# Trust and Decision-Making with Explainable AI in Immersive Technologies: A Systematic Literature Review

HILLMER CHONA*, The Pennsylvania State University, USA
YIHAO ZHOU, The Pennsylvania State University, USA
PING XU, The Pennsylvania State University, USA
JEFFREY SAMUEL SCHULMAN JR., The Pennsylvania State University, USA
TING YU WU, The Pennsylvania State University, USA
CHENGLIN WENG, The Pennsylvania State University, USA
SIYU WU, The Pennsylvania State University, USA
CHARAN PUSHPANATHAN PRABAVATHI, The Pennsylvania State University, USA

The integration of Explainable Artificial Intelligence (XAI) into Extended Reality (XR) environments, including Virtual Reality (VR) and Augmented Reality (AR), offers innovative solutions for decision-making by enhancing transparency and trust in AI systems. This systematic literature review examines the role of XAI in immersive environments, focusing on trust calibration, user reliance, and explainability techniques. Key findings highlight the importance of user-centered explanations that are interactive, visual, and adaptive, aligning with user expertise and task-specific requirements. Ethical challenges such as overreliance, anthropomorphism, and bias require robust design frameworks that ensure fairness, accountability, and appropriate trust calibration. Techniques, like shared situational awareness and abnormality detection, were identified as key enablers for fostering balanced human-AI collaboration. By synthesizing recent advancements, this paper identifies areas for future research towards creating dynamic, transparent, and inclusive XAI systems in immersive technologies. The findings underscore the transformative potential of XAI in XR to foster informed decisions and ethical human-AI collaboration.

CCS Concepts: • **Human-centered computing** → **Virtual reality**; **Empirical studies in HCI**; • **Computing methodologies** → **Explainable AI**.

Additional Key Words and Phrases: Artificial Intelligence, Explainability, Interpretability, Trust, XAI, XR, VR, AR

## 1 Introduction

Explainable AI (XAI) refers to methods and techniques in the field of artificial intelligence (AI) that offer insights into the functioning of AI models, thereby making their decisions understandable to users [7, 32]. This aspect of AI is crucial for decision-making in sectors where transparency and trust are paramount [74], such as healthcare [34, 62] and autonomous systems [37, 53]. The core purpose of XAI is to describe an AI model's rationale, characterize its strengths and weaknesses, and predict its behavior in various scenarios [31]. By improving the interpretability of AI systems, XAI

*corresponding author

Authors' Contact Information: Hillmer Chona, The Pennsylvania State University, College of Information Sciences and Technology, University Park, PA, USA, hac5324@psu.edu; Yihao Zhou, The Pennsylvania State University, College of Information Sciences and Technology, University Park, PA, USA, yihaozhou@psu.edu; Ping Xu, The Pennsylvania State University, College of Information Sciences and Technology, University Park, PA, USA, ymx5173@psu.edu; Jeffrey Samuel Schulman Jr., The Pennsylvania State University, College of Information Sciences and Technology, University Park, PA, USA, jeffschulman@psu.edu; Ting Yu Wu, The Pennsylvania State University, College of Information Sciences and Technology, University Park, PA, USA, tjw6302@psu.edu; Chenglin Weng, The Pennsylvania State University, College of Information Sciences and Technology, University Park, PA, USA, cvw5844@psu.edu; Siyu Wu, The Pennsylvania State University, College of Information Sciences and Technology, University Park, PA, USA, sfw5621@psu.edu; Charan Pushpanathan Prabavathi, The Pennsylvania State University, College of Information Sciences and Technology, University Park, PA, USA, cjp6449@psu.edu.

helps stakeholders validate and trust AI-made decisions, ensuring that they align with ethical standards and practical expectations.

The integration of emergent technologies such as XR—encompassing VR [64], and AR [65]—with XAI frameworks presents innovative opportunities for immersive explainable decision-making [75]. These technologies provide a unique platform for visualizing and interacting with AI systems, thereby enabling users to engage with and understand decision-making in a more intuitive and impactful way. VR can create completely controlled environments where every element is an aspect of the model's data and decisions, allowing for deep dives into simulations that mirror real-life scenarios or abstract AI concepts [64]. AR, on the other hand, overlays digital information onto the real world, which can be used to display real-time data and decision paths directly within a user's field of view [65]. The integration enhances the potential for creating a more dynamic form of interaction, where decision-makers can manipulate AI insights and outputs in their physical space, potentially leading to more agile, informed, and understandable decision-making.

However, while the field is promising, it faces challenges. Such systems often require processing large amounts of sensitive (including personally-identifiable) information, in such settings as healthcare or finance, which can raise issues such as privacy leakage and trustworthiness crises, as Scharowski et al. explained in the challenges of building trustworthy AI systems [69]. In addition, as the methods for implementing XAI evolve, such as those described by Jahn et al., which include novel algorithms for generating counterfactual explanations, the question of how to integrate these newly developed XAI algorithms into the current immersive environment remains an open question, and calls for an adaptable and flexible framework for guidance [31].

In this paper, to better understand how to apply explainable AI in an immersive environment, we conducted a systematic literature review (SLR) using the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) framework [72]. Our aim was to systematically analyze the current state of the art regarding how XAI is integrated into XR environments, assess the impact of XAI on trust, reliance, and decision-making, and discuss its implications.

## 2 Methods

To achieve our research objectives, we conducted a systematic literature review. Following the PRISMA guidelines [72], our process involved three main stages: identification, screening, and inclusion. During the identification phase, we designed a review protocol that utilized keywords to guide database searches. In the screening phase, we applied inclusion and exclusion criteria to refine the dataset systematically. The final inclusion stage involved data extraction and full-text review. Figure 1 provides a detailed visual representation of the entire PRISMA workflow.

### 2.1 Scope and Keywords

To define the scope of this review, we included works focusing on XAI systems within immersive environments, such as Virtual Reality (VR), Augmented Reality (AR), and Mixed Reality (MR). We specifically targeted studies that addressed trust, reliance, and decision-making because these are key factors in understanding user acceptance and interaction with XAI systems. Immersive XR environments were chosen as the focus due to their unique ability to integrate explainability techniques within dynamic, interactive, and spatially immersive contexts, which can influence user perception and decision-making processes. We developed a series of keywords to reflect these criteria. The search string was structured as shown in Table 1.

