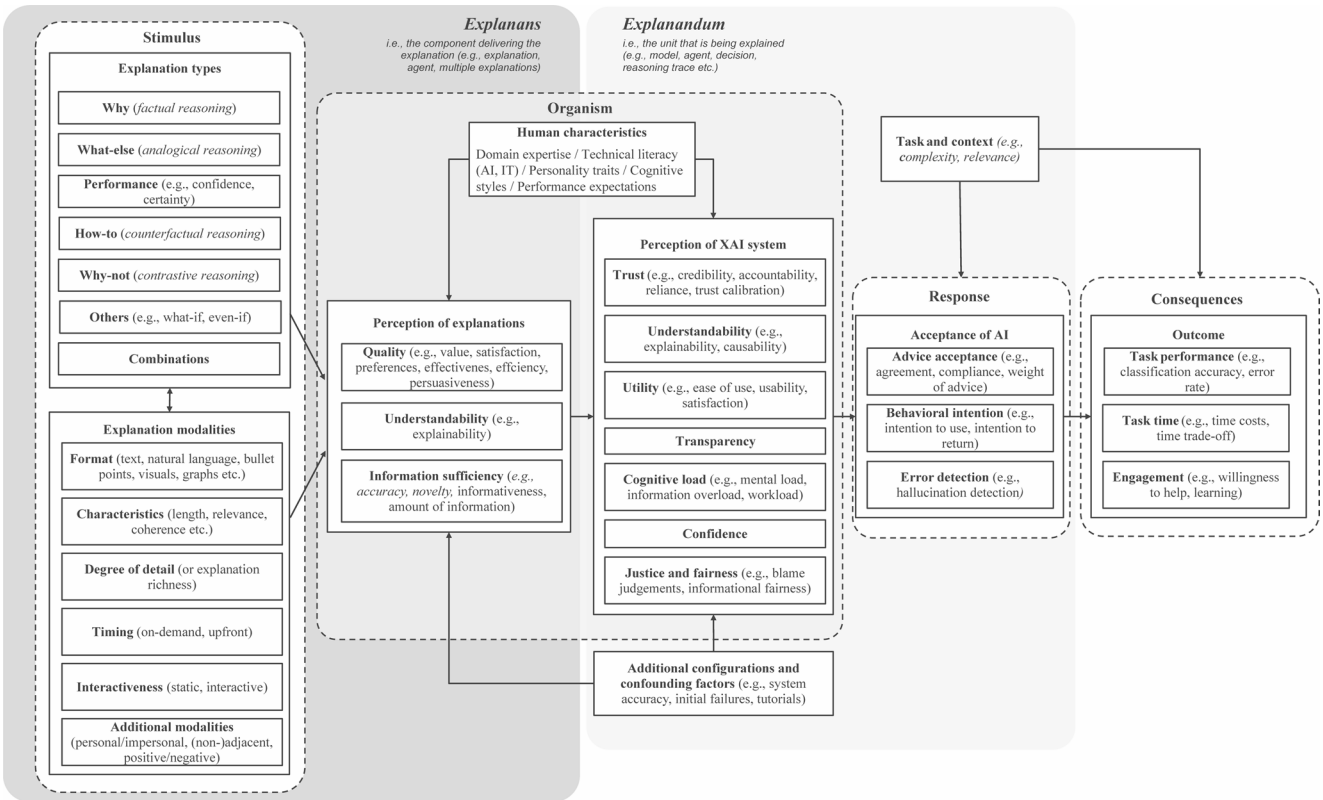

Fig. 4 Synthesized conceptual framework of human–XAI interaction based on the S-O-R-C model

2011; Schneider et al., 2023; Ochmann et al., 2024). We systematize the causal chain of XAI into four main theoretical constructs: (1) stimuli, (2) organism, (3) response, and (4) consequences. We designated XAI configurations as the independent external (1) stimuli in experimental XAI studies (Sect. 6.1). Researchers adjust these configurations to observe the effects of various explanation types, combinations of explanations, and modalities. Next, we divided the core causal chain of human–XAI interaction into three components. The first component, XAI perceptions, covers how users observe, comprehend, and assess explanations, solely focusing on explanations as the primary research subject. In the second part, researchers explore how users perceive the XAI system as a whole, including its explanations. Both perceptions of the XAI configurations and the full XAI system refer to the non-observable (2) organism of our human–XAI interaction framework (Sect. 6.2).

We found that the perception of the explanations can be separated from that of the XAI system, which comprises the AI system and the explanations as additional cues. Thus, we conceptualize explanation types and modalities (i.e., stimuli) as well as users' perceptions of these explanations as constructs that constitute the explanans of the system. Following philosophical tradition (Hempel & Oppenheim, 1948; Miller, 2019), the explanans refers to the entity or mechanism that provides the explanation. Our model acknowledges that multiple explanans (e.g., individual explanations or explanation elements) may contribute to clarifying a single AI output, such as a decision, recommendation, or underlying model functionality. This target of explanation is referred to as the explanandum- the unit that is to be explained. Our human–XAI interaction framework emphasizes that experimental XAI research often focuses either on the explanans (i.e., explanation features) or the explanandum (i.e., the overall system), yet the theoretical and practical impact emerges from their interplay. For instance, GenAI systems can dynamically generate and adapt their reasoning while presenting a chain-of-thought, making the interaction between explanation features and system behavior especially salient in shaping user understanding and trust. The link between explanans and explanandum thus captures the explanatory mechanism by which explanations influence user cognition and behavior. The figure indicates the conceptual conflation within the organism construct.

We then summarize AI acceptance as the observable (3) response, capturing how users interact with the system, either by accepting or rejecting its outputs, and their intention to use the system or its advice, which also constitutes an important component of the explanandum (Sect. 6.3). As the examined studies show, the response is driven by the cognitive and affective processes (organisms). Finally, these configurations, perceptions, and behaviors yield measurable (4) consequences in interactions between humans and AI, such as task performance, time, and engagement (Sect. 6.4). In addition to the S-O-R-C, we found that human characteristics, task specifics, and context significantly moderate the human–XAI interaction (Sect. 6.5). While some configurations, like the accuracy of the AI model, aroused researchers' interest, they do not represent the main subject of interest. In the following, we present each component in detail.

6.1 Stimuli - XAI Configurations

Our analysis highlights the diversity of conceptualizations of explanans as external stimuli in our retrieved sample. Concepts from the literature on knowledge-based systems (Gregor & Benbasat, 1999), such as trace or line of reasoning, justification, control, and terminology, have been explored to enhance human decision-making (Matt & Schlusche, 2022; Aslan et al., 2022). An additional conceptualization lies between local and global explanations, where some studies focus on providing insights into specific model predictions (local), while others emphasize understanding the overall system behavior (global) (Lai et al., 2022; Wastensteiner et al., 2021; Zhang et al., 2020). Results from that stream showed that local explanations are more effective in trusting AI systems to make fair decisions (Dodge et al., 2019; Herm, 2023) and that they have less of a priming effect (Lai et al., 2022). Moreover, the white-box versus black-box dichotomy further contributes to this diversity, with some investigations aiming for transparency by revealing the internal workings of the model (white-box), while others prioritize an opaque approach (black-box) (Poursabzi-Sangdeh et al., 2021; Musto et al., 2021). Research revealed that white-box cannot positively influence the understanding of the model (Cheng et al., 2019) and the ability to identify errors (Poursabzi-Sangdeh et al., 2021) as they might expose complexity. Pafla et al. (2024) compared between human-generated and machine-generated explanations. Overall, the systematization and derivation of explanans in HCI and IS research remain scattered.

6.1.1 Explanation Types

To structure the fragmented field, we map the user-centered categorization of explanations (i.e., why, what-else, why not) (Liao et al., 2020; Wang et al., 2019) with human reasoning types (Miller, 2019) and thereby systematize our insights on explanation types as one categorization of XAI stimuli.

The most often referred to type of explanation explains why (factual reasoning) a certain decision was made. Here literature utilizes feature attribution methods (e.g. Zhang et al., 2023), feature-based approaches (e.g., Yurrita et al.,

2023; Orzikulova et al., 2024), trace explanations (Mao & Benbasat, 2001; Aslan et al., 2022), and other instantiations such as keyword highlighting (e.g. Wang et al., 2022; Lai & Tan, 2019) or natural language based explanations (e.g., Chien et al., 2022). Driven by the rationale that questioning the AI decision with "why" remains in line with humans' factual reasoning, this type has a positive effect on the human–XAI interaction. For example, Meske and Bunde (2022) exemplified a positive impact on informativeness, trustworthiness, and mental models. However, the result of our sampled literature shows that not all experiments were able to reconstruct a significant impact (Lai et al., 2022; Zhang et al., 2021). Rader et al. (2018) highlighted that why-explanations can negatively influence perceived control.

What-else (analogical reasoning) explanations represent the second most frequently discussed type of XAI feature (e.g. Leichtmann et al., 2024; Silva et al., 2023). Analogical reasoning also appears in the literature as example-based, case-based, similarity-based, or evidence-based explanations. What-else explanations could be instantiated by providing alternative solutions (Sivaraman et al., 2023) or showing similar user profiles (Tsai et al., 2021). Tsai et al. (2021) showed that example-based explanations outperformed feature-based explanations. Again, there are possible negative effects related to analogy-based explanations. In terms of justice and fairness perceptions, case-based explanations show a negative impact when presented next to other explanation types (Binns et al., 2018; Yurrita et al., 2023). Similarly, Jiang et al. (2022) showed that in case of high user uncertainty, providing divergent information such as alternatives harms humans' decision making.

Confidence scores (i.e., information on the system's certainty for a specific prediction) can act as a cue for human users to evaluate the AI response (Steyvers et al., 2025). van der Waa et al. (2020) proposed and examined different confidence measures from the perspective of XAI. As such, confidence scores can act as a measure for improving trust calibration (Zhang et al., 2020). However, that type does not provide users with any insights into the system's reasoning and thus does not increase users' understanding of the system (Cramer et al., 2008).

How-to (counterfactual reasoning) explanations provide users with hypothetical examples by presenting the required changes of input factors to get a different prediction (Naiseh et al., 2023; Silva et al., 2023; Weller et al., 2025). For example, Wang et al. (2022) provided a mechanism for substituting words in user inputs to increase users' performance in an AI-assisted ideation task. In addition, Lee and Chew (2023) revealed that counterfactual explanations are effective in reducing both experts' and laypeople's overreliance, and Silva et al. (2023) found that they were significantly rated as being more explainable than the baseline. Nonetheless, Herm (2023) showed that local how-to explanations demand more mental effort compared to prior mentioned what-else explanations. Schutter and Cremer (2024) distinguish between would-, could-, and should-counterfactuals, illustrating the breadth of explanatory types in this reasoning approach.

Why-not (contrastive reasoning) describes why a certain prediction was not made (Wang et al., 2022; Zhang & Lim, 2022; Pafla et al., 2024). In the realm of recommender systems, contrastive explanations can be represented as trade-off explanations that facilitate the comparison of different recommendations (Cramer et al., 2008; Pu & Chen, 2007). Wilkinson et al. (2021) discovered that why explanations were preferred over why-not explanations in terms of justifying some advice. Apart from this, the experimental evidence on contrastive reasoning in human–XAI interactions remains scarce.

How explanations are categorized as global explanations because they explain the overall model reasoning, often referred to as the process (Jang et al., 2024), regardless of the given prediction (Zoeten et al., 2024). Concerning research on knowledge-based systems, how explanations are categorized as strategic explanations (Aslan et al., 2022). How explanations exert the greatest influence in terms of mental effort (Herm, 2023).Lai et al.(2022) explored that global explanations can hence hamper human–XAI performance and Rader et al.(2018) stated that explanations can cause doubts about AI's consistency.

There are multiple additional explanation types worth exploring: what-if explanations (Anderson et al., 2020) offer insights into future or altered inputs, while even-if explanations (Kenny & Keane, 2021) were absent in our sample, but suggest potential for future research. Other types found include terminological (Aslan et al., 2022), deep (Mao & Benbasat, 2001), and what explanations (Rader et al., 2018). Additionally, combinations of explanation types remain undiscovered. Most researchers choose to test single explanations using between-subjects experiments or present all treatments separately in within-subjects experiments. Only Anderson et al. (2020) and Bhattacharya et al. (2024) considered combining explanations to increase the impact on human–XAI interaction. Current human–XAI research offers limited insight into how combined explanations (i.e., multiple explanans) create complementary or synergistic effects. Therefore, we highlight the need to view stimuli not as isolated factors, but as interacting components that jointly shape user behavior.

6.1.2 Explanation Representation Modalities

Beyond the variety of explanation types, XAI research also emphasizes the importance of explanation modalities, including text, numbers, bullet points, visuals, graphs, feature histograms, highlighted inputs, and rule sets (e.g., Lu et al., 2023; Ha & Kim, 2024; Pafla et al., 2024). In the most recent discussions, focus was placed on the significance of natural language explanations (Lima et al., 2023; Lu et al., 2023). Human-understandable summaries of the model's reasoning, as discussed by Chien et al. (2022), are recognized as more effective than graph-based explanations. In addition to the format, Förster et al. (2020) examined the importance of specific characteristics in designing and shaping the external stimuli of XAI systems. The researchers differentiated between length, relevance, coherence, and generality and found that concrete, coherent, and relevant explanations are most appreciated. Musto et al. (2021) found that in movie recommendation systems, effective justifications are long and summarize key information. Another central modality is the degree of detail incorporated within the explanations. For example, the number of explanation elements (Musto et al., 2021; Guesmi et al., 2022) such as visual chunks (Abdul et al., 2020) is often related to a trade-off between cognitive load and accuracy (Herm et al., 2022). Other researchers also refer to explanation richness (Pu & Chen, 2007) or the degree of transparency (Zhang et al., 2021).