We conducted the search using Google Scholar to ensure wide coverage of peer-reviewed journal and conference papers. A Python script was developed to scrape the metadata of the papers. Initially, 315 papers were identified using

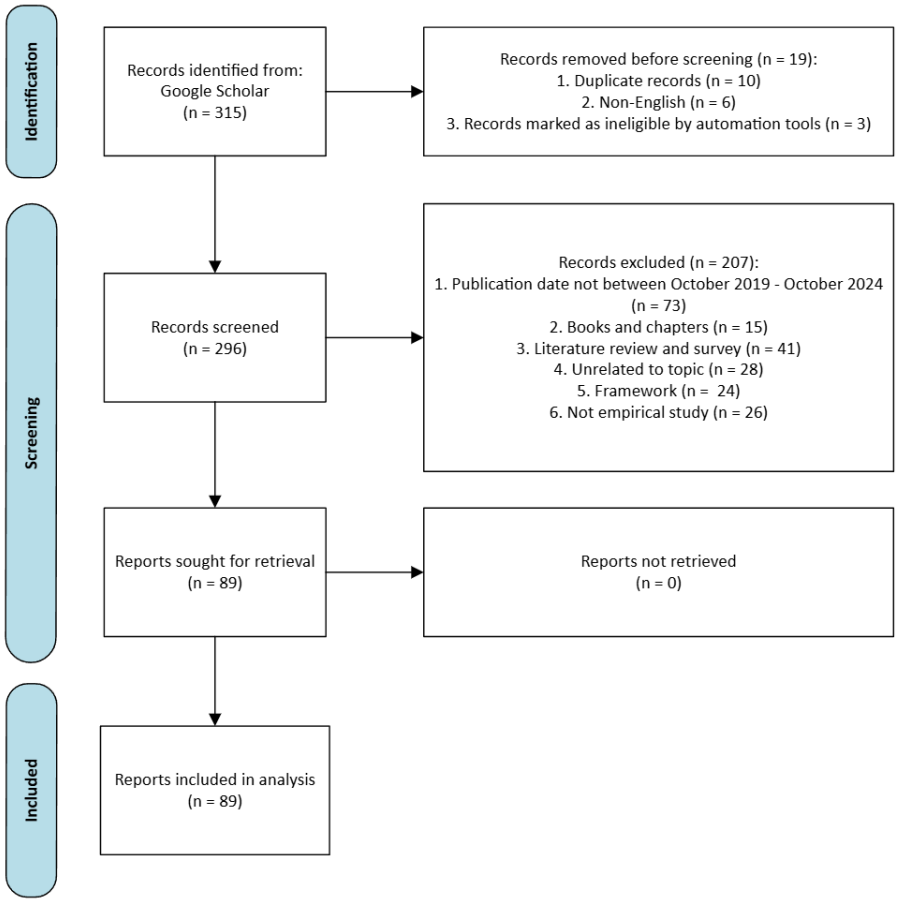


Fig. 1. PRISMA workflow diagram.

the search string. Then, we screened for duplicates and managed the retrieved papers using Zotero Reference Manager, ensuring that the resulting dataset was refined for further screening. During this stage, papers written in non-English were also excluded. After identification, 296 papers remained.

### 2.2 Screening and Inclusion

To systematically screen identified articles, we applied the exclusion criteria as follows:

- Not empirical studies;
- Books or chapters;
- Literature reviews or surveys;
- Frameworks;
- Not related to XAI and immersive environments;
- Not between October 2019 - October 2024;

| Category | Search String |
|---|---|
| **General XAI Terms** | ("EXPLAINABLE AI" OR "XAI" OR "EXPLAINABLE ARTIFICIAL INTELLIGENCE" OR "EXPLAINABLE INTELLIGENT AGENTS" OR "autonomous systems" OR "Interpretable Machine Learning" OR "Interpretable ML" OR "interpretable Artificial Intelligence" OR "Interpretable AI") |
| **Trust and Reliance** | ("APPROPRIATE TRUST" OR "APPROPRIATE RELIANCE") |
| **Targeted Task** | ("DECISION-MAKING" OR "DECISION MAKING") |
| **Immersive Context** | ("VIRTUAL" OR "IMMERSIVE ENVIRONMENT" OR "VIRTUAL REALITY" OR "AUGMENTED REALITY" OR "MIXED REALITY" OR "XR" OR "EXTENDED REALITY") |
| **Database Filter** | source:ACM OR source:IEEE OR source:apa OR source:aaai OR source:ijcai OR source:springer OR source:sciencedirect |

Table 1. The Search String in Google Scholar

The papers we included are: empirical studies, journal papers or conference proceedings, topics related to XAI and XR, publication date between October 2019 to October 2024.

We restricted the search to a time window limited to the prior five years to ensure the inclusion of the most recent research trends. After applying the time filter, the remaining papers were manually reviewed to ensure relevance to our research objectives. Papers that lacked XAI integration, such as those focusing on visualizations without AI methods, or failed to address trust, reliance, or decision-making were excluded. Overall, the screening process resulted in 207 papers being removed, leaving a final dataset of 89 relevant papers (see Figure 1).

### 2.3 Data Extraction and Analysis

For the 89 papers, we used a shared spreadsheet to organize their metadata. To identify recurring themes, our team evenly divided the data set among members. Each member summarized their assigned papers and identified five relevant keywords, such as "Cognitive Trust" and "Human-Agent Teams". We then held group discussions to share the findings and collectively determined five recurring themes, which are presented in the Findings section. Following this, each paper was assigned one to two themes, and all team members were assigned a portion of the work conducting an in-depth reading and analysis of a specific theme. Finally, we integrated all analyses to construct a cohesive Findings section that reports the identified themes and insights from the dataset.

### 2.4 Use of Generative AI

We utilized Generative AI (i.e., ChatGPT-4) to generate initial summaries, section headings, and paper drafting. These AI-generated outcomes were later manually refined to ensure that they accurately reflected the findings and objectives of the study.

## 3 Findings

In this section, we present our findings derived from a comprehensive analysis of the retrieved papers. From the existing body of research, we identified and synthesized five key themes that address different aspects of explainable AI (XAI) in immersive environments: Trust-Building Mechanisms in Immersive Decision-Making, Explainability and

Interpretability Tailored for XR Interfaces, Ethical Considerations and Challenges in XR-Driven Decision-Making, Perspectives on Human-AI Team Collaboration, and Applications of XAI in XR Environments. These themes encapsulate the essential dimensions of XAI's role in XR systems, highlighting the mechanisms that foster trust, the importance of adaptive explanations, ethical complexities, collaborative dynamics, and the practical implementations of XAI across diverse immersive contexts.