The analysis shows that prior work offers diverse explanation formats, providing inspiration for future XAI modality research. The set of retrieved papers examined completeness and soundness (i.e., the extent to which the complete system is addressed and the extent to which each part of an explanation describes the AI system) (Kulesza et al., 2013), presentation order, and morphological clarity (i.e., the extent to which a visualization delineates features by adjusting their appearance or discarding specific data) (Hudon et al., 2021), personal vs. impersonal (i.e., varying the source of the explanation between human and non-human sources) (Kunkel et al., 2019), adjacent vs. non-adjacent (i.e., showing the explanation integrated in the underlying input vs. displaying it separated from the input data) (Hudon et al., 2021), and positive vs. negative framing (i.e., considering that the AI decisions are inherently positive or negative for a user) (Hadash et al., 2022). Regarding the interaction modalities in XAI systems, researchers distinguish between interactive and static explanation interfaces (e.g., Martijn et al., 2022). Naiseh et al. (2023) argue that users perceive XAI systems as a new interactive system that needs to be customizable and adjustable. Closely related to that topic is the type of explanation provision. The explanation can be provided on-demand or upfront (Tsai & Brusilovsky, 2021). Finally, XAI research has shown that personalized explanations are perceived as more helpful (Schröppel & Förster, 2024). Overall, the multi-modality and interactivity of state-of-the-art GenAI models further complicate this trend, prompting a view of explanans as a multi-categorical construct - one that poses significant difficulties for focused experimental research.

6.2 Organism - Human Perceptions of XAI

Experimental research highlights two key effects of explanations in human–XAI interaction related to the organisms component in the S-O-R-C model: (1) perception of explanations and (2) XAI system perception. Perceived explanations shape system perception, which in turn influences user responses. The organism component serves as a bridge between the explanans and the explanandum. While users' perceptions of the provided explanations reflect the explanans, their perceptions of the broader XAI system, such as the AI's decisions, reasoning processes, or the model itself, correspond to the explanandum.

6.2.1 Perception of Explanations

The perceptions of XAI configurations, measured through subjective metrics like explanation quality, understandability, and information sufficiency (Table 3), directly relate to the explanans provided in AI experiments. Assessing these perceptions helps evaluate the effectiveness of XAI features and treatments independently of the AI system's broader factors.Quality. The perceived quality of explanations has been discussed by several researchers by considering the value of an explanation (Graefe et al., 2023), the satisfaction (Lu et al., 2023; Kaufman et al., 2025), the preferences (e.g., Ashktorab et al., 2019), perceived usefulness (Zhang et al., 2023), effectiveness and efficiency (e.g. Wilkinson et al., 2021; Musto et al., 2021). For instance, Dikmen and Burns (2022) measured explanation quality by surveying confidence, human likeness, adequate justification, and understanding. Within the field of recommender systems, the concept of persuasiveness is frequently mentioned.

Understanding of explanations. XAI research emphasizes understandability as central to conveying complex system behavior and a fundamental criterion for explanans. Beyond perceived understandability - commonly assessed in surveys (e.g., Hadash et al., 2022; Zhang et al., 2023) - some studies employ objective measures, such as memorization tasks (Wastensteiner et al., 2021) and comprehension tests (Abdul et al., 2020). Authors also refer to related notions like explainability (Lima et al., 2023), which contributes to some conceptual variation and lack of consistency across the XAI literature.

Information sufficiency. A multitude of authors studied the perceived degree of information that is transferred by the explanations in addition to the mere prediction. Therefore, we refer to information sufficiency by summarizing multiple variables, such as information accuracy (Lu et al., 2023), the novelty of information (Lu et al., 2023; Rader et al., 2018), perceived informativeness (Meske & Bunde, 2022)and the amount of information (Dikmen & Burns, 2022).

6.2.2 Perception of the XAI System

Experimental XAI research also investigates users' overall perceptions of the AI system (Table 4) as shaped by the explanans. It connects the explanans to the actual explanandum, which is typically related to explanation goals (Laato et al., 2022).

Trust Trust is one of the most cited outcomes in human–XAI research, with many studies showing that explanations aim to - and often do - increase user trust (e.g., Lee et al., 2019; Silva et al., 2023; Tutul et al., 2024). While others demonstrated that there is no direct but an indirect effect through justification quality and perceived transparency (Wilkinson et al., 2021). Trust is positively correlated with perceived understanding (Cramer et al., 2008) and transparency (Jussupow et al., 2021) and depends on different explanation quality measures such as confidence and understandability (Dikmen & Burns, 2022; Kunkel et al., 2019). Multiple authors investigated related concepts of trust such as credibility (Chien et al., 2022), accountability (Shin, 2021), and reliability (e.g. Naiseh et al., 2023; He et al., 2023). The latter stream aims at achieving appropriate trust (He et al., 2023). Overall, these sub-variables fall under trust as the umbrella term. The core mechanism, referred to as trust calibration (Zhang et al., 2023; Naiseh et al., 2023), involves users adjusting their trust in system advice over time to achieve appropriate trust (e.g., Yang et al., 2020). In contrast to appropriate trust, researchers discovered either overtrust or under-trust. Leichtmann et al. (2024) and Naiseh et al. (2023) showed that without explanations users tend to overly on AI recommendations. Despite the positive findings, a few researchers could not find any effect (e.g., Westphal et al., 2023; Leichtmann et al., 2023) or even examine a negative effect of explanations on trust (e.g., Zhang et al., 2021; Papenmeier et al., 2022). Especially, in cases of low and medium accuracy, explanations do not add any value or even harm the user's trust intention.

Understanding of AI The dependent variable understanding of AI describes to which degree the users comprehend the AI system's actions. Researchers either referred to understandability (e.g., Cramer et al., 2008; Leichtmann et al., 2024), explainability (Lima et al., 2023), interpretability (Frost et al., 2025), or causability (Aslan et al., 2022). Furthermore, research exhibits simulatability as a more specific form of understanding an AI model's behavior (Wang & Yin, 2021; Silva et al., 2023). Apart from performance metrics, explanations can help users to better understand the capabilities and limits of AI (Leichtmann et al., 2024). Perceived understanding and explainability act as a mediator for trust, competence, and acceptance (Cramer et al., 2008; Silva et al., 2023).

Utility Multiple authors have studied how explanations influence the users' evaluation of AI systems (Ebermann et al., 2023). We summarize utility as an overarching category including perceived competence (e.g., Naiseh et al., 2023), user experience (Lai et al., 2022), ease of use (Matt & Schlusche, 2022; Meske & Bunde, 2022), perceived usefulness (e.g., Sivaraman et al., 2023; Meske & Bunde, 2022), prediction quality (Tsai et al., 2021), satisfaction (Guesmi et al., 2022), and usability (Lee & Chew, 2023).

Transparency Perceived transparency associated with XAI is established in research (Lu et al., 2023; Shin, 2021; Campos et al., 2024). Notably, the existing literature consistently demonstrates the positive influence of explanations on transparency (Tsai & Brusilovsky, 2021). Aslan et al. (2022) contribute to this understanding by establishing a link between perceived transparency, explanations, and causability. Furthermore, the literature underscores the consequential association between perceived transparency and trust, as emphasized by Shin (2021) and Brunk et al. (2019).

Cognitive Load Cognitive load - users' mental effort when interacting with XAI - is a key concern, with the goal of minimizing overload. Our analysis reveals trade-offs between cognitive effort and explanation accuracy (Abdul et al., 2020). Increasing the information density to increase transparency comes with the cost of perceived information overload (Chien et al., 2022). Again, within the selected literature, authors consider different concepts of cognitive load, e.g., mental effort (Wastensteiner et al., 2021; Herm, 2023), information overload (Rader et al., 2018; Poursabzi-Sangdeh et al., 2021), and workload (Lai et al., 2022).

Confidence A less studied but growing field involves the user's confidence in their decision-making (e.g., Lee & Ram, 2023; Kim et al., 2024). Lai et al. (2022) examined the human's confidence in creating delegation rules but could not indicate any improvement by global or local explanations. Similar findings are provided by Panigutti et al. (2022), who emphasized that explanations did not

significantly increase or decrease human confidence. Yang et al. (2020) revealed that non-experts had more confidence in their decisions.

Fairness and Justice The constructs of fairness and justice are central to the finance and legal domain, where humans are more strongly influenced by AI. In a within-subjects study, Binns et al. (2018) demonstrated the effectiveness of explanation styles on the perception of fairness and justice. Lima et al. (2023) examined the effect of explainability on blame judgments, where they concluded that explanations alone do not significantly influence the perceptions of whom to blame. Similar to transparency, Shin (2021) included fairness as a mediator of trust in their conceptual model. Yurrita et al. (2023) demonstrated a positive effect of explanations on perceptions of informational fairness.

6.3 Response – Acceptance of AI

Finally, perceptions of explanations and the AI system influence users' responses to XAI stimuli. Acceptance of AI advice and error detection are direct behavioral outcomes, while usage intentions (e.g., intention to use) still reflect subjective behavioral beliefs and intentions. According to our analysis, responses to the XAI system and its advice are part of the explanandum because these reactions directly relate to the unit being explained. In this sense, the explanandum—that is, what the explanations aim to clarify—is ultimately reflected in how users act upon, revise, or adopt AI recommendations. These responses, in turn, serve as measurable indicators of the extent to which the explanandum is understood, trusted, or considered appropriate for action.

Acceptance of AI Advice (Reliance) The degree of acceptance of AI advice is crucial, as it refers to the degree to which humans accept or reject the advice produced by the AI system as well as changing their decisions (Aslan, 2024). In contrast to trust, the literature emphasizes the objectively measurable actual agreement (e.g., Naiseh et al., 2023; Zhang et al., 2020) and compliance (e.g., Westphal et al., 2023) of users with AI. The majority of experiments demonstrated the effectiveness of explanations on the acceptance of AI (Sivaraman et al., 2023; Jiang et al., 2022). However, Wilkinson et al. (2021) reported that there are no direct but rather indirect effects mediated by the quality of justification styles, hence arguing for considering the perceptions of explanations as outlined in the previous chapter. Sivaraman et al. (2023) emphasized that explanations meaningfully complement AI advice. In contrast, Westphal et al. (2023) and Silva et al. (2023) could not point out a significant trend of compliance in each XAI condition. Our analysis showed that the approaches to objectively measure AI advice acceptance are more sophisticated. For example, the weight of advice (WOA) has been applied to examine how AI advice influences humans' decision-making, by measuring the changes in human decisions before and after showing AI advice (Panigutti et al., 2022; Poursabzi-Sangdeh et al., 2021). Additionally, Naiseh et al. (2023) measured if users switched from the AI recommendation, and Danry et al. (2023) observed weighted discernment. Finally, Lee and Chew (2023) measured acceptance in detail by differentiating between (dis-) agreement of right AI advice and (dis-) agreement of wrong AI advice. Research on acceptance of AI advice relates to the rapidly growing field of ***AI reliance*** and the quest for understanding appropriate reliance on AI advice (Bo et al., 2025; Kim et al., 2025).

Behavioral Intention Additionally, a comparable number of publications examined the behavioral intentions to use an AI system (e.g., Leichtmann et al., 2024; Guesmi et al., 2022). The rationale behind that is the goal of strategically improving the usage of AI systems through explainability. Aslan et al. (2022) showed that a higher level of causability increases the intention to utilize the AI system. Similar findings were presented by Ebermann et al. (2023). However, Panigutti et al. (2022) could not confirm these findings.

Error Detection A few researchers studied the effectiveness of XAI in terms of recognizing incorrect AI advice (e.g., Wang & Yin, 2021; Naiseh et al., 2023). For example, Lee and Chew (2023) found that counterfactuals are suitable for identifying wrong AI advice. However, Rader et al. (2018) reported that explanations were less effective in verifying the correctness of the system's output. Poursabzi-Sangdeh et al. (2021) show that high information load hinders users from spotting model errors, even if they can simulate its behavior. Despite its link to AI acceptance, error detection is often overlooked or treated separately.

6.4 Consequences - Task-Related Outcomes

Beyond observable responses, experiments often measure the real-world consequences of human–XAI interaction. Key consequences include task performance (accuracy and quality), task time, and user engagement with the system.

Task Performance Overall, a small portion of authors measured the impact of explanations on outcome-related variables such as task performance (e.g., Dikmen & Burns, 2022; Spitzer et al., 2024). In general, task performance is referred to as the number of correctly predicted instances, often named human classification performance (e.g., Lai et al., 2020; Lu & Zhang, 2025). Still, task performance can come in different facets depending on the domain and task.

For instance, Bauer et al. (2023) and Dikmen and Burns (2022) examined investment performance, Herm (2023) investigated diagnostic accuracy, and Bhattacharya et al. (2024) studied the impact on model's performance when expert users configure AI systems. The results are somewhat contrary. While one group of researchers highlighted the positive impact of explanations due to the better understanding and the appropriate reliance (e.g., Wastensteiner et al., 2021; Silva et al., 2023), another group found that explanations can also decrease or at least can not improve overall task performance because of information overload, confirmation biases, and priming effects (e.g., Bauer et al., 2023; Zhang & Lim, 2022). Hamm et al. (2023) define performance as a construct composed of quality and time, where quality is considered equivalent to the correct classification. Other studies referred to specific outcomes of XAI use such as cyber resilience (Sadeghi et al., 2024), error measures in forecasting setups (Weller et al., 2025) or assessing cancer risks (Pálfi et al., 2024).