### 3.1 Trust-Building Mechanisms in Immersive Decision-Making

*3.1.1 Foundations and Theoretical Advances in Trust-Building.* In immersive environments, users engage with AI-driven decision support systems in ways that differ significantly from traditional desktop or mobile interfaces. By directly embedding AI explanations and data representations into a user's physical or simulated surroundings, these technologies can foster a richer sense of presence and situational understanding, thus influencing how trust is formed and maintained.

Decision-making is an essential process in a wide range of domains. Accounting professionals rely on automation for complex financial analyses [2], while manufacturing teams depend on AI-guided robotics to manage high-risk tasks [11, 52]. In healthcare, immersive platforms have the potential to integrate patient data, diagnostic tools, and AI decision aids into a spatially rich environment, offering clinicians enhanced interpretability of recommendations [32, 34, 62, 73]. Similarly, cybersecurity analysts can visualize threat intelligence within their operational context, improving their understanding and calibration of trust in automated detection systems [7, 23].

Early work on trust in intelligent automation emphasized system reliability and demonstrated that when automated decision support systems (DSS) are highly accurate, users tend to give them more trust [49]. Over time, researchers realized that trust cannot be fully explained by accuracy alone. Instead, trust is context-sensitive and influenced by users' professional expertise, personal experiences, and the social ecosystem in which the technology is deployed [2]. For example, Juvina et al. developed a unified computational model based on the Adaptive Control of Thought-Rational (ACT-R [1]) cognitive architecture to capture the learning process behind trust—building factors such as trait propensity and repeated interactions [32]. Their work provides a conceptual backbone and merges findings from interpersonal trust in psychology and human-machine trust research in computer science. This transition from simple reliability models to broader socio-cognitive frameworks reveals the complexity of trust and sets the stage for more integrative theories.

Following this trend, later theoretical perspectives move beyond dyadic user-technology relationships. Cameron et al. utilize the Social Triad Model to situate trust within a larger sociotechnical ecosystem by introducing the "Deployer" (i.e., the organization or individual behind the robot or AI system) as a key influence on user trust [11]. Such frameworks emphasize that trust is not just "engineered" through technical features but also emerges from the broader social and organizational context, including interpersonal relationships, communication practices, and cultural norms.

Recent studies have deepened our understanding of how trust manifests and evolves. For example, He et al. find that a debugging intervention aimed at helping users critically evaluate AI outputs may sometimes backfire and reduce reliance due to users' over-fixation on system weaknesses [27]. Similarly, Lochner et al. noticed that while automation with high accuracy can bolster trust, it may inadvertently reduce users' vigilance in error detection, creating a "Catch-22" [2] situation [49]. Such findings strengthen the argument that trust building is not only about making systems more accurate or transparent but also linked with timing, user skill levels, and contextual cues.

[1] See https://acs.ist.psu.edu/ for software and documents
[2] A catch-22 is a paradoxical situation from which an individual cannot escape because of contradictory rules or limitations. The term was coined by Joseph Heller, who used it in his 1961 novel Catch-22.

*3.1.2 Contextual, Cognitive, and Affective Drivers of Trust-Building.* Within immersive environments, trust-building extends beyond static dashboards. Users can spatially navigate information, engage with 3D models, manipulate visualized data streams, and observe how AI logic changes under different scenarios. Proactive immersive assistants may adapt their guidance to user expertise, task complexity, and cognitive load [44]. In a VR training module for healthcare decision-making, for example, subtle AR/VR cues can explain the rationale behind a diagnostic suggestion and help clinicians calibrate trust in the AI's recommendations while maintaining professional judgment [34, 62]. Likewise, within autonomous systems deployed in domains like robotics or unmanned vehicles, immersive explanations can show how AI decisions unfold in real-time navigation tasks [37, 53].

Given the affective and narrative nature of immersive media, emotional responses and the way scenarios are framed fundamentally shape whether users perceive AI as a trusted teammate rather than a mere computational tool. Framing human-machine teaming scenarios through narrative lenses helps users in understanding agent roles, facilitating more nuanced mental models and fostering trust [52]. Emotions experienced during interactions, such as anxiety, hostility, and positive affect, directly influence trust in autonomous agents [22]. Ferronato et al. find that higher risk propensity correlates with a greater likelihood of trusting automation, but only if users also perceive transparency and reliability [23]. Similarly, value alignment between users and agents enhances trust, indicating that perceived moral or philosophical compatibility matters [57].

Another unignorable part is social factor. Users' perception of the technology's deployer can shape their openness to trusting robots or AI systems [11], while exposure to conversational agents over time can humanize these systems, leading to higher trust and acceptance [8]. In low-stakes scenarios like VR shopping assistants, even minor trust violations can provoke feelings of vulnerability that affect long-term trust formation [26]. These relational and affective dimensions show that trust is not only rational or performance-based but also social, emotional, and value-laden.

Alongside the factors mentioned above, explanations, which is the key of XAI, remain a cornerstone of trust-building. Multiple studies demonstrate that well-designed explanations can reduce overreliance on AI [74] and improve user acceptance of AI outputs, particularly when paired with interpretive aids such as attention heatmaps and robot gestures [53]. Providing probabilities for classification tasks can sometimes enhance trust in correct classifications, although it may not always help users detect incorrect ones [19]. Similarly, traceability—making the decision-making processes of smart charging agents more transparent—increases reported trust and subjective predictability, but not necessarily actual user understanding [5]. These cases reveal a persistent challenge: explanations can reassure users without truly enhancing their discernment, risking unwarranted trust.

*3.1.3 Future Directions and Key Challenges in Trust-Building.* To tackle the challenges, counterfactual explanations and abnormality spotlighting can direct users to critical decision features, encouraging deeper understanding and trust grounded in insight rather than blind faith [31]. Aligning human and AI situation awareness also fosters better collaboration: when users understand the reasoning environment and the AI's informational context, they can more effectively calibrate trust, especially in high-stakes or unreliable scenarios [73, 85]. These findings resonate with earlier arguments that trust-building explanations must be action-specific, context-aware, and cognitively accessible.

Methodologically, trust research has expanded from questionnaires and scenario-based experiments to more nuanced and real-time assessment techniques. Physiological measures such as electroencephalogram (EEG) and galvonic skin response (GSR) can measure body response in various trust states, enabling dynamic trust calibration and responsive system design [1]. Studies show that concise trust scales, multi-dimensional measures, and even single-item assessments reliably detect trust fluctuations following system errors [59]. Longitudinal and sequential interaction studies document

trust development over time, confirming that repeated exposure, consistent performance, and tailored explanations gradually stabilize trust [33]. Simulation and formal models, including Partially Observable Markov Decision Process (POMDP)-based frameworks, guide policies for trust-aware assistance-seeking, optimizing when and how systems should request human intervention or provide just-in-time explanations [54].

Drawing on research outputs, the researchers have proposed some design implications. To avoid algorithm aversion and deception, it is necessary to introduce transparency measures that do not mask competence differences or encourage misplaced trust [7, 24]. In addition, risk profiles and domain contexts are deeply important. For example, users respond differently to robots with distinct risk approaches in hazardous environments [9], and risk levels shape trust attitudes more acutely when the stakes are high [29].

Despite progress, unresolved questions remain. One is how to ensure that explanations and transparency do not inadvertently cultivate false confidence or overreliance. Some explanation types can be "misleading," inadvertently boosting trust even when a system is faulty [30]. Similarly, untrustworthy AI can exploit transparency cues to appear more reliable than it is [7]. Addressing these ethical and safety risks involves designing systems that encourage a balance between user engagement, skepticism, and verification.

Another avenue for future research is exploring trust dimensions in cultural, moral, and social settings. While value alignment boosts trust [57], how do differing cultural backgrounds, professional norms, or ethical frameworks affect these dynamics? Studies will need to integrate insights from psychology, anthropology, and ethics to develop frameworks that ensure trust is earned through genuine competence, honesty, and alignment with user values rather than superficial human-likeness or emotional manipulation [26, 56].

The continuous refinement of trust measurement and sensing techniques also promises progress. Integrating psychophysiological signals, gaze behavior tracking [12], and narrative-driven design [52] can support adaptive systems that adjust communication strategies based on real-time trust indicators. With improved sensing and modeling capabilities, future systems can proactively calibrate trust, stepping in with explanations or seeking human assistance precisely when needed.

### 3.2 Explainability and Interpretability Tailored for XR Interfaces

Explainability and interpretability have emerged as cornerstone principles in the design of user-centric interfaces, playing a pivotal role in fostering trust and usability in human-computer interaction (HCI). As AI systems become increasingly complex and integrated into dynamic environments, ensuring that users can comprehend the rationale behind AI decisions is paramount. This is particularly critical in domains such as healthcare, education, and robotics, where decision-making often involves high stakes, uncertainty, and diverse user needs. Addressing these challenges requires a shift from static, one-size-fits-all explanations to more adaptive, transparent, and interactive designs tailored to the specific contexts in which AI systems operate.

In recent years, the development of XAI has focused on creating interfaces that enhance users' ability to understand, trust, and interact effectively with AI systems. This section synthesizes insights from a range of studies exploring interactive, visual, and socio-culturally sensitive approaches to XAI. It delves into the impact of participatory learning methods, the design of dynamic and multi-modal explanations, and the calibration of trust through personalized systems. We aim to illuminate best practices and identify gaps in the current landscape of explainability in AI-enabled interfaces by examining the interplay between cognitive diversity, uncertainty management, and real-time adaptability.

*3.2.1 Interactive and User-Centric Explanations.* Interactive systems significantly improve user understanding and engagement by allowing dynamic exploration of AI behavior. Hence, interactive and user-centric explanations are central to creating explainable AI interfaces in XR environments. Studies emphasize the value of enabling users to explore AI systems dynamically and receive explanations tailored to their needs. Crisan et al. introduced interactive model cards that empower users to explore AI models through a human-centered approach, fostering deeper transparency and trust [18]. Similarly, Qian and Unhelkar demonstrated how interactive robot policy explanations in integrated virtual and physical environments enhance user comprehension and adaptability [64]

Complementing these findings, Weitz et al. extended this approach by showcasing the effectiveness of participatory machine learning shows in engaging end-users [78]. Their study revealed that participatory learning enhances comprehension of AI systems, enabling users to critically interact with decision-making processes. This participatory approach aligns with tools like ChatrEx, a conversational chatbot interface that tailors explanations to user needs [37], and Gupta et al.'s work on conversational AI in decision-support systems [25], which fosters trust and transparency.

Liu et al. added further depth by exploring interactive explanations in human-AI decision-making [48]. Their research revealed that interactive systems not only improve the perceived usefulness of AI assistance but also highlight critical challenges, such as the potential reinforcement of human biases. This insight emphasizes the importance of designing two-way interaction interfaces that balance engagement with user trust calibration.

*3.2.2 Visual and Multi-Modal Explanation Approaches.* Visual explanations are particularly effective in enhancing interpretability in complex domains. XR's immersive nature enables novel techniques such as layered visualizations, motion-based graphics, and real-time overlays to illustrate AI decisions. For example, Wentzel et al. developed the DITTO framework, which uses layered visualizations to support medical decision-making by illustrating treatment outcomes for cancer patients [79]. Similarly, Chen et al. demonstrated how contrastive visual explanations, which compare alternative decisions, improve user understanding and confidence in AI-driven robotic planning [16]. Schrills and Franke explored the impact of color-coded visual explanations, showing that intuitive designs improve trust and observability in decision-making interfaces [71].

Further advancing this approach, Kim et al. introduced generative machine learning methods to create motion-based visual explanations suitable for dynamic environments like XR interfaces [38]. Adding to these insights, He et al. examined how analogies combined with stated system accuracy help explain AI decisions [28]. Their findings reveal that while analogies improve users' understanding of accuracy, they may not necessarily lead to increased reliance on the system. This underscores the complexity of designing visual explanations that balance intelligibility with trust calibration. Roundtree et al. investigated human-collective visualization transparency in systems involving hundreds of entities. They found that abstract visualizations, which simplify data representations, improve operator performance and situational awareness by reducing cognitive overload [67]. This aligns with broader findings that visual transparency mechanisms in large-scale systems must strike a balance between detail and abstraction to ensure user comprehension.

*3.2.3 Trust Calibration Through Explanation Design.* Trust calibration ensures users appropriately rely on AI systems while avoiding overtrust or skepticism. Mehrotra et al. found that value similarity between users and AI enhances trust, emphasizing the importance of aligning AI behavior with user expectations [57]. Similarly, Böckle et al. also underscored the need for personalized explanations, finding that user-specific traits, such as openness to experience, significantly influence trust levels [10].

Building on this, Wang and Yin identified three key principles for effective AI-assisted decision-making explanations: improving understanding, recognizing uncertainty, and calibrating trust [76] Their study highlighted that feature-based explanations are particularly effective for expert users in high-stakes contexts, where precision and clarity are essential. Additionally, Alizadeh et al. studied explanatory dialogues in voice assistants, demonstrating that transparent communication fosters long-term trust and accountability [4].