Task Time This variable comprises the time to complete a particular task or to respond (Müller et al., 2025) and refers to the efficiency of solving a task (Wilkinson et al., 2021; Spitzer et al., 2024). One cluster of researchers refers to time costs to draw attention to the fact that interactive XAI interfaces will increase the time users try to discover the system (Cheng et al., 2019; Tan et al., 2012). As such, there is a trade-off between time to read and understand explanations and task time. However, according to Zhang and Lim (2022), explanation can improve task performance without degrading completion time. Herm (2023) found that users take the longest to complete tasks without explanations, while why and why-not explanations are the most time-consuming. However, time varies based on specific XAI configurations.

Engagement Finally, we summarize engagement as the degree to which the users interact with the system (e.g., Lai et al., 2022). This includes the willingness to help the systems (i.e., by contributing feedback in a sense of hybrid intelligence) (e.g., Lee et al., 2019), the extent to which users discover new knowledge (Musto et al., 2021), and how explanations facilitate learning (e.g., Tsai et al., 2021; Spitzer et al., 2024). Weller et al. (2025) found that factual explanations are associated with more beneficial overruling of the AI by users.

6.5 Human Characteristics and Task Context

Human agent characteristics are less prominently examined within our retrieved set of papers. The most prominent factor refers to domain expertise (e.g., Bauer et al., 2023; Yurrita et al., 2023; Lu & Zhang, 2025; Arnold et al., 2006). Prior knowledge in the field of observation has crucial implications for the demand for explanations (Mao & Benbasat, 2001), the effectiveness of explanations (Wang & Yin, 2021), and appropriate trust (Lee & Chew, 2023). For example, Wang and Yin (2021) argue that domain knowledge is required to understand the explanations. Lee and Chew (2023) found that laypersons tend to over-rely on AI advice.

Another relevant human characteristic can be summarized as technical literacy which comprises prior experience in AI (e.g., Cheng et al., 2019; Ashktorab et al., 2019; Riefle et al., 2024), IT knowledge (e.g., Aslan et al., 2022), and AI knowledge in particular (e.g., Zhang et al., 2023; Leichtmann et al., 2024). Furthermore, a portion of publications considered cognitive styles (e.g., Naveed et al., 2018; Spitzer et al., 2024; Giboney et al., 2015). In this regard, research has focused on different constructs such as decision-making styles (Naveed et al., 2018) or the need for cognition (NFC) (e.g., Martijn et al., 2022). People with higher cognitive reflection tend to better discern and seek additional sources (Danry et al., 2023). However, Riefle et al. (2024) utilized the constructs of rational and intuitive cognitive styles proposed by Hamilton et al. (2016) and found that an increase in rational cognitive style significantly decreases the objective comprehension of AI model behavior. In addition, several other variables were considered including personality traits (e.g., Martijn et al., 2022; Shulner-Tal et al., 2024; Nimmo et al., 2024; Okoso et al., 2025) or performance expectations (Matt & Schlusche, 2022; Bauer et al., 2023; Zoeten et al., 2024). Human characteristics shape key internal responses (organism) in human–XAI interaction, especially perceptions of explanations and the system, making them vital for XAI design and evaluation. This line of research can be further extended to the context of group decision-making. For instance, Brito Duarte et al. (2025) compared the use of explanations in individual versus group settings and found that the absence of explanations tends to amplify overreliance on AI-generated outputs.

In comparison to human characteristics, the role of task context is rather neglected within human-XAI research. Still, task complexity was incorporated as a mediator in prior research (e.g., Westphal et al., 2023). Dikmen and Burns (2022) measured perceived task difficulty and Ebermann et al. (2023) controlled for task conditions by selecting distinct levels of uncertainty. Finally, task relevance was considered in research where AI advice can have a crucial consequence (e.g., Leichtmann et al., 2024; Yurrita et al., 2023). However, our analysis reveals that there are various contextual factors influencing the interaction with XAI systems (Jenkin et al., 2024).

7 Discussion

By conducting a theoretical review (Paré et al., 2015) of experimental studies on human–XAI interaction, we conceptualized the state of experimental research and identified key knowledge gaps that inform future directions for the field. Grounded in the S-O-R-C model, we introduced a conceptual framework that provides a causal, process-oriented lens to trace how explanation types and modalities (stimuli) influence user perceptions (organism), subsequent behaviors (responses), and tangible outcomes (consequences) (RQ1). One of the core contributions is the conceptual link between explanation-level perceptions and perceptions of the overarching XAI system, which we characterize as explanans (stimuli and organism) and explanandum (organism and response), enabling clearer theorization and measurement. Our analysis shows that XAI effects vary by explanation design, system type, and task context, revealing inconsistencies and a lack of structured theory use. These findings inform five promising research directions that help shape the future of XAI (RQ2), particularly in the context of emerging AI entities (e.g., LLMs, AI agents) and evolving human-AI delegation mechanisms.

7.1 Main Findings

Through our analysis of the applied stimuli (1), we found an inconsistent categorization of explanation types across the sample of papers. This inconsistency primarily stems from the numerous features and instantiations applicable to explaining decision-making processes. Our analysis highlights the strong relationship between explanation types and modalities, with the chosen modality significantly influencing the effectiveness of the explanations. However, there is a small portion of researchers that considered both explanation types and modalities (e.g., Pafla et al., 2024; Wang et al., 2022) and prior research lacks a systematization of dependencies and combination effects concerning explanation types and modalities. As such, our analysis also raises questions about combining multiple explanations, as most experimental research focuses on single features or comparing treatments. Few studies to date (Anderson et al., 2020; Bhattacharya et al., 2024; Martijn et al., 2022) have investigated the combination of explanation types and discusses the synergies or conflicts, which become more relevant when humans delegate to multiple AI agents. Furthermore, the absence of widely used theoretical frameworks in XAI hinders researchers from building a robust theoretical basis. Thus far, the most prominent theoretical foundations for explanations in XAI remain the works of Liao et al. (2020) and Miller (2019), which draw from social science perspectives and human reasoning frameworks. However, we identified several useful theories. Toulmin's argumentation model, for instance, offers a lens to conceptualize explanations as structured arguments, clarifying their logical coherence and persuasiveness (Arnold et al., 2006; Lee & Ram, 2023; Naveed et al., 2018). Likewise, signaling theory helps interpret explanation elements as cues about the system's credibility and intent, shaping trust and acceptance (Aslan et al., 2022). Cognitive fit theory adds to this perspective by explaining performance variations as a function of alignment between explanation format and task demands (e.g., Ebermann et al., 2023; Giboney et al., 2015; Herm, 2023). These theories help clarify fragmented findings in the literature, provide a structured basis for interpreting and designing XAI stimuli in experimental and real-world settings, and explain how such stimuli shape user perceptions - represented as the organism in our human–XAI interaction framework.

Regarding the organisms (2) component, our findings indicate a consensus among scholars regarding the mediating effect of explanation quality on the subsequent perception of the overall AI system and its predictions (Kunkel et al., 2019). The underlying rationale is that explanations must be valuable to users (Graefe et al., 2023). Measurements vary from persuasiveness (Musto et al., 2021) and satisfaction (Lu et al., 2023) to justification efficiency (Wilkinson et al., 2021). For example, in research on recommender systems (e.g., Tsai & Brusilovsky, 2021; Naveed et al., 2018), the explanations themselves are most important for increasing the advice acceptance rate or the purchase rate. Hence, researchers in that field investigate the persuasiveness of explanations. Our findings related to the organism offer insights into the connection between explanation features (i.e., explanans) and the broader system being explained (i.e., explanandum), thereby shedding light on how explanations shape users' perceptions and behaviors. This new perspective contrasts with prior findings by Haque et al. (2023) and suggests investigating explanans more granularly without neglecting their interrelatedness with the underlying AI model or AI Agent – the explanandum. Theoretical frameworks such as dual-processing theory and the heuristic–systematic model help explain how users shift between intuitive and analytical modes of processing, affecting how explanation features are cognitively evaluated and how this, in turn, informs the perception of the AI system (Ha & Kim, 2024; Lu & Zhang, 2025; Naiseh et al., 2023). Moreover, mental models theory helps explain how users form internal representations of both the explanation and the system's behavior (Bauer et al., 2023; Kulesza et al., 2013; Martijn et al., 2022).

Additionally, we found that human–XAI research focuses on the impact of explanations on appropriate reliance on AI advice as users' responses (3). The research aligns with

examining explainability and transparency as key determinants of appropriate reliance in AI (e.g., Shin, 2021; Jussupow et al., 2021; Silva et al., 2023). This relates to the perceptions of XAI, establishing the main causal chain: explanation quality influences explainability (i.e., understanding the AI's decision rationales and predictions) and transparency, which in turn impacts behavioral trust. This structure motivates a more sophisticated analysis of human–XAI interactions. Consistently, studies found that explanations increase system understandability (e.g., Naiseh et al., 2023; Zhang et al., 2021). However, some studies showed that higher understandability does not always correlate with higher trust (Zhang et al., 2021). Human–XAI research on users' responses to different explanation configurations can be divided into two groups: First, responses can be measured by observing human interactions with AI systems and collecting behavioral data such as acceptance and rejection rates (e.g., Cramer et al., 2008; Giboney et al., 2015; Lu & Zhang, 2025). Interestingly, studies examining users' ability to detect errors found that explanations hampered users' performance in identifying errors while increasing overall performance (e.g., Naiseh et al., 2023; Poursabzi-Sangdeh et al., 2021). Second, prior studies examined behavioral intentions (e.g., Aslan et al., 2022; Panigutti et al., 2022) and overall satisfaction with AI systems (e.g., Sivaraman et al., 2023; Guesmi et al., 2022). However, most researchers did not find significant effects of explanations on behavioral intentions (e.g., Panigutti et al., 2022; Leichtmann et al., 2024). Based on our findings, we conceptualize user responses as a critical component of the explanandum, as they reflect how effectively explanations support human decision-making, particularly in distinguishing between correct and incorrect outcomes. To better understand how user perceptions (organism) give rise to observable behaviors (responses), established theoretical frameworks provide critical conceptual guidance. The effort-accuracy tradeoff framework, for example, suggests that users allocate cognitive resources based on the perceived balance between effort and expected decision quality, shaping advice acceptance (Mao & Benbasat, 2001). Similarly, expectation-confirmation theory explains how unmet expectations diminish usage intentions (Matt & Schlusche, 2022), while the heuristic-systematic model accounts for how users shift between intuitive and analytical reasoning when judging AI outputs (Naiseh et al., 2023; Shin, 2021). Finally, cognitive load theory clarifies how overload or poorly structured explanations can hinder appropriate reliance (Anderson et al., 2020; Westphal et al., 2023). By clarifying the cognitive and motivational processes behind user perceptions and actions, these theories deepen our understanding of how individuals interpret, respond to, and ultimately engage with the AI system and its outputs—the explanandum.

To improve human performance as the main consequence (4), several researchers hypothesize that providing explanations should benefit overall task performance (e.g., Hamm et al., 2023; Herm, 2023). In many cases, task performance is closely related to appropriate reliance on AI advice. For instance, Lai and Tan (2019) measure human performance as classification accuracy. In recommender systems, performance is often related to the consumption rate (e.g., Lee & Ram, 2023; Chen et al., 2023). However, classification accuracy cannot be considered in isolation. Task performance can be represented as a function of accuracy and decision-making duration (e.g., Cheng et al., 2019; Lee & Chew, 2023), resulting in efficiency. As AI technologies evolve and exhibit greater levels of agency, new challenges emerge in how humans engage with these systems - such as prompting LLMs, delegating tasks to AI agents, or supervising multiple agents - highlighting the need for further research into their long-term consequences.

7.2 Application of the Conceptual Framework of Human–XAI Interaction

The conceptual framework provides researchers and practitioners with a theory-driven scaffold (Paré et al., 2015) to systematically design, evaluate, and extend human–XAI interactions. In particular, it provides a foundation for exploring the explainability of emerging AI systems, such as GenAI-based agents and agentic AI systems composed of multiple AI agents, as well as the evolving forms of human-AI delegation that accompany these technological advances. By distinguishing between stimuli (e.g., explanation types and modalities), internal psychological processes (e.g., trust, understanding, cognitive load), and behavioral as well as real-world consequences, the human–XAI interaction framework guides hypothesis formulation and identification of mediators and moderators such as user expertise or task complexity. Researchers can apply it to explore how specific explanation designs affect user responses, while practitioners can use the systematization of stimuli to tailor explanation strategies to user needs and context. As emerging technologies, such as GenAI systems, increasingly offer interactive and multimodal outputs, the framework helps evaluate how explanation formats and multiple explanations influence trust calibration and decision-making by breaking down the link between explanans and explanandum through our conceptualization of stimuli, organism, response, and consequences. It further supports the identification of research gaps and offers measurable constructs for examining why explanations succeed or fail, making it a practical and conceptual tool for advancing both empirical research and design in the field of XAI.

To illustrate the application of the human–XAI interaction framework, consider a radiologist delegating to a multi-agent, LLM-enabled AI system to support lung cancer diagnosis from chest X-rays. In this scenario, different AI agents deliver varied explanation types: a visual Imaging Agent offers why explanations (e.g., "spiculated mass in the upper left lobe") grounded in radiological features, while a Counterfactual Agent provides how-to or what-if explanations (e.g., "if the lesion had smoother margins, it would fall below the malignancy threshold"). These explanations are presented through diverse modalities such as visual overlays (e.g., heatmaps), structured text, and probabilistic scores, either upfront or on-demand, forming the stimuli in this example. The organism layer reflects the radiologist's internal cognitive processes - such as perceived quality and clarity of each agent's explanation - which are shaped by contextual factors like the radiologist's domain expertise and the complexity of the case (e.g., easily vs. difficult-to-identify cancer). Perceptions of individual explanations shape users' perceived transparency of the overall multi-agent system, including its constituent agents and their decisions, while the system's complexity may contribute to information overload. In the response layer, these perceptions translate into behaviors such as the degree of agreement with AI output and the ability to detect incorrect advice. Finally, the consequence layer captures downstream effects, including diagnostic accuracy (i.e., task performance), decision time, and the radiologist's engagement with the AI agents in terms of continued interactions. This example illustrates how the conceptual framework supports the structured design and evaluation of explanation features across dynamic, high-stakes tasks, and clarifies the distinct role each layer plays in capturing human–XAI interaction, also in emerging settings of complex multi-agent collaborations (Fig. 5).