*3.2.4 Cognitive and Socio-Cultural Factors.* Effective explanation design must account for cognitive and socio-cultural diversity to ensure explainability and inclusivity. Kopecka et al. examined preferences for AI explanations based on cognitive styles, finding that holistic thinkers prefer goal-based explanations while analytic thinkers favor belief-based ones [41]. These findings highlight the importance of designing explanations that accommodate different cognitive styles.

Cultural sensitivity is equally crucial. Okolo et al. studied community health workers in rural India, finding that culturally tailored explanations significantly improve comprehension and trust [62]. In addition, He et al. emphasized the relevance of analogical reasoning to bridge cognitive gaps, providing intuitive explanations that align with users' common sense knowledge [28]. These insights underscore the need for adaptive designs in XR that cater to diverse cognitive and cultural backgrounds.

*3.2.5 Handling Uncertainty in Complex Systems.* Managing uncertainty in AI systems is critical for ensuring user confidence and interpretability. Cau et al. analyzed the impact of logical reasoning styles on decision-making under uncertainty, advocating for explanations that balance simplicity and depth [15]. Nguyen et al. explored automated rationale generation in AI systems, emphasizing the role of clear and concise human-language explanations in fostering user understanding [60].

Lawless et al. emphasized the importance of real-time adaptability in high-stakes environments [46]. Their study highlighted the benefits of interactive decision-support systems that leverage large language models to provide real-time explanations, enabling users to address uncertainty dynamically.

*3.2.6 Applications in Critical Domains.* Explainability is pivotal in domains like healthcare, robotics, and education. Wentzel et al. and Kaptein et al. demonstrated how visual and interactive explanations enhance decision-making in medical settings, particularly in treatment planning and long-term health support for children [34, 79]. Detecting swarm degradation was addressed by Capiola et al., who proposed measuring human and machine performance to offer actionable feedback in robotic systems [14]. Babel et al. emphasized trust-building through verbal and non-verbal cues in social robotics [6]. Alizadeh et al. and Kim et al. highlighted the importance of participatory and generative AI-assisted tools in fostering critical engagement and AI literacy among users [4, 39].

*3.2.7 Balancing Simplicity and Depth.* Explanations must provide sufficient detail for informed decision-making while avoiding information overload. Interactive systems, such as ChatrEx and interactive model cards, demonstrate the potential for balancing these needs.

*3.2.8 Adaptability and Context-Sensitivity.* Dynamic interfaces, such as those used in XR or robotics, require real-time adaptability to user interactions and environmental changes. Future research should explore adaptive explanation systems that personalize content based on user behavior.

*3.2.9 Inclusivity and Diversity.* Catering to diverse user populations, including varying cognitive styles, cultural backgrounds, and personality traits, remains a priority. Analogical reasoning and culturally sensitive designs offer promising pathways for achieving inclusivity.

The reviewed studies collectively underscore the transformative potential of explainability and interpretability in shaping user-centric AI systems. Across domains such as healthcare, robotics, and education, explainable AI has proven essential in enabling users to make informed decisions, calibrate trust appropriately, and engage with technology confidently. The research highlights how interactive tools, such as participatory learning models and conversational systems, empower users by making AI processes transparent and relatable. Similarly, visual and multi-modal explanations, including contrastive and generative approaches, have shown promise in enhancing user comprehension while addressing the complexities of dynamic and high-stakes environments.

Yet, significant challenges remain in the quest for more inclusive, adaptive, and context-sensitive explainability solutions. Balancing the need for simplicity and depth is an ongoing challenge, as overly detailed explanations risk overwhelming users while overly simplified ones may obscure critical insights. Furthermore, the diversity of users—spanning cognitive styles, cultural backgrounds, and varying levels of AI literacy—demands innovative approaches, such as analogical reasoning and culturally tailored designs, to make AI explanations more accessible and equitable.

### 3.3 Ethical Considerations and Challenges in XR-Driven Decision-Making

Integrating XAI into immersive technologies, such as AR, VR, and XR, presents significant opportunities and challenges. XR is increasingly used in high-stakes domains, such as training, healthcare, and decision-making, where AI systems influence user perceptions, choices, and outcomes. As AI adoption grows, ethical decision-making has emerged as a critical concern, emphasizing the need for transparency, accountability, and fairness in AI systems.

A central element of ethical AI design is trust calibration, in which users develop confidence in AI recommendations while avoiding overreliance or unwarranted skepticism. Humer et al. explored XAI methods like Grad-CAM and nearest-neighbor examples, demonstrating their impact on user trust and decision accuracy in specific tasks [30]. Srivastava et al. emphasized the importance of shared situational awareness (SSA), showing how human-AI alignment in contextual understanding fosters ethical decision-making [73].

The immersive nature of the XR amplifies these dynamics. XR blends reality with virtual simulations, creating heightened user presence and engagement, which can magnify trust or skepticism toward AI systems. As Lopez et al. argue, these environments demand ethical frameworks that ensure that AI systems remain accountable and trustworthy in collaborative human-AI interactions [50]. Moreover, the potential for anthropomorphism, as discussed by Natarajan et al., introduces further ethical complexity and requires careful attention to user-AI dynamics [58].

*3.3.1 Trust Calibration Through XAI Methods.* Trust calibration ensures that users develop an appropriate confidence level in AI recommendations. Overtrust leads to blind acceptance of erroneous outputs [30, 58, 70, 73], whereas distrust may cause users to disregard accurate and valuable insights [20, 30, 56, 70, 73]. Explanation techniques like Grad-CAM and nearest-neighbor examples influence trust by making AI reasoning more transparent and understandable [30]. In XR contexts, such methods align user expectations with the system's capabilities, particularly in high-stakes applications.

*3.3.2 Transparency and Ethical Outcomes.* Transparency in AI systems enables users to understand the reasoning behind decisions and fosters trust and ethical engagement[30, 73]. However, transparency in XR environments presents unique challenges, owing to the complexity and dynamic nature of immersive interactions. The SSA framework enhances transparency by aligning the situational analysis of AI with the user's understanding of the task context. This

alignment reduces misunderstandings, promotes trust, and minimizes ethical risks such as reliance on flawed or biased AI recommendations. Humer et al. compared various XAI techniques and found that task-specific explanations improve user comprehension of AI decisions; however, their effectiveness depends on context and user expertise. For example, novice users may benefit from simple visual overlays, highlighting critical decision points, whereas experts may require detailed data visualization [30]. Kopecka et al. extended this concept by examining how cognitive styles and cultural factors influence preferences for AI explanations, emphasizing the need for adaptable user-centric designs [41].

*3.3.3 Ethical Challenges in XR-Specific Contexts.* The immersive and emotionally engaging nature of XR introduces additional ethical challenges to XAI. Users often over-rely on AI systems because of the perceived authority or anthropomorphic design of virtual agents. Anthropomorphic AI agents can foster trust, but they also risk creating unwarranted confidence in the system's capabilities [58]. This underscores the importance of transparency mechanisms that counterbalance the persuasive effects of anthropomorphic cues. In XR applications, stakes are higher because misaligned decisions can have immediate and impactful consequences. Srivastava et al.'s research on SSA demonstrates that ensuring alignment between AI and user situational understanding mitigates these risks by keeping users attentive and engaged in decision-making[73]. Jahn proposed leveraging abnormality detection as a user-centric XAI method to address bias and support appropriate trust calibration in complex scenarios[31]. Eckhardt's "Garbage In, Garbage Out" principle further underscores the importance of robust data practices in XR systems[20]. Bias in AI systems can be amplified in immersive environments, leading to ethical concerns such as reinforcing stereotypes or misleading users. Transparent context-aware explanations tailored to the XR environment can help mitigate these risks.

*3.3.4 Overtrust and Distrust in Immersive Systems.* In XR applications, managing user trust is critical for ensuring balanced interactions with AI systems. Overtrust, where users rely excessively on AI outputs, can lead to harmful outcomes, particularly in high-stakes scenarios, such as healthcare training or autonomous operations. Conversely, distrust can render AI systems ineffective as users dismiss valid recommendations. Humer et al. identified overtrust as a significant challenge, emphasizing how poorly designed explanations can reinforce unwarranted confidence in AI systems[30]. For example, visual explanation techniques (e.g. saliency maps, Grad-CAM) may inadvertently highlight irrelevant areas, misleading users to accept incorrect decisions. In immersive XR environments, a heightened sense of presence can amplify these effects, making context-sensitive trust calibration even more urgent. Fostering SSA between users and AI systems reduces over-trust by aligning user understanding with AI reasoning [73]. Individual differences in trust propensity also affect user interactions with AI, as Matthews et al. found in their study of autonomous robots [56]. This variability complicates the development of one-size-fits-all solutions, necessitating adaptive mechanisms to address the user-specific trust dynamics in XR.

*3.3.5 Bias and Cognitive Limitations.* AI systems reflect the biases inherent in their training data, which can become more pronounced in immersive XR environments. "Garbage In, Garbage Out' highlights how biases in data can propagate through AI models, leading to unethical outcomes. For instance, biased AI systems in XR may reinforce stereotypes or generate discriminatory results in simulations designed for diverse user groups[20]. Cognitive overload exacerbates these problems. XR environments often immerse users in highly interactive and visually stimulating scenarios that can overwhelm their decision-making capacity. Such an overload degrades SSA, impairing users' ability to critically assess AI outputs and make informed ethical decisions [73]. This underscores the importance of designing XAI systems that simplify complex information and present it in user-friendly formats.

*3.3.6 Ethical Implications of Anthropomorphism.* Anthropomorphic AI agents in XR add another layer of complexity to ethical decision-making. While anthropomorphism can enhance user engagement and trust, it also risks creating unrealistic expectations regarding AI capabilities. Anthropomorphic features often lead users to perceive AI systems as more competent or empathetic than they truly are, which can result in overreliance[58]. Ethical concerns surrounding anthropomorphism extend beyond trust calibration. Users may feel deceived or manipulated if they later discover that an AI's human-like traits are designed primarily to influence their behavior. This raises questions regarding user autonomy and informed consent, which are critical components of ethical AI design.

*3.3.7 Task-Specific Risks in Immersive Environments.* XR applications often involve high-stakes or emotionally charged scenarios where task-specific risks become significant. Humer et al.'s research revealed that the effectiveness of XAI methods varies across tasks, with some explanations more suitable for high-risk applications. For example, users of surgical training simulations may require precise, data-driven explanations, whereas entertainment-focused XR systems can prioritize simpler and less technical outputs. Failing to tailor XAI systems to specific tasks increases the likelihood of ethical breaches. Mitigating these risks with SSA, ensuring that users remain actively engaged in the decision-making process, allows for risk management and safety considerations [73]. Moreover, Jahn proposed using abnormality detection in XAI to address edge cases and enhance trust calibration in complex tasks [31].

### 3.4 Perspectives in Human-AI Team Collaboration

*3.4.1 Enhancing Human-Agent Teams Communication.* Human expect intelligent agents to produce proactive, effective, concrete, and human-like communications [84]; moreover, they prefer bidirectional interactions [45]. To enhance these interactions, researchers have explored the use of techniques, such as inner-speech [63] and Large Language Model [81], to provide agents with more natural communication skills that can increase human acceptance.

*3.4.2 Facilitating Interactions in Human-AI Teams Through Mental Models.* Mental models used to characterize both humans and agents reveal that humans tend to overestimate the performance of agents [36]. To mitigate this negative trend, studies have explored the application of Theory of Mind (ToM), which helps humans better understand and predict the behavior of agents. Research shows that incorporating ToM can reduce human blindness to agents' responses [66]. Furthermore, studies suggest that ToM encourages critical thinking, leading to better informed and adequate decision making [47]. Additionally, to promote transparency, Lu et al [51] proposed a model designed to generate comprehensive outcomes, ensuring that agents' reasoning are clearer and more accessible to humans.

*3.4.3 Influences of Model, Task, and Human Factors on Team Decision-Making.* Artificial Intelligence (AI) agents have become valuable tools in supporting decision-making processes. Understanding how the model, task, and human factors influence these processes is essential to foster a strong human-AI relationship. Studies have demonstrated that providing users with the frequency of model uncertainties can help reduce confirmation bias [13], that supporting complex processes with AI agents improves efficiency [68], and that exposing humans to stressors, such as time pressure or social expectations, does not significantly impact their reliance on AI agents during decision-making [86].

*3.4.4 Building and Measuring Trust in Human-AI Teams.* Trust is essential for a team's success. Since humans understand that AI can make mistakes, adding capabilities to AI agents, such as communicating uncertainty [43] apologizing when failing [42], and presenting plans for improvements [50] (Lopez et al, 2023), can help build trust. Trust in AI agents is also influenced by reputational factors, for example, the process that agents support [40]. Furthermore, Marble et al [55] found that, regardless of the performance of agents, humans tend to trust other humans more than AI agents. In

efforts to find measures to monitor trust, Eloy et al [21] identified potential correlations between neural processes and human trust.