7.3 Research Agenda

Given our findings and the resulting human–XAI interaction framework, we reveal five overarching future research directions for experimental research on XAI in IS and HCI. In particular, they provide a foundation for exploring the explainability of emerging AI systems, such as GenAI-based agents and agentic AI systems composed of multiple AI agents, as well as the evolving forms of human-AI delegation that arise alongside these technological advances. While prior research has focused on explanations in the context of recommender systems, decision support tools, or simple conversational agents, there is now a need to extend XAI to GenAI systems that allow users to delegate complex tasks to multiple, highly adaptive AI agents. Our research agenda delineates key pathways for advancing XAI to meet these emerging challenges.

7.3.1 Further Research on Stimuli

Future Research Avenue 1: Theoretical Foundations for XAI Phenomena The first avenue for future research concerns the diverse configurations of explanations in AI systems. As highlighted in our literature review, there is considerable variation in explanation types, instantiations, and representations. However, our analysis reveals that many of these explanations lack a solid theoretical foundation. They are often designed and evaluated without reference to established theoretical frameworks, which limits both their interpretability and generalizability. To bridge explanans and explanandum, deeper theoretical engagement is needed, particularly to account for users' perceptions and behavioral responses as the mediating organism. Current explanation design in XAI often relies on broader social science perspectives (Miller, 2019) or reasoning-based categorizations (Liao et al., 2020), limiting its ability to address the complexity of human–XAI interaction. We argue that drawing

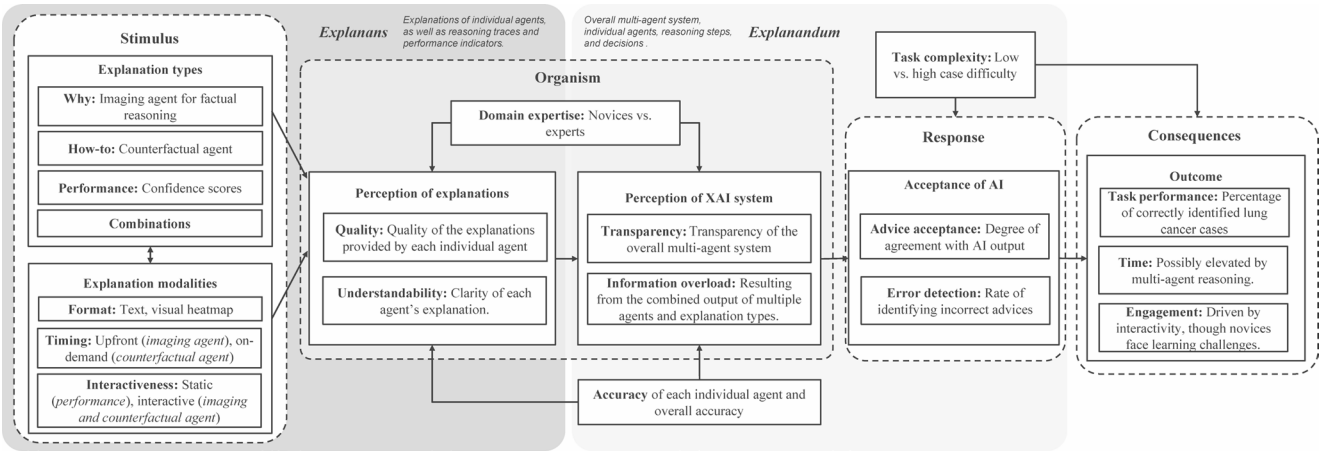

Fig. 5 Illustrative example for a multi-agent, LLM-enabled AI system to support lung cancer diagnosis

on established theories from cognitive psychology (e.g., dual-process models, cognitive load theory), communication studies (e.g., Toulmin's argumentation model, signaling theory), and decision science (e.g., expectation-confirmation frameworks, mental models) can provide stronger conceptual grounding. Further research is needed to understand how these theories can link design features to users' cognitive and behavioral responses, thereby informing the development of more effective and human-centered explanations - especially in the context of emerging GenAI systems and delegation scenarios. Furthermore, future research on XAI configurations should explore the combined effects of multiple explanation and reasoning types, and identify potential conflicts that may arise in settings involving multiple AI agents. Our human–XAI interaction framework can effectively guide future research in designing theory-based experiments and explaining causal chains comprehensively.

Future Research Avenue 2: Emerging Class of Explanations for GenAI Systems (e.g., LLMs, Agentic AI) While prior research has predominantly examined the effects of explanations in decision-support and recommender systems, our analysis points out that literature lacks the investigation of new types of AI systems such as GenAI (Kim et al., 2025; Steyvers et al., 2025) or agentic AI (Acharya et al., 2025) as well as the novel forms of delegation. For example, conversational agents such as ChatGPT differentiate from traditional decision-support, knowledge-based, and recommender systems by their conversational and highly interactive style (Wang et al., 2023). We raise the question of how existing XAI design approaches (Ashktorab et al., 2019; Wilkinson et al., 2021; Jiang et al., 2022) can be adapted to GenAI systems, when interactions with these information systems are becoming more conversational and intuitive. There are still unanswered questions regarding how to incorporate XAI into AI agents and emerging multi-agent systems, how to design such stimuli in a user-centric way, and how employees perceive provided rationales and explanations based on human-like dialogues and language. Early studies have demonstrated the potential of deliberative AI systems that facilitate dimension-specific opinion elicitation, support iterative refinement of decisions, and enable structured human–AI dialogues (Ma et al., 2025). The advances in agentic AI regarding tool usage and reasoning capabilities paired with greater degrees of autonomy (Acharya et al., 2025) call for building upon our conceptual framework to understand these new forms of interactions and delegation from a perspective of explainability and transparency. A pivotal area of exploration is determining how LLMs can convert technical explanations into comprehensible natural language (Chen et al., 2019). This translation is essential for bridging the gap between complex AI processes and user understanding. Moreover, enabling LLMs to clarify their data processing methods and reasoning steps is essential for transparency and trust (Zhao et al., 2024). Another research direction is to explore how on-demand explanations, prompted by user interactions with GenAI, enhance understanding. Investigating transfactual reasoning in this context could offer valuable insights into user comprehension and effective AI communication.

7.3.2 Further Research on Organism and Response

Future Research Avenue 3: Exploring the Relationship Between Explanans and Explanandum Despite the increasing number of empirical research (Haque et al., 2023), the synthesis of literature spanning experimental studies reveals that the underlying mechanisms of human–XAI interaction - connecting explanans and explanandum - remain fragmented. This is demonstrated by the large number of conflicting results and findings regarding certain effects and variables. For instance, explanations can improve users' understanding of the AI (Leichtmann et al., 2024), but may be ineffective if users lack domain knowledge (Wang & Yin, 2021). Thereby, our analysis highlights the critical role of context in shaping individuals' cognitive and affective responses to XAI. Leveraging established theories can help clarify these contextual influences on user behavior, particularly by informing the development of theory-driven XAI interfaces and conceptually grounded models. Our human–XAI interaction framework can guide researchers in their future studies on better understanding the relationship between explanans and explanandum. In addition, our findings raise important questions about trade-offs and potential conflicting effects. For example, while increased information density can enhance transparency, it may also lead to information overload (Chien et al., 2022). These conflicting effects and trade-offs highlight promising directions for future research aimed at examining the complex interplay between explanans and explanandum in terms of complex AI agents and multi-agent systems.

Future Research Avenue 4: User-Related Examinations of Explanation Effects The results of previous research showed that explanations possess different effects on different types of end users and stakeholders (Langer et al., 2021). As GenAI agents become increasingly integrated into both leisure and work contexts (Brynjolfsson et al., 2025), it is essential to investigate how different user groups delegate tasks and interact with interactive XAI interfaces, particularly to ensure these systems are adaptable to users with diverse skill levels and degrees of digital literacy (Ehsan et al., 2024). A critical challenge lies in creating XAI interfaces that are accessible to both experts and non-experts,

necessitating designs that seamlessly adapt to a wide expertise spectrum. Additionally, developing adaptive and personalized explanations tailored to individual users' cognitive styles and personal characteristics is crucial (Riefle et al., 2024). This approach ensures that XAI systems communicate effectively across diverse user backgrounds, fostering a more inclusive and user-friendly AI experience. Including human characteristics in conceptual models can hence provide more thorough analyses of causal relations and effects of explanations.

7.3.3 Further Research on Consequences

Future Research Avenue 5: Engagement in Human–XAI Interaction Given the nature of emerging GenAI systems, we expect to see a greater impact of AI technologies and human-AI delegation on the workplace and personal development of humans. Future XAI research, for example, should examine how explanations influence the prompting skills and performance of users interacting with LLMs. Part of this shift towards delegating tasks and processes to AI systems raises questions regarding the role of explanations on such human-in-the-loop and human-on-the-loop configurations. Humans are tasked with supervising AI systems, providing and maintaining training data, and ensuring continuous development. Further research is required to understand how XAI systems effect the willingness to delegate decisions and tasks (Baird & Maruping, 2021). Overall, there is a lack of research examining the real-world impacts and consequences of human–XAI interaction. Following the S-O-R-C model, consequences may impact the non-observable organism and response, resulting in a cycle of effects in the long term. Yet, research did not try to account for the impacts of the task-related outcomes on, for example, long-term perceptions and behavioral intentions concerning the XAI system and corresponding delegation behaviour.

8 Implications and Limitations

8.1 Implications for Theory

Our study contributes to the XAI literature by offering a human–XAI interaction framework for researchers and practitioners to undertake impactful studies on XAI and to enhance delegation to (Gen)AI systems through improved explanations. The framework and the aggregation of recurring measurements in XAI literature offer an entry point for future research and prepare scholars for conceptualizing experiments and contributing to the dedicated literature by considering stimuli, organisms, responses, and consequences. In contrast to the vast number of prior reviews on XAI (e.g., Brasse et al., 2023; Schwalbe & Finzel, 2023), our study emphasizes solely experimental research and complements prior surveys on empirical user studies (Haque et al., 2023; Rong et al., 2023) and user-centered XAI design (Laato et al., 2022) by discussing the how and why of XAI effects.

We integrate insights from HCI and IS to delineate clear pathways for researchers in both fields, and to motivate more empirical and theoretical research on the effects and the inner workings of human–XAI interaction, especially by the IS community (Brasse et al., 2023). By synthesizing the research on making AI systems more transparent and explainable to end users, we provide researchers and practitioners the foundations for building and evaluating responsible AI (Dennehy et al., 2023) and overcoming the typical barriers of responsible AI, such as transparency or fairness (Merhi, 2023). Further empirical research on XAI can help make AI systems, particularly emerging forms such as GenAI and agentic AI, more transparent, fair, and non-discriminatory, thereby supporting efforts to reduce the digital divide (Gupta et al., 2022; Vassilakopoulou & Hustad, 2023).

Our main contribution lies in the aggregation of studied effects within a novel conceptual framework of human–XAI interaction based on the S-O-R-C framework (Mehrabian & Russell, 1974). While prior reviews on XAI in IS and HCI have mapped the what, who, and why of explainability (e.g., Brasse et al., 2023; Laato et al., 2022), the ***how*** (i.e., the mechanisms linking explanations to user perceptions, behaviors, and outcomes) remained under-theorized. We addressed this gap by synthesizing experimental evidence and conceptualizing explanation features as causal stimuli in human–AI interaction. Based on the S-O-R-C model, our framework enables a causal unpacking of how explanation types and modalities (stimuli) shape user perceptions (organism), behaviors (responses), and real-world impacts (consequences). Additionally, we incorporate the philosophical concepts of explanans and explanandum (Hempel & Oppenheim, 1948) into our conceptual framework to better capture the unobservable mechanisms within the organism. This integration helps clarify how specific explanatory elements (explanans) shape users' perceptions of the AI system or its decisions (explanandum).

In contrast to prior research, we structure explanation types according to types of human reasoning, and we found that experimental research lacks a systematic derivation of explanation configurations and their differentiated effects. These reasoning types can help to categorize stimuli and homogenize scattered XAI research. The extracted modalities help to consider any additional or alternative configurations that are closely related to the underlying explanation types. Our human–XAI interaction framework proposes

a combined analysis of explanation types and modalities because of their strong interrelatedness. Furthermore, we shed light on the inner states of users interacting with XAI systems by separating the organism component into perceptions of explanations and perceptions of the XAI system. Thus, by analyzing the conceptual models, we found that the processing of the explanations and their representations influences the perceptions of the overall system, including these explanations. Our contribution lies in detaching the explanations from the underlying AI model. However, we acknowledge that explanations are closely connected to the AI model. This knowledge is crucial when developing and designing explanations for new types of AI systems (i.e., LLMs, agentic AI). Our study moves beyond the assumptions of monolithic system-level acceptance models like TAM and TPB by accounting for both granular, feature-level and system-level interactions, making it especially suited to contemporary, modular AI systems such as LLMs, GenAI, and agentic AI applications. The organism, in turn, determines the response, i.e., the behavior and intentions of the user. Our results emphasize measuring the responses objectively by tracking the classification accuracy or the changes in the user's actual behavior. Finally, we found that the consequences of XAI use are crucial for the HCI and IS communities to study real-world impacts. Thereby, we extend prior conceptualizations proposed by Haque et al. (2023).

8.2 Implications for Practice

Practitioners can especially benefit from making AI decisions more understandable to their customers, thereby increasing customer trust and subsequent behavioral intentions (Papagiannidis et al., 2023). Our results help AI developers and operators to design and assess XAI systems systematically. Additionally, the findings regarding the different configurations of XAI guide practitioners to individualize and adapt XAI systems purposefully.

The results highlight the need to understand the user's characteristics as these moderate the impact of XAI treatments on targeted outcome variables. Thus, XAI practice must better collect and analyze human behavior, preferences, and cognitive styles to adapt explanation types and modalities to specific users or user groups. Insights on the perception of XAI can guide practitioners to reflect on their XAI design and evaluate it against measures such as explanation quality, understandability, and information sufficiency. Our findings offer practitioners a catalog of suitable subjective and objective measurements to evaluate designed explanations.

Furthermore, our analysis reveals that apart from task performance, explanations can also stimulate engagement with AI or learning through understanding the AI's decisions. This outlines the potential for organizations to build deeper interactions between humans and AI. Overall, the resulting conceptual framework and proposed future research directions provide companies aiming to develop responsible GenAI systems with a comprehensive foundation for addressing XAI design and emerging forms of human-AI delegation.

8.3 Limitations

Our literature review comes with certain limitations. First, the scope of this SLR is not exhaustive. Extending the search process to other domains could provide additional findings. Furthermore, analogous to the selection of outlets, our selection of keywords may not cover all possible areas of XAI research. Furthermore, the number of findings ultimately coded and included in our dataset was limited. Hence, we do not intend to imply any form of direct causation between the components of the explanation and the results experienced by users. In addition, applying complementary coding schemes (Jeyaraj et al., 2006) is expected to provide more insights into relationships between variables and could also improve the understanding of the contradictory findings. Finally, we acknowledge that the majority of findings stem from studies published in conference proceedings, reflecting the current scarcity of experimental XAI research in top-ranked journals.

9 Conclusion

In conclusion, we contribute an overarching and contemporary perspective on the effects of explanations in human–XAI interaction by conducting a theoretical review on experimental XAI research. Our novel conceptual framework of human–XAI interaction offers other researchers a foundation for designing further experiments and creating targeted value for the HCI and IS community in times of growing human-AI delegation. Overall, the literature review fills the gap in understanding how XAI design impacts user responses and task-related outcomes and paves the way to further fruitful examinations of XAI concerning GenAI, LLM-based, and Agentic AI systems at the frontier of AI-driven augmentation.

Appendix A – Concept Matrix

Fig. 6 Concept matrix

	Task		Human characteristics					Explanation configurations									Human-XAI interaction											Outcomes		
								Types							Design		Perception of XAI configurations			Perception of XAI system							Acceptance of AI			Task outcome		
Author/s, year	Relevance	Complexity	Domain expertise	Technical literacy	Personality traits	Cognitive styles	Performance expectations	Why	How-to	What-if	What-else	Why-not	How	Performance	Others	Explanation modalities	Combinations	Quality	Understandability	Information sufficiency	Others	Trust	Understanding of AI	Utility	Transparency	Cognitive load	Human confidence	Justice / Fairness	Advice acceptance	Recognition of errors	Behavioral intention	Task performance	Task time	Engagement
Abdul et al., 2020								x								x			x		x	x				x					x		
Anderson et al., 2020			x					x		x							x									x							
Arnold et al., 2006		x	x					x					x		x	x																	
Ashktorab et al., 2019		x		x				x							x			x															
Aslan et al., 2024								x																					x	x				
Aslan et al., 2022				x				x					x		x									x	x					x			
Bauer, 2023			x					x				x									x										x			
Bhattacharya et al., 2024								x						x			x					x	x			x					x			
Binns et al., 2018								x	x		x				x										x		x	x						
Bo et al., 2025														x		x											x		x			x	x	
Brunk et al., 2019																																		
Chen et al., 2023								x			x											x	x								x			
Cheng et al., 2019			x	x																														
Chien et al., 2022								x						x									x											
Cramer et al., 2008											x	x		x								x	x	x	x			x		x				
Dai et al., 2023																															x			
Danry et al., 2023						x		x							x	x				x		x							x					
De Brito Duarte et al., 2025								x														x	x				x		x			x	x	
De Campos et al., 2024								x			x		x			x						x	x	x	x			x						
De Schutter and De Cremer 2024										x																		x						
De Zoeten et al., 2024							x	x					x									x	x		x		x							
Dikmen et al., 2022		x	x															x		x	x	x									x			
Dodge et al., 2019						x		x	x		x		x														x	x						
Ebermann et al., 2023		x				x																				x				x				
Ehsan et al., 2024				x				x							x	x			x			x					x							
Förster et al., 2020		x										x				x		x																
Frost et al., 2025								x	x											x		x	x	x							x			
Gibney et al., 2015						x										x		x				x		x			x		x			x		
Gönül et al., 2006								x					x	x	x	x		x				x							x					x
Graefe et al., 2023								x					x			x		x				x			x				x					
Guesmi et al., 2022			x	x	x	x												x				x		x	x				x					
Hadash et al., 2022															x	x			x															
He et al., 2023				x				x								x						x												
Ha and Kim, 2024								x								x						x									x			
Hamm et al., 2023								x														x			x					x	x			
Herm, 2023																x		x			x													
Herm et al., 2023								x	x		x	x				x										x					x	x		
Hudon et al., 2021		x						x								x										x	x						
Jang et al., 2024					x								x			x						x		x						x			
Jenkin et al., 2024		x													x													x					x
Jiang et al., 2022		x						x			x	x		x														x					
Jussupow et al., 2021																x						x			x									
Kaufmann et al., 2025	x	x	x					x						x			x	x				x					x		x					
Kim et al., 2024														x								x			x		x		x		x	x	x	
Kim et al., 2025								x									x						x	x			x		x		x	x	x	
Kulesza, 2013								x					x	x		x																		
Kunkel et al., 2019					x						x							x				x												
Lai et al., 2019								x			x			x	x							x									x			
Lai et al., 2022								x					x			x								x	x	x	x			x			x	
Lai et al., 2020								x						x		x															x			
Lee and Chew, 2023			x						x							x								x					x		x	x		
Lee et al., 2023	x																					x					x				x	x		
Lee et al., 2019																			x			x			x									x
Leichtmann et al., 2023			x	x				x			x					x			x			x	x		x					x	x			
Leichtmann et al., 2024			x											x		x						x									x	x		
Lima et al., 2023			x					x								x			x								x							
Lu et al., 2023								x			x					x		x		x					x									
Lu and Zhang, 2025			x					x																					x			x		
Mao and Benbasat, 2001			x					x							x	x		x											x			x		
Ma et al., 2025								x								x							x	x		x	x		x			x		
Martijn et al., 2022			x		x	x		x			x			x			x	x		x														
Matt and Schlusche, 2022							x															x		x										
Meske and Bunde, 2022								x						x		x		x		x		x		x		x								
Müller et al., 2024								x																x							x	x		
Musto et al., 2021								x							x	x		x				x			x									x
Naiseh et al., 2023								x	x		x		x	x					x			x	x	x	x			x	x	x	x			
Naveed et al., 2018						x		x			x				x	x				x									x					
Nimmo et al., 2024					x			x														x	x											x
Okoso et al., 2025					x											x											x		x					
Orzikulova et al., 2024								x														x		x							x			
Pafla et al. 2024								x				x			x	x		x				x					x				x	x		
Pálfi et al., 2024								x																							x			
Panigutti et al., 2022								x								x		x				x					x		x					
Papenmeier et al., 2022								x							x							x	x		x									
Poursabzi-Sangdeh et al., 2021			x												x	x										x			x	x				
Pu and Chen, 2007												x				x		x				x		x		x				x		x		
Rader et al. 2018				x				x					x		x					x					x	x			x	x				
Riefle et al., 2024				x		x		x			x				x							x												
Sadeghi et al., 2024								x																	x						x			
Schoeffer et al., 2024								x																			x		x	x	x			
Schröppel and Förster, 2024								x										x																
Shin 2021															x			x				x			x		x							
Shulner-Tal et al., 2025					x			x		x	x				x							x					x							
Silva et al., 2023								x	x		x			x				x				x	x		x				x			x	x	
Sivaraman et al., 2023								x			x					x								x			x		x					
Spitzer et al., 2024						x		x																								x	x	
Steyvers et al., 2024														x		x	x										x							
Tan et al., 2012								x					x														x					x	x	
Tsai and Brusilovsky, 2021								x			x					x		x				x		x	x					x				x
Tsai et al., 2021								x			x					x		x				x		x	x					x				x
Tutul et al., 2025				x	x			x					x		x							x							x			x		
van der Waa et al., 2020														x				x		x		x							x					
Wang and Benbasat, 2007								x				x	x									x												
Wang and Yin, 2021			x					x	x		x											x			x					x				
Wang et al., 2022								x	x			x			x																x	x		
Wastensteiner et al., 2021								x							x				x							x					x	x		
Weller et al., 2025								x	x																	x					x			
Westphal et al., 2023		x				x		x							x				x			x	x		x						x			
Wilkenson et al., 2021								x				x			x			x		x	x										x		x	
Woodcock, 2021			x					x	x				x		x							x												
Yang et al., 2020					x						x				x			x				x					x							
Yurrita et al., 2023			x	x				x	x		x				x												x							
Zhang and Lim, 2022									x			x														x	x					x	x	
Zhang et al., 2021																		x				x			x		x							
Zhang et al., 2020														x		x						x							x					
Zhang et al., 2023				x				x	x							x		x	x			x								x				
Zhao et al., 2024								x																										
Sum	**2**	**9**	**19**	**12**	**9**	**10**	**2**	**74**	**14**	**3**	**23**	**11**	**16**	**17**	**23**	**49**	**6**	**28**	**10**	**9**	**5**	**51**	**19**	**21**	**27**	**15**	**20**	**9**	**37**	**7**	**18**	**35**	**19**	**10**

Appendix B – Coding Scheme

Table 6 Full set of codes for meta data and theoretical foundation

Level 1	Level 2	Level 3
Type		
	Within-subjects	
	Between-subjects	
Other considerations		
	Number of features	
	Attention checks	
	Compensation	
	Timing of measurements	
	Exclude low and high durations	
	Randomization effects	
	Algorithm's accuracy	
	Learning effect	
	Confounding factors	
Type experiment		
	Real system	
	Wizard of oz	
	Vignette	
	Field Experiment	
	Online experiment	
Demographics		
Source of participants		
	Qualtrics	
	Prolific	
	University	
	MTurk	
N (number of participants)		
System		
	Decision support system	
	Recommender system	
	Conversational agent	
	Knowledge-based system (KBS)	
	Others	
		Intelligent system/automation
		User-adaptive systems
Domain		
	HR	
	Automotive	
	University	
	Finance	
	Healthcare	
	Diverse/multiple	
	Others	
		Supply chain
		Smart home
		Social media
Theoretical foundation		
	Cognitve learning theory	
	Social transparency	
	Cogntive dissonance theory	
	Explanation evaluation framework	
	Technology acceptance theory	
	Toulmin's model (argumentation theory)	

Table 6 (continued)

Level 1	Level 2	Level 3
	dual-process theory	
	Expectation-confirmation theory	
	Effort-accuracy tradeoff framework	
	Psychological contract violation	
	Cognitive Theory of Multimedia Learning	
	Cognitive load	
	Cognitive fit theory	
	Dunning-Kruger effect	
	Social sciences	
	Fairness theory	
	Socratic questioning	
	Mental models framework	
	Signaling theory	
	Grounding in communication	

Table 7 Full set of codes for the stimuli component

Level 1	Level 2	Level 3	Level 4
Explanation types			
	Why (factual reasoning)		
		Saliency maps	
		Natural language-based	
		Decision tree	
		Content-based explanations	
		Prediction rationale	
		Logical reasoning	
		Feature attribution methods	
		Feature-based/Feature-importance	
			Feature contribution
		Trace explanations	
		Keyword confirmation explanation	
		Highlighting keywords	
	What-else (analogical reasoning)		
	Performance (e.g., confidence, certainty)		
	How-to (counterfactual reasoning)		
	Why-not (contrastive reasoning)		
	How (global explanations)		
	What-if (transfactual)		
	Others		
		Data-centric vs. model-centric	
		Rule-based	
		Black-box vs. clear model	
		Showing the feature model	
		Degree of details (e.g., advanced, intermediate, basic)	
		What (e.g., AI existence, next steps)	
		Faithful and random explanations	
		Deep explanations	
		AI-framed questioning	
		Terminological	
		Out of vocabulary	
Explanation concepts (stimuli)			
	Feedforward vs. feedback explanations		
	User-centered (questions)		

Table 7 (continued)

Level 1	Level 2	Level 3	Level 4
	Explanation elements		
	Contextualized access to knowledge		
	Post-hoc vs. built-in		
	Local vs. global		
	Brief explanations vs. interactive visualization		
	Model-driven vs. knowledge-driven		
	On-demand vs. always (upfront)		
	Causal explanations vs. questioning		
	Interactive vs. static explanations		
	White-box vs. black-box		
	Model agnostic vs. model-specific		
	Trace or line of reasoning, justification, control, terminology		
Treatment configuration			
	Explanation (yes/no)		
	Comparing types of explanations		
	Combination of explanations		
	Explanation modalities		
		Format	
			Graphs
			Visual
			Text, natural language
		Degree of detail	
			Number of aspects
			Scope
			Completeness
			Degree of transparency
			Explanation complexity (Cognitive load)
		Interactivenesss	
			Static
			Interactive
		Characteristics of explanations	
			Relevant
			General
			Coherent
			Short
			Soundness
		Additional modalities	
			Adjacent vs. non-adjacent
			Impersonal vs. personal condition
			Presentation order
			Framing of explanations (positive/negative)
			Morphological clarity