### 3.5 Applications of XAI in XR Environments

Research related to the applications of XAI in XR environments collectively reveals a sophisticated exploration of human-machine interactions in diverse domains, demonstrating that trust is a nuanced, context-dependent construct fundamentally shaped by system design, transparency and communication strategies [17, 80, 83]. These studies span critical areas including robotics, healthcare, augmented reality, decision support systems, and educational technologies, consistently highlighting the importance of understanding human cognitive processes, emotional responses, and contextual expectations when developing intelligent systems. The broader implications of these studies extend well beyond immediate technological applications. They represent a fundamental reimagining of how we design and conceptualize intelligent systems, shifting from a purely functional approach to one that prioritizes human experience and collaborative potential.

At the heart of these investigations lies a complex understanding of trust as a dynamic, context-dependent phenomenon. Researchers discovered that technological effectiveness isn't solely determined by technical performance, but by sophisticated factors like communication transparency, adaptive interaction, and the ability to align with human cognitive and emotional expectations [3, 75]. For instance, anthropomorphic design does not universally improve trust, with functionality and reliability often mattering more than human-like appearance [80, 83], while natural language interactions and self-assessment mechanisms can significantly improve human-machine collaboration [82]. The research indicated that the delicate balance required in providing technological assistance demonstrates that explanation strategies and interactions must be carefully tailored to specific user populations and interaction contexts [35]; and transparent communication, self-assessment mechanisms, and adaptive interfaces are crucial in building and maintaining trust. The broader implications of these studies extend well beyond immediate technological applications. They represent a fundamental reimagining of how we design and conceptualize intelligent systems, shifting from a purely functional approach to one that prioritizes human experience and collaborative potential [61, 75]. The research [77] highlights the importance of transparency, usability, and relatable analogies in bridging the knowledge gap between AI developers and everyday users. Whether examining human-robot collaboration in high-stakes search-and-rescue scenarios, developing AI-assisted writing tools for students, or exploring trust dynamics in navigation robots, these studies emphasize the need for human-centered design that balances technological capabilities with user understanding, emotional engagement, and cognitive load management.

By emphasizing context, transparency, and adaptive communication, these research findings provide a critical roadmap for developing technologies that do not merely perform tasks, but genuinely understand and respond to human psychological and cognitive needs. As AI and robotic systems become increasingly sophisticated, this research offers invaluable insights into creating technologies that can truly integrate with human workflows, decision-making processes, and emotional landscapes.

## 4 Limitations and Future Work

While this systematic literature review aims to provide a comprehensive synthesis of research at the intersection of Explainable AI (XAI), Extended Reality (XR), and decision-making, certain limitations must be acknowledged:

- **Search Scope and Source Limitation:** The review relies solely on Google Scholar as the database for article retrieval. While Google Scholar provides a broad range of literature, it may not cover all relevant studies indexed in other specialized databases such as IEEE Xplore, PubMed, or ACM Digital Library. This could result in the exclusion of potentially significant research.
- **Language Bias:** Only studies published in English were included, potentially omitting relevant research published in other languages, particularly from non-English-speaking regions, where XR and AI research is gaining traction.
- **Publication Bias:** The inclusion criteria favored peer-reviewed journal articles and conference papers, which may have excluded valuable insights from gray literature, such as technical reports, theses, or industry white papers.
- **Publication Date Bias:** This study considered only publications from the last five years. This strategy may have excluded relevant and foundational work that could have contributed significantly to understanding the evolution of the intersection of XAI and VR technologies.
- **Evolving Nature of the Field:** Both XAI and XR are rapidly evolving fields. Studies published after the conclusion of this review process may introduce new findings or perspectives that could significantly influence result interpretation.
- **Variability in Methodological Rigor:** The included studies exhibit variability in methodological rigor, study designs, and evaluation criteria, making it challenging to generalize findings across diverse contexts and applications.
- **Limited Contextual Focus:** The review focuses on the intersection of XAI, XR, and decision-making. While this scope ensures relevance to the research questions, it may have excluded studies exploring broader implications of XAI or XR in unrelated domains.
- **Potential Bias in Screening and Selection:** Despite using a systematic approach and PRISMA guidelines, the screening and selection process may be influenced by subjective judgments during the inclusion/exclusion of studies.
- **Lack of Empirical Generalizability:** Many of the included studies are based on experimental setups or simulated environments, limiting the applicability of their findings to real-world scenarios.

In future work, we plan to expand our article retrieval by utilizing multiple academic databases, thereby ensuring a more comprehensive and varied selection of sources. We also intend to broaden the time frame of our search aiming to find foundational resources that may offer valuable insights, but were excluded in the current study. In addition, our goal is to refine our search criteria to reduce subjective judgments and enhance the objectivity of our results. By implementing these strategies, we expect to provide a more thorough and diverse synthesis of research at the intersection of XAI, XR, and decision-making.

## 5 Conclusion

The integration of Explainable AI (XAI) into Extended Reality (XR) environments holds transformative potential for improving decision-making, trust, and user comprehension in immersive systems. Through this systematic literature review, we synthesized findings from multiple studies, identifying recurring themes such as trust-building mechanisms, explainability tailored for XR interfaces, ethical considerations, human-AI team collaboration, and practical applications across critical domains. These themes provide a comprehensive understanding of the complex interplay between AI, XR, and human cognition, emphasizing the importance of user-centric design, and ethically grounded approaches in designing explainable systems.

Trust-building in XR environments emerges as a critical factor for meaningful human-AI collaboration. Trust itself relates dynamically through extended interactions, much like humans interacting with others. This evolution demands observation and analysis over time. Building trust between users and robotic systems is rooted in understanding and responding to user needs at multiple levels. By gathering insights from psychology, computer science, and social science, developers can better understand human trust formation processes and design systems that align technical capabilities with user expectations. This interdisciplinary approach fosters the creation of advanced, adaptable interaction patterns and trustworthy systems. When developers deeply understand user behaviors and needs, they can design system architectures that build confidence, align with user goals, and foster reliable partnerships. Achieving truly intelligent systems capable of earning and maintaining user trust is a complex, long-term challenge. Future research should emphasize comparative studies that evaluate trust formation across different user groups, including non-users and expert subgroups, to identify diverse perspectives and optimize trust-building mechanisms.

Ethical concerns such as trust calibration, bias, transparency, and anthropomorphism require careful attention in the design of XAI systems. Effective trust calibration is critical to prevent both overreliance and distrust in AI systems, especially in high-stakes domains such as healthcare. Explanation techniques align with user expertise to mitigate these risks. Transparency reduces misunderstandings by aligning AI reasoning with user situational awareness. Bias in AI systems, exacerbated by poor data practices, can propagate unfair outcomes. Anthropomorphic AI agents further complicate ethical considerations, fostering trust but risking unrealistic user expectations. Ensuring fairness requires dynamic, user-centered systems that adapt to task-specific needs and diverse cognitive preferences. Ethical XAI design in immersive systems must prioritize transparency, fairness, and user alignment to foster trust and accountability.