Table 8 Full set of codes for the organism component

Level 1	Level 2	Level 3	Level 4
Perception of explanations			
	Quality		
		User satisfaction and usability	
		Persuasiveness	
		Explanation selection and preference	
		Perceived goodness of explanation	
	Understandability of explanations (e.g., explainability)		
	Information sufficiency		
		Informativeness	
		Information accuracy	
		Novelty of information	
	Others		
		Information processing	
		Explanation use	
		Perceived changes in explanations	
Perception of AI			
	Trust		
		Trusting beliefs	
		Perceived accountability	
		Reliance on AI	
			Truthfulnes
			Over-reliance
			Self-reliance
			Perceived reliability
			Appropriate Reliance
		Trust meter	
		Trust intentions	
		Credibility	
		Appropriate trust	
		Trust calibration	
	Understanding of AI		
		Interpretability	
		Simulatability	
		Simulation of model behavior	
		Comprehensibility	
		Self-reported understanding	
		Objective Understanding	
	Usefulness/quality (utility)		
		User experience	
		Perceived ease of use	
		Perceived usefulness of the system	
		Perceived (technical) Competence	
	Transparency (e.g., causability)		
	Cognitive load		
		Mental effort	
		Negative mood	
		Workload	
		Information overload	
	Perceived justice and fairness		
		Justice	
		Fairness	
	Confidence		
		Confidence in AI	
		Confidence in one's own decision	
	Others		

Table 8 (continued)

Level 1	Level 2	Level 3	Level 4
		Curiosity	
		Second chance	
		Metal model faithfulness (accuracy)	
		Social competence	

Table 9 Full set of codes for the response and consequence components

Level 1	Level 2	Level 3
Acceptance of AI (response)		
	Acceptance/reliance	
		Adherence
		Ratio of decision consistency
		Judgement congruency
		Weighted discernment
		Weight of advice
		Acceptance of recommendations
		Acceptance of the system
	Usage intention/behavioral intention	
		Explanation influence
	Error detection (i.e., recognizing incorrect AI advice)	
	Explanation access (i.e., use of explanations)	
Task outcome (consequences)		
	Task performance	
		Human classification performance
	Task Time	
	Engagement	
		Learning
		Willingness to help

Table 10 Full set of codes for human characteristics, task characteristics, and additional factors

Level 1	Level 2	Level 3
Human characteristics		
	Domain expertise	
	Technical literacy	
		Prior AI experience/literacy
		IT knowledge
		AI knowledge
	Cognitive styles/decision-making styles	
		Modes of information processing
		Cognitive reflection
		Need for cognition
		Cognitive abilities
	Personality traits	
		Trust propensity/disposition to trust
		Social orientation toward chatbots
		Personal innovativeness
	Performance expectations	
	Decision-making domain	
	Historical accuracy	
Task characteristics		
	Relevance of decision	
	Task complexity	
		Task type (level of knowledge assistance)
		User's epistemic uncertainty
		Data complexity
	Others	
		Service frustration
		Self-assessment (i.e., human estimation of their performance)
Additional configurations and confounding factors		
	AI performance (i.e., accuracy)	
	Presentation of the ML model	
	Initial failures	
	Tutorials	
	Answer correctness	
	Degree of human agency	

Acknowledgements Not applicable.

Authors' contributions All authors devised this study's main conceptual ideas. P.R. performed the source selection as well as the search procedure as part of the literature review. P.R. together with M.M.L. conducted the paper analysis iteratively and derived the concept matrix as well as the conceptual framework through coding the retrieved papers. C.P. and J.M.L. verified the results of the paper analyses and supervised the findings of this work as well as the derived avenues for future research. All authors were involved in discussing the results and contributing to the manuscript.

Funding Open Access funding enabled and organized by Projekt DEAL. This work was partially funded by the German Federal Ministry of Education and Research (Bundesministerium für Bildung und Forschung) until the 31 st of December 2023 under Grant agreement No. 02K18D060.

Data Availability Not applicable.

Declarations

Ethics Approval and Consent to Participate Not applicable.

Consent for Publication Not applicable.

Competing interests The authors have no conflicts of interest to declare that are relevant to the content of this article.

References

Abdul, A., von der Weth, C., Kankanhalli, M., & Lim, B. Y. (2020). COGAM: Measuring and moderating cognitive load in machine learning model explanations. In *Proceedings of the 2020 CHI conference on human factors in computing systems* (pp. 1–14). ACM.

Abedin, B., Meske, C., Junglas, I., Rabhi, F., & Motahari-Nezhad, H. R. (2022). Designing and managing human-AI interactions. *Information Systems Frontiers, 24*(3), 691–697.

Acharya, D. B., Kuppan, K., & Divya, B. (2025). Agentic AI: Autonomous intelligence for complex goals-A comprehensive survey. *IEEE Access*. https://doi.org/10.1109/access.2025.3532853

Anderson, A., Dodge, J., Sadarangani, A., Juozapaitis, Z., Newman, E., Irvine, J., et al. (2020). Mental models of mere mortals with explanations of reinforcement learning. *ACM Transactions on Interactive Intelligent Systems, 10*, 1–37. https://doi.org/10.1145/3366485

Arnold, V., Clark, N., Collier, P. A., Leech, S. A., & Sutton, S. G. (2006). The differential use and effect of knowledge-based system explanations in novice and expert judgment decisions. *MIS quarterly*, 79–97.

Arora, R. (1982). Validation of an SOR model for situation, enduring, and response components of involvement. *Journal of Marketing Research*, *19*(4), 505–516.

Arrieta, A., Díaz-Rodríguez, N., Ser, D., Bennetot, J., Tabik, A., Barbado, S., Garcia, A., S., et al. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, *58*, 82–115.

Ashktorab, Z., Jain, M., Liao, Q. V., & Weisz, J. D. (2019). Resilient chatbots: repair strategy preferences for conversational breakdowns. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. ACM.

Aslan, A. (2024). The effect of AI explanations on medical experts detecting misdiagnosis by AI systems. In *European Conference on Information Systems (ECIS)*. Aisel.

Aslan, A., Greve, M., Braun, M., & Kolbe, L. (2022). Doctors' dilemma - understanding the perspective of medical experts on AI explanations. In *International Conference on Information Systems*. Aisel.

Baird, A., & Maruping, L. M. (2021). The next generation of research on IS use: a theoretical framework of delegation to and from agentic IS artifacts. *MIS quarterly*, *45*(1).

Bauer, K., Zahn, M., & Hinz, O. (2023). Expl(AI)ned: The impact of explainable artificial intelligence on users' information processing. *Information Systems Research*. https://doi.org/10.1287/isre.2023.1199

Bhattacharya, A., Stumpf, S., Gosak, L., Stiglic, G., & Verbert, K. (2024). EXMOS: explanatory model steering through multifaceted explanations and data configurations. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. ACM.

Binns, R., van Kleek, M., Veale, M., Lyngs, U., Zhao, J., & Shadbolt, N. (2018). 'It's Reducing a human being to a percentage': perceptions of justice in algorithmic decisions. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (pp. 1–14). ACM. https://doi.org/10.1145/3173574.3173951

Bo, J. Y., Wan, S., & Anderson, A. (2025). To rely or not to rely? Evaluating interventions for appropriate reliance on large language models. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. ACM.

Brasse, J., Broder, H. R., Förster, M., Klier, M., & Sigler, I. (2023). Explainable artificial intelligence in information systems: A review of the status quo and future research directions. *Electronic Markets, 33*(1), Article 26.

Brunk, J., Mattern, J., & Riehle, D. M. (2019). Effect of transparency and trust on acceptance of automatic online comment moderation systems. In *2019 IEEE 21st Conference on Business Informatics (CBI)* (pp. 429–435). Moscow, Russia: IEEE. https://doi.org/10.1109/CBI.2019.00056

Brynjolfsson, E., Li, D., & Raymond, L. (2025). Generative AI at work. *The Quarterly Journal of Economics*. https://doi.org/10.1093/qje/qjae044

Campos, L., Fernández-Luna, J. M., & Huete, J. F. (2024). An explainable content-based approach for recommender systems: A case study in journal recommendation for paper submission. *User Modeling and User-Adapted Interaction, 34*(4), 1431–1465.

Chen, C., Tian, A. D., & Jiang, R. (2023). When post hoc explanation knocks: Consumer Responses to Explainable AI Recommendations. *Journal of Interactive Marketing, 59*(3), 234–250.

Chen, H., Chen, X., Shi, S., & Zhang, Y. (2019). Generate natural language explanations for recommendation. In *Proceedings of SIGIR 2019 Workshop on ExplainAble Recommendation and Search, Paris, France*. ACM.

Cheng, H. F., Wang, R., Zhang, Z., O'Connell, F., Gray, T., Harper, F. M. (2019). Explaining decision-making algorithms through UI: strategies to help non-expert stakeholders. In *Proceedings of the 2019 CHI conference on human factors in computing systems* (pp. 1–12). ACM. https://doi.org/10.1145/3290605.3300789

Chien, S.-Y., Yang, C.-J., & Yu, F. (2022). XFlag: Explainable Fake News Detection Model on Social Media. *International Journal of Human-Computer Interaction, 38*, 1808–1827. https://doi.org/10.1080/10447318.2022.2062113

Chromik, M., & Butz, A. (2021). Human-XAI interaction: a review and design principles for explanation user interfaces. *Human-Computer Interaction-INTERACT 2021, Bari, Italy* (pp. 619–640). Springer.

Cooper, H. M. (1998). *Synthesizing research: a guide for literature reviews* (Vol. 2). Sage.

Cramer, H., Evers, V., Ramlal, S., van Someren, M., Rutledge, L., Stash, N., et al. (2008). The effects of transparency on trust in and acceptance of a content-based art recommender. *User Modeling and User-Adapted Interaction, 18*, 455–496. https://doi.org/10.1007/s11257-008-9051-3

Dai, J., Zhang, C., Aliakseyeu, D., Peeters, S., &Ijsselsteijn, W. A. (2023). The effect of explanation design on user perceptionof smart home lighting systems: A mixed-method investigation. In *Proceedingsof the 2023 CHI conference on human factors in computing systems (*pp. 1–14).

Dalvi-Esfahani, M., Mosharaf-Dehkordi, M., Leong, L. W., Ramayah, T., & Kanaan-Jebna, A. M. J. (2023). Exploring the drivers of XAI-enhanced clinical decision support systems adoption: Insights from a stimulus-organism-response perspective. *Technological Forecasting and Social Change*. https://doi.org/10.1016/j.techfore.2023.122768

Danry, V., Pataranutaporn, P., Mao, Y., & Maes, P. (2023). Don't just tell me, ask me: AI systems that intelligently frame explanations as questions improve human logical discernment accuracy over causal ai explanations. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM. https://doi.org/10.1145/3544548.3580672

de Brito Duarte, R., Abreu, M. C., Campos, J., & Paiva, A. (2025). The Amplifying effect of explainability in AI-assisted decision-making in groups. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* (pp. 1–15). ACM.

de Zoeten, M. C., Ernst, C. P. H., & Rothlauf, F. (2024). The effect of explainable AI on AI-trust and its antecedents over the course of an interaction. In *European Conference on Information Systems (ECIS)*. Aisel.

Dell'Acqua, F., McFowland, E., Mollick, E. R., Lifshitz-Assaf, H., Kellogg, K., Rajendran, S. (2023). Navigating the jagged technological frontier: Field experimental evidence of the effects of AI on knowledge worker productivity and quality. *Harvard Business School Technology & Operations Mgt. Unit Working Paper*(24–013).

Dennehy, D., Griva, A., Pouloudi, N., Dwivedi, Y. K., Mäntymäki, M., & Pappas, I. O. (2023). Artificial intelligence (AI) and information systems: Perspectives to responsible AI. *Information Systems Frontiers, 25*(1), 1–7.

Dikmen, M., & Burns, C. (2022). The effects of domain knowledge on trust in explainable AI and task performance: A case of peer-to-peer lending. *International Journal of Human-Computer Studies, 162*, Article 102792. https://doi.org/10.1016/j.ijhcs.2022.102792

Dodge, J., Liao, Q. V., Zhang, Y., Bellamy, R. K. E., & Dugan, C. (2019). Explaining models: an empirical study of how explanations impact fairness judgment. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. ACM.

Ebermann, C., Selisky, M., & Weibelzahl, S. (2023). Explainable AI: The effect of contradictory decisions and explanations on users' acceptance of AI systems. *International Journal of Human-Computer Interaction, 39*, 1807–1826. https://doi.org/10.1080/10447318.2022.2126812

Ehsan, U., Passi, S., Liao, Q. V., Chan, L., Lee, I. H., Muller, M. (2024). The Who in XAI: How AI background shapes perceptions of AI explanations. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. ACM.

Elshan, E., Zierau, N., Engel, C., Janson, A., & Leimeister, J. M. (2022). Understanding the design elements affecting user acceptance of intelligent agents: Past, present and future. *Information Systems Frontiers*, *24*(3), 699–730.

Ferreira, J. J., & Monteiro, M. S. (2020). What are people doing about XAI user experience? A survey on AI explainability research and practice. In *Design, User Experience, and Usability. Design for Contemporary Interactive Environments* (pp. 56–73).

Förster, M., Klier, M., Kluge, K., & Sigler, I. (2020). Evaluating explainable artificial intelligence – what users really appreciate. In *European Conference on Information Systems (ECIS)*. Aisel.

Frost, L., Förster, M., & Klier, M. (2025). What's past is prologue: a novel method to explain AI-based time series forecasts. In *European Conference on Information Systems (ECIS)*. Aisel.

Giboney, J. S., Brown, S. A., Lowry, P. B., & Nunamaker, J. F., Jr. (2015). User acceptance of knowledge-based system recommendations: Explanations, arguments, and fit. *Decision Support Systems, 72*, 1–10.

Goldfried, M. R., & Sprafkin, J. N. (1976). Behavioral personality assessment. In J. T. Spence, R. C. Carson, & J. W. Thibaut (Eds.), *Behavioral approaches to therapy* (pp. 295–321). General Learning Press.

Gönül, M. S., Önkal, D., & Lawrence, M. (2006). The effects of structural characteristics of explanations on use of a DSS. *Decision support systems*, *42*(3), 1481-1493.

Graefe, J., Paden, S., Engelhardt, D., & Bengler, K. (2023). Human-centered explainability for intelligent vehicles—A user study. *International Journal of Human-Computer Interaction, 39*, 3237–3253. https://doi.org/10.1080/10447318.2023.2215098

Gregor, S., & Benbasat, I. (1999). Explanations from intelligent systems: theoretical foundations and implications for practice. *MIS quarterly*, 497–530.

Grodal, S., Anteby, M., & Holm, A. L. (2021). Achieving rigor in qualitative analysis: The role of active categorization in theory building. *Academy of Management Review, 46*, 591–612. https://doi.org/10.5465/amr.2018.0482

Guesmi, M., Chatti, M. A., Vorgerd, L., Ngo, T., Joarder, S., Ain, Q. U. (2022). Explaining user models with different levels of detail for transparent recommendation: a user study. In *Adjunct Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization* (pp. 175–183). Barcelona Spain: ACM. https://doi.org/10.1145/3511047.3537685

Gupta, M., Parra, C. M., & Dennehy, D. (2022). Questioning racial and gender bias in AI-based recommendations: Do espoused national cultural values matter? *Information Systems Frontiers*, *24*(5), 1465–1481.

Ha, T., & Kim, S. (2024). Improving trust in AI with mitigating confirmation bias: effects of explanation type and debiasing strategy for decision-making with explainable AI. *International Journal of Human-Computer Interaction*, 1–12.

Hadash, S., Willemsen, M. C., Snijders, C., & Ijsselsteijn, W. A. (2022). Improving understandability of feature contributions in model-agnostic explainable AI tools. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (pp. 1–9). ACM. https://doi.org/10.1145/3491102.3517650

Hamilton, K., Shih, S.-I., & Mohammed, S. (2016). The development and validation of the rational and intuitive decision styles scale. *Journal of Personality Assessment, 98*(5), 523–535.

Hamm, P., Klesel, M., Coberger, P., & Wittmann, H. F. (2023). Explanation matters: An experimental study on explainable AI. *Electronic Markets, 33*(1), Article 17.

Hannigan, T. R., McCarthy, I. P., & Spicer, A. (2024). Beware of botshit: How to manage the epistemic risks of generative chatbots. *Business Horizons*. https://doi.org/10.1016/j.bushor.2024.03.001

Haque, A. B., Islam, A. N., & Mikalef, P. (2023). Explainable artificial intelligence (XAI) from a user perspective: A synthesis of prior literature and problematizing avenues for future research. *Technological Forecasting and Social Change, 186*, Article 122120.

He, G., Aishwarya, N., & Gadiraju, U. (2025). Is conversational XAI all you need? Human-AI decision making with a conversational XAI assistant. In *Proceedings of the 30th International Conference on Intelligent User Interfaces*. ACM.

He, G., Kuiper, L., & Gadiraju, U. (2023). Knowing about knowing: an illusion of human competence can hinder appropriate reliance on AI systems. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM. https://doi.org/10.48550/ARXIV.2301.11333

Hempel, C. G., & Oppenheim, P. (1948). Studies in the logic of explanation. *Philosophy of Science, 15*(2), 135–175.

Herm, L. V. (2023). Impact of explainable AI on cognitive load: insights from an empirical study. In *European Conference on Information Systems (ECIS 2023)*. https://doi.org/10.48550/ARXIV.2304.08861

Herm, L. V., Heinrich, K., Wanner, J., & Janiesch, C. (2022). Stop ordering machine learning algorithms by their explainability! A user-centered investigation of performance and explainability. *International Journal of Information Management*, *69*.

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., et al. (2025). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems, 43*(2), 1–55.

Hudon, A., Demazure, T., Karran, A., Léger, P.-M., & Sénécal, S. (2021). Explainable Artificial Intelligence (XAI): How the Visualization of AI Predictions Affects User Cognitive Load and Confidence. In F. D. Davis, R. Riedl, J. vom Brocke, P.-M. Léger, A. B. Randolph, & G. Müller-Putz (Eds.), *Information Systems and Neuroscience* (Vol. 52, pp. 237–246). Springer International Publishing.

Jang, W., Chang, Y., Kim, B., Lee, Y. J., & Kim, S. C. (2024). Influence of personal innovativeness and different sequences of data

presentation on evaluations of explainable artificial intelligence. *International Journal of Human-Computer Interaction, 40*, 4215–4226. https://doi.org/10.1080/10447318.2023.2209995

Jenkin, T., Kelley, S., Ovchinnikov, A., & Ying, C. (2024). Explanation seeking and anomalous recommendation adherence in human-to-human versus human-to-artificial intelligence interactions. *Decision Sciences, 55*, 653–668. https://doi.org/10.1111/deci.12658

Jeyaraj, A., Rottman, J. W., & Lacity, M. C. (2006). A review of the predictors, linkages, and biases in IT innovation adoption research. *Journal of information technology, 21*(1), 1–23.

Jiang, J., Kahai, S., & Yang, M. (2022). Who needs explanation and when? Juggling explainable AI and user epistemic uncertainty. *International Journal of Human-Computer Studies, 165*, Article 102839. https://doi.org/10.1016/j.ijhcs.2022.102839

Jussupow, E., Meza Martinez, M. A., Maedche, A., & Heinzl, A. (2021). Is this system biased? - How users react to gender bias in an explainable AI system. In *International Conference on Information Systems (ICIS)*. Austin, Texas.

Kaufman, R. A., Broukhim, A., Kirsh, D., & Weibel, N. (2025). What did my car say? impact of autonomous vehicle explanation errors and driving context on comfort, reliance, satisfaction, and driving confidence. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* (pp. 1–17). ACM.

Keele, S. (2007). Guidelines for performing systematic literature reviews in software engineering. *EBSE Technical Report*.

Kenny, E. M., & Keane, M. T. (2021). On generating plausible counterfactual and semi-factual explanations for deep learning. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 35, pp. 11575–11585).

Kim, S. S. Y., Liao, Q. V., Vorvoreanu, M., Ballard, S., & Vaughan, J. W. (2024). I'm not sure, but… examining the impact of large language models' uncertainty expression on user reliance and trust. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (pp. 822–835).

Kim, S. S. Y., Vaughan, J. W., Liao, Q. V., Lombrozo, T., & Russakovsky, O. (2025). Fostering appropriate reliance on large language models: The role of explanations, sources, and inconsistencies. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* (pp. 1–19). ACM.

Kulesza, T., Stumpf, S., Burnett, M., Yang, S., Kwan, I., & Wong, W. K. (2013). Too much, too little, or just right? Ways explanations impact end users' mental models. In *2013 IEEE Symposium on Visual Languages and Human Centric Computing* (pp. 3–10). San Jose, CA, USA: IEEE. https://doi.org/10.1109/VLHCC.2013.6645235

Kunkel, J., Donkers, T., Michael, L., Barbu, C. M., & Ziegler, J. (2019). Let me explain: impact of personal and impersonal explanations on trust in recommender systems. In *Proceedings of the 2019 CHI conference on human factors in computing systems* (pp. 1–12). ACM. https://doi.org/10.1145/3290605.3300717

Laato, S., Tiainen, M., Najmul Islam, A. K., & Mäntymäki, M. (2022). How to explain AI systems to end users: A systematic literature review and research agenda. *Internet Research, 32*(7), 1–31.

Lai, V., & Tan, C. (2019). On human predictions with explanations and predictions of machine learning models: a case study on deception detection. In *FAT* '19: Conference on Fairness, Accountability, and Transparency, Atlanta, GA, USA, January 29–31*. https://doi.org/10.48550/ARXIV.1811.07901

Lai, V., Carton, S., Bhatnagar, R., Liao, Q. V., Zhang, Y., & Tan, C. (2022). Human-AI collaboration via conditional delegation: a case study of content moderation. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. ACM. https://doi.org/10.48550/ARXIV.2204.11788

Lai, V., Liu, H., & Tan, C. (2020). Why is 'Chicago' deceptive? Towards building model-driven tutorials for humans. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM. https://doi.org/10.48550/ARXIV.2001.05871

Langer, M., Oster, D., Speith, T., Hermanns, H., Kästner, L., Schmidt, E., et al. (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*, Article 103473.

Lee, K., & Ram, S. (2023). Explainable deep learning for false information identification: an argumentation theory approach. *Information Systems Research*, isre.2020.0097. https://doi.org/10.1287/isre.2020.0097

Lee, M. H., & Chew, C. J. (2023). Understanding the effect of counterfactual explanations on trust and reliance on AI for human-AI collaborative clinical decision making. In *Proceedings of the ACM on Human-Computer Interaction*. https://doi.org/10.48550/ARXIV.2308.04375

Lee, Y., Ha, M., Kwon, S., Shim, Y., & Kim, J. (2019). Egoistic and altruistic motivation: How to induce users' willingness to help for imperfect AI. *Computers in Human Behavior*, *101*, 180–196. https://doi.org/10.1016/j.chb.2019.06.009

Leichtmann, B., Hinterreiter, A., Humer, C., Streit, M., & Mara, M. (2024). Explainable artificial intelligence improves human decision-making: Results from a mushroom picking experiment at a public art festival. *International Journal of Human-Computer Interaction, 40*, 1–18. https://doi.org/10.1080/10447318.2023.2221605

Leichtmann, B., Humer, C., Hinterreiter, A., Streit, M., & Mara, M. (2023). Effects of explainable artificial intelligence on trust and human behavior in a high-risk decision task. *Computers in Human Behavior*. https://doi.org/10.1016/j.chb.2022.107539

Li, Y., Gan, Z., & Zheng, B. (2023). How do artificial intelligence chatbots affect customer purchase? Uncovering the dual pathways of anthropomorphism on service evaluation. *Information Systems Frontiers, 27*(1), 1–18.

Liao, Q. V., Gruen, D., & Miller, S. (2020). Questioning the AI: informing design practices for explainable AI user experiences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM. https://doi.org/10.1145/3313831.3376590

Lima, G., Grgić-Hlača, N., & Cha, M. (2023). Blaming humans and machines: what shapes people's reactions to algorithmic harm. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM. https://doi.org/10.48550/ARXIV.2304.02176

Lu, H., Ma, W., Wang, Y., Zhang, M., Wang, X., Liu, Y., et al. (2023). User perception of recommendation explanation: Are your explanations what users need? *ACM Transactions on Information Systems, 41*, 1–31. https://doi.org/10.1145/3565480

Lu, T., & Zhang, Y. (2025). 1+1 > 2? Information, humans, and machines. *Information Systems Research, 36*, 394–418. https://doi.org/10.1287/isre.2023.0305

Ma, S., Chen, Q., Wang, X., Zheng, C., Peng, Z., Yin, M. (2025). Towards human-ai deliberation: design and evaluation of llm-empowered deliberative ai for ai-assisted decision-making. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* (pp. 1–23). ACM.

Mao, J.-Y., & Benbasat, I. (2001). The effects of contextualized access to knowledge on judgement. *International Journal of Human-Computer Studies, 55*, 787–814. https://doi.org/10.1006/ijhc.2001.0507

Martijn, M., Conati, C., & Verbert, K. (2022). Knowing me, knowing you: Personalized explanations for a music recommender system. *User Modeling and User-Adapted Interaction, 32*, 215–252. https://doi.org/10.1007/s11257-021-09304-9

Matt, C., & Schlusche, C. (2022). How transparency measures can attenuate initial failures of intelligent decision support systems. In *European Conference on Information Systems (ECIS)*. Aisel.

Mehrabian, A., & Russell, J. A. (1974). A verbal measure of information rate for studies in environmental psychology. *Environment and Behavior*, *6*(2), 233.

Merhi, M. I. (2023). An assessment of the barriers impacting responsible artificial intelligence. *Information Systems Frontiers, 25*(3), 1147–1160.

Meske, C., & Bunde, E. (2022). Design principles for user interfaces in AI-based decision support systems: The case of explainable hate speech detection. *Information Systems Frontiers*. https://doi.org/10.1007/s10796-021-10234-5

Meske, C., Abedin, B., Klier, M., & Rabhi, F. (2022). Explainable and responsible artificial intelligence. *Electronic Markets*, *32*(4), 2103–2106.

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence, 267*, 1–38.

Müller, R., Thoß, M., Ullrich, J., Seitz, S., & Knoll, C. (2025). Interpretability is in the eye of the beholder: Human versus artificial classification of image segments generated by humans versus XAI. *International Journal of Human-Computer Interaction*, *41*(4), 2371–2393.

Musto, C., Gemmis, M., Lops, P., & Semeraro, G. (2021). Generating post hoc review-based natural language justifications for recommender systems. *User Modeling and User-Adapted Interaction, 31*, 629–673. https://doi.org/10.1007/s11257-020-09270-8

Naiseh, M., Al-Thani, D., Jiang, N., & Ali, R. (2023). How the different explanation classes impact trust calibration: The case of clinical decision support systems. *International Journal of Human-Computer Studies, 169*, Article 102941. https://doi.org/10.1016/j.ijhcs.2022.102941

Naveed, S., Donkers, T., & Ziegler, J. (2018). Argumentation-based explanations in recommender systems: conceptual framework and empirical results. In *Adjunct Publication of the 26th Conference on User Modeling, Adaptation and Personalization* (pp. 293–298). Singapore Singapore: ACM. https://doi.org/10.1145/3213586.3225240

Nimmo, R., Constantinides, M., Zhou, K., Quercia, D., & Stumpf, S. (2024). User characteristics in explainable AI: the rabbit hole of personalization? In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (pp. 1–13). ACM.

Ochmann, J., Michels, L., Tiefenbeck, V., Maier, C., & Laumer, S. (2024). Perceived algorithmic fairness: An empirical study of transparency and anthropomorphism in algorithmic recruiting. *Information Systems Journal*, *34*(2), 384–414.

Okoso, A., Yang, M., & Baba, Y. (2025). Do expressions change decisions? Exploring the impact of AI's explanation tone on decision-making. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* (pp. 1–22). ACM.

Orzikulova, A., Xiao, H., Li, Z., Yan, Y., Wang, Y., Shi, Y. (2024). Time2stop: adaptive and explainable human-ai loop for smartphone overuse intervention. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (pp. 1–20). ACM.

Pafla, M., Larson, K., & Hancock, M. (2024). Unraveling the dilemma of AI errors: exploring the effectiveness of human and machine explanations for large language models. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. ACM. https://doi.org/10.1145/3613904.3642934

Pálfi, B., Arora, K., Prociuk, D., & Kostopoulou, O. (2024). Risk prediction algorithms and clinical judgment: Impact of advice distance, social proof, and feature-importance explanations. *Computers in Human Behavior, 153*, Article 108102.

Panigutti, C., Beretta, A., Giannotti, F., & Pedreschi, D. (2022). Understanding the impact of explanations on advice-taking: a user study for AI-based clinical Decision Support Systems. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (pp. 1–9). ACM. https://doi.org/10.1145/3491102.3502104

Papagiannidis, E., Enholm, I. M., Dremel, C., Mikalef, P., & Krogstie, J. (2023). Toward AI governance: Identifying best practices and potential barriers and outcomes. *Information Systems Frontiers*, *25*(1), 123–141.

Papenmeier, A., Kern, D., Englebienne, G., & Seifert, C. (2022). It's complicated: The relationship between user trust, model accuracy and explanations in AI. *ACM Transactions on Computer-Human Interaction, 29*, 1–33. https://doi.org/10.1145/3495013

Paré, G., Trudel, M.-C., Jaana, M., & Kitsiou, S. (2015). Synthesizing information systems knowledge: A typology of literature reviews. *Information & Management, 52*(2), 183–199.

Poursabzi-Sangdeh, F., Goldstein, D. G., Hofman, J. M., Vaughan, W., J. W., & Wallach, H. (2021). Manipulating and measuring model interpretability. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM. https://doi.org/10.1145/3411764.3445315

Pu, P., & Chen, L. (2007). Trust-inspiring explanation interfaces for recommender systems. *Knowledge-Based Systems*, *20*, 542–556. https://doi.org/10.1016/j.knosys.2007.04.004

Rader, E., Cotter, K., & Cho, J. (2018). Explanations as mechanisms for supporting algorithmic transparency. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (pp. 1–13). ACM. https://doi.org/10.1145/3173574.3173677

Riefle, L., Hemmer, P., Benz, C., & Vössing, M. (2024). The role of cognitive styles for explainable AI. In *European Conference on Information Systems (ECIS)*. Aisel.

Rong, Y., Leemann, T., Nguyen, T. T., Fiedler, L., Qian, P., Unhelkar, V. (2023). Towards human-centered explainable ai: a survey of user studies for model explanations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Rzepka, C., & Berger, B. (2018). *User interaction with AI-enabled systems*. A systematic review of IS research.

Sadeghi, K., Ojha, D., Kaur, P., Mahto, R. V., & Dhir, A. (2024). Explainable artificial intelligence and agile decision-making in supply chain cyber resilience. *Decision Support Systems, 180*, Article 114194.

Schneider, J., Meske, C., & Vlachos, M. (2023). Deceptive XAI: Typology, creation and detection. *SN Computer Science, 5*(1), Article 81.

Schoeffer, J., De-Arteaga, M., & Kuehl, N. (2024). Explanations, fairness, and appropriate reliance in human-AI decision-making. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (pp. 1-18).

Schröppel, P., & Förster, M. (2024). Exploring XAI users' needs: a novel approach to personalize explanations using contextual bandits. In *European Conference on Information Systems (ECIS)*. Aisel.

Schutter, L., & Cremer, D. (2024). How counterfactual fairness modelling in algorithms can promote ethical decision-making. *International Journal of Human-Computer Interaction, 40*, 33–44. https://doi.org/10.1080/10447318.2023.2247624

Schwalbe, G., & Finzel, B. (2023). A comprehensive taxonomy for explainable artificial intelligence: A systematic survey of surveys on methods and concepts. *Data Mining and Knowledge Discovery, 38*(5), 1–59.

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*, Article 102551. https://doi.org/10.1016/j.ijhcs.2020.102551

Shulner-Tal, A., Kuflik, T., Kliger, D., & Mancini, A. (2024). Who made that decision and why? Users' perceptions of human versus AI decision-making and the power of explainable-AI.

International Journal of Human-Computer Interaction, 41, 1–18. https://doi.org/10.1080/10447318.2024.2348843

Silva, A., Schrum, M., Hedlund-Botti, E., Gopalan, N., & Gombolay, M. (2023). Explainable Artificial Intelligence: Evaluating the Objective and Subjective Impacts of xAI on Human-Agent Interaction. *International Journal of Human-Computer Interaction, 39*, 1390–1404. https://doi.org/10.1080/10447318.2022.2101698

Sivaraman, V., Bukowski, L. A., Levin, J., Kahn, J. M., & Perer, A. (2023). Ignore, trust, or negotiate: understanding clinician acceptance of AI-based treatment recommendations in health care. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM. https://doi.org/10.48550/ARXIV.2302.00096

Spitzer, P., Kühl, N., Goutier, M., Kaschura, M., & Satzger, G. (2024). Transferring domain knowledge with (X)AI-based learning systems. In *European Conference on Information Systems (ECIS)* (Vol. 2). Aisel.

Steyvers, M., Tejeda, H., Kumar, A., Belem, C., Karny, S., Hu, X., et al. (2025). What large language models know and what people think they know. *Nature Machine Intelligence*. https://doi.org/10.1038/s42256-024-00976-7

Tan, W. K., Tan, C. H., & Teo, H. H. (2012). Consumer-based decision aid that explains which to buy: Decision confirmation or overconfidence bias? *Decision Support Systems*, *53*(1), 127–141.

Tsai, C. H., You, Y., Gui, X., Kou, Y., & Carroll, J. M. (2021). Exploring and promoting diagnostic transparency and explainability in online symptom checkers. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (pp. 1–17). ACM. https://doi.org/10.1145/3411764.3445101

Tsai, C.-H., & Brusilovsky, P. (2021). The effects of controllability and explainability in a social recommender system. *User Modeling and User-Adapted Interaction, 31*, 591–627. https://doi.org/10.1007/s11257-020-09281-5

Tutul, A. A., Nirjhar, E. H., & Chaspari, T. (2024). Investigating trust in human-AI collaboration for a speech-based data analytics task. *International Journal of Human–Computer Interaction*, 1–19. https://doi.org/10.1080/10447318.2024.2328910

van der Waa, J., Schoonderwoerd, T., van Diggelen, J., & Neerincx, M. (2020). Interpretable confidence measures for decision support systems. *International Journal of Human-Computer Studies*. https://doi.org/10.1016/j.ijhcs.2020.102493

Vassilakopoulou, P., & Hustad, E. (2023). Bridging digital divides: A literature review and research agenda for information systems research. *Information Systems Frontiers*, *25*(3), 955–969.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*.

vom Brocke, J., Simons, A., Niehaves, B., Niehaves, B., Reimer, K., Plattfaut, R. (2009). Reconstructing the giant: on the importance of rigour in documenting the literature search process. In *European Conference on Information Systems (ECIS)*. Aisel.

Vössing, M., Kühl, N., Lind, M., & Satzger, G. (2022). Designing transparency for effective human-AI collaboration. *Information Systems Frontiers*, *24*(3), 877–895.

Wang, B., Li, G., & Li, Y. (2023). Enabling conversational interaction with mobile ui using large language models. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (pp. 1–17). ACM.

Wang, D., Yang, Q., Abdul, A., & Lim, B. Y. (2019). Designing theory-driven user-centric explainable AI. In *Proceedings of the 2019 CHI conference on human factors in computing systems* (pp. 1–15). ACM.

Wang, X., & Yin, M. (2021). Are explanations helpful? A comparative study of the effects of explanations in AI-assisted decision-making. In *26th International Conference on Intelligent User Interfaces* (pp. 318–328). College Station TX USA: ACM. https://doi.org/10.1145/3397481.3450650

Wang, Y., Venkatesh, P., & Lim, B. Y. (2022). Interpretable directed diversity: leveraging model explanations for iterative crowd ideation. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (pp. 1–28). ACM. https://doi.org/10.1145/3491102.3517551

Wastensteiner, J., Weiss, T. M., Haag, F., & Hopf, K. (2021). Explainable AI for tailored electricity consumption feedback - an experimental evaluation of visualizations. In *European Conference on Information Systems (ECIS), Marrakesh, Morocco*. https://doi.org/10.48550/ARXIV.2208.11408

Webster, J., & Watson, R. T. (2002). Analyzing the past to prepare for the future: Writing a literature review. *MIS Quarterly*. https://doi.org/10.2307/4132319

Weller, N., Schmitt, A., Fahse, T. B., & Blohm, I. (2025). Improving AI-assisted decision-making: insights into example-based explanations and cognitive load in sales forecasting. In *European Conference on Information Systems (ECIS)*. Aisel.

Westphal, M., Vössing, M., Satzger, G., Yom-Tov, G. B., & Rafaeli, A. (2023). Decision control and explanations in human-AI collaboration: Improving user perceptions and compliance. *Computers in Human Behavior, 144*, Article 107714. https://doi.org/10.1016/j.chb.2023.107714

Wilkinson, D., Alkan, Ö., Liao, Q. V., Mattetti, M., Vejsbjerg, I., Knijnenburg, B. P., et al. (2021). Why or why not? The effect of justification styles on chatbot recommendations. *ACM Transactions on Information Systems, 39*, 1–21. https://doi.org/10.1145/3441715

Woodcock, C., Mittelstadt, B., Busbridge, D., & Blank, G. (2021). The impact of explanations on layperson trust in artificial intelligence–driven symptom checker apps: experimental study. *Journal of medical Internet research*, *23*(11).

Xiao, B., & Benbasat, I. (2011). Product-related deception in e-commerce: a theoretical perspective. *MIS quarterly*, 169–195.

Yang, F., Huang, Z., Scholtz, J., & Arendt, D. L. (2020). How do visual explanations foster end users' appropriate trust in machine learning? In *Proceedings of the 25th International Conference on Intelligent User Interfaces* (pp. 189–201). Cagliari Italy: ACM. https://doi.org/10.1145/3377325.3377480

Yurrita, M., Draws, T., Balayn, A., Murray-Rust, D., Tintarev, N., & Bozzon, A. (2023). Disentangling fairness perceptions in algorithmic decision-making: the effects of explanations, human oversight, and contestability. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (pp. 1–21). ACM. https://doi.org/10.1145/3544548.3581161

Zhang, T., Yang, X. J., & Li, B. (2023). May I ask a follow-up question? Understanding the benefits of conversations in neural network explainability. *International Journal of Human-Computer Interaction, 41*, 5623–5647. https://doi.org/10.48550/ARXIV.2309.13965

Zhang, W., & Lim, B. Y. (2022). Towards relatable explainable AI with the perceptual process. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (pp. 1–24). ACM. https://doi.org/10.1145/3491102.3501826

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. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (pp. 295–305). https://doi.org/10.48550/ARXIV.2001.02114

Zhang, Z., Genc, Y., Wang, D., Ahsen, M. E., & Fan, X. (2021). Effect of AI explanations on human perceptions of patient-facing AI-powered healthcare systems. *Journal of Medical Systems, 45*, Article 64. https://doi.org/10.1007/s10916-021-01743-6

Zhao, H., Chen, H., Yang, F., Liu, N., Deng, H., Cai, H., et al. (2024). Explainability for large language models: A survey. *ACM Transactions on Intelligent Systems and Technology*, *15*(2), 1–38.

Zierau, N., Elshan, E., Visini, C., & Janson, A. (2020). A review of the empirical literature on conversational agents and future research directions. In *International Conference on Information Systems (ICIS)*. Aisel.

Philipp Reinhard is a Research Associate and PhD Candidate in Information Systems at the University of Kassel, Germany. He holds master's degrees in Information Systems and in Entrepreneurship & Innovation Management from TU Darmstadt. His research focuses on the human-centered design of AI-based services, including generative AI and agentic AI. His work has appeared in top-tier IS journals such as Information Systems and the Journal of Service Research.

Mahei M. Li is a Project Lead at the University of St. Gallen and Research Group Lead at the University of Kassel. His research examines digital transformation, service systems, and GenAI and Agentic AI implementation, design, and management. He serves on the board of the AIS SIG Services.

Christoph Peters is a Full Professor of Information Systems, esp. Digital Process Management at the University of the Bundeswehr Munich, Germany. His research focuses on digital and AI-based services and processes, as well as human-centered technology. His work has been published in JMIS, BISE, JOSM, MISQE, CAIS, ICIS, and ECIS.

Jan Marco Leimeister is Chair Professor and Managing Director of the Institute of Information Systems and Digital Business at the University of St. Gallen, Switzerland. He is also Director at the Research Center for IS Design at the University of Kassel, Germany. His work covers digital and AI-driven value creation.