Furthermore, our review underscores the growing importance of human-AI collaboration in immersive settings. The relationship between humans and AI agents has evolved toward partnership interactions, where humans consider these technologies as a team member. Our literature review shows that humans expect these agents to incorporate human-like behaviors, in addition to the traditional accuracy in their predictions. From this review, we highlight key capabilities for AI agents, such as bidirectional effective communication, providing uncertainty levels, and apologizing when inaccurate outputs are presented, as essential factors that can contribute to building transparency and fostering appropriate reliance.

As the integration of XAI into XR environments continues to evolve, the critical role of explainability and interpretability in human-computer interaction becomes increasingly evident. By enabling users to comprehend AI-driven systems' decisions and processes, these techniques foster trust, usability, and meaningful engagement across diverse applications, including healthcare, robotics, and education. Current research highlights the importance of designing interfaces that prioritize user-centric explanations, whether interactive, visual, or multimodal. Across all domains,

effective XAI strategies are those that cater to user needs while balancing simplicity and depth, providing actionable insights while avoiding cognitive overload.

As AI-driven XR systems become more complex, the need for adaptive, transparent, and inclusive explainability methods will continue to grow. Future research must focus on refining real-time adaptable systems that respond dynamically to user interactions and contextual changes. Balancing simplicity and depth in explanations remains critical, as overly simplified interfaces risk obscuring essential details, while overly complex ones may overwhelm users. Furthermore, inclusivity must remain a priority, ensuring that XAI systems account for diverse user needs, cognitive styles, and socio-cultural contexts. As XAI continues to shape immersive XR applications, its success will depend on its ability to foster trust, promote understanding, and enable meaningful collaboration between humans and machines.

## References


[1] Kiran Akash, Wen-Li Hu, Nikhil Jain, and Tahira Reid. 2018. A Classification Model for Sensing Human Trust in Machines Using EEG and GSR. *ACM Transactions on Interactive Intelligent Systems* 8, 3 (2018), 1–20. https://doi.org/10.1145/3132743

[2] Sami Ala-Luopa, Tomas Olsson, and Kaisa Väänänen. 2024. Trusting Intelligent Automation in Expert Work: Accounting Practitioners' Experiences and Perceptions. *Computer Supported Cooperative Work (CSCW)* (2024). https://doi.org/10.1007/s10606-024-09499-6 Publisher: Springer.

[3] Gene M Alarcon, Sarah A Jessup, Scott K Meyers, Dexter Johnson, and Walter D Bennette. 2024. Trustworthiness Perceptions of Machine Learning Algorithms: The Influence of Confidence Intervals. In *2024 IEEE 4th International Conference on Human-Machine Systems (ICHMS)*. IEEE, 1–6.

[4] Fatemeh Alizadeh, Peter Tolmie, Min Lee, and Philipp Wintersberger. 2024. Voice Assistants' Accountability through Explanatory Dialogues. *Proceedings of the 6th ACM International Conference on Conversational Agents* (2024). https://doi.org/10.1145/3640794.3665557

[5] Christian Attig, Tobias Schrills, Michael Gödker, and Philipp Wollstadt. 2023. Enhancing Trust in Smart Charging Agents—The Role of Traceability for Human-Agent Cooperation. In *International Conference on Human-Computer Interaction*. Springer, 255–267. https://doi.org/10.1007/978-3-031-48057-7_19

[6] Friederike Babel, Julia Kraus, Lisa Miller, Matthias Kraus, and Nicolai Wagner. 2021. Small Talk with a Robot? The Impact of Dialog Content, Talk Initiative, and Gaze Behavior of a Social Robot on Trust, Acceptance, and Proximity. *International Journal of Social Robotics* (2021). https://doi.org/10.1007/s12369-020-00730-0

[7] Nikola Banovic, Zhenhui Yang, Anirudh Ramesh, and Anhong Liu. 2023. Being Trustworthy Is Not Enough: How Untrustworthy Artificial Intelligence (AI) Can Deceive the End-Users and Gain Their Trust. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW2 (2023), 3579460. https://doi.org/10.1145/3579460 Publisher: ACM.

[8] Joshua W. Bonny and Karen T. Wynne. 2024. Increasing Human-Likeness and Acceptance of Conversational Autonomy through Experience. *IEEE 4th International Conference on Autonomous Systems* (2024). https://doi.org/10.1109/ICAS.2024.10555683

[9] Tom Bridgwater, Manuel Giuliani, Anna van Maris, Gareth Baker, and Ross J. Anderson. 2020. Examining Profiles for Robotic Risk Assessment: Does a Robot's Approach to Risk Affect User Trust? *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (2020). https://doi.org/10.1145/3319502.3374804

[10] Moritz Böckle, Kwaku Yeboah-Antwi, and Ioannis Kouris. 2021. Can You Trust the Black Box? The Effect of Personality Traits on Trust in AI-Enabled User Interfaces. In *International Conference on Human-Computer Interaction*. Springer, 10–21. https://doi.org/10.1007/978-3-030-77772-2_1

[11] David Cameron, Emily C. Collins, Stevienna de Saille, Iveta Eimontaite, Alice Greenwood, and James Law. 2024. The Social Triad Model: Considering the Deployer in a Novel Approach to Trust in Human–Robot Interaction. *International Journal of Social Robotics* (2024). https://doi.org/10.1007/s12369-023-01048-3

[12] Siyuan Cao and Chi Mei Huang. 2022. Understanding User Reliance on AI in Assisted Decision-Making. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (2022), 3555572. https://doi.org/10.1145/3555572 Publisher: ACM.

[13] Siyuan Cao, Anhong Liu, and Chi Mei Huang. 2024. Designing for Appropriate Reliance: The Roles of AI Uncertainty Presentation, Initial User Decision, and User Demographics in AI-Assisted Decision-Making. *Proceedings of the ACM on Human-Computer Interaction* 8, CSCW1 (2024), 3637318. https://doi.org/10.1145/3637318 Publisher: ACM.

[14] August Capiola, Dexter Johnson, Izz aldin Hamdan, Joseph B. Lyons, and Elizabeth L. Fox. 2023. Detecting Swarm Degradation: Measuring Human and Machine Performance. In *Virtual, Augmented and Mixed Reality*, Jessie Y. C. Chen and Gino Fragomeni (Eds.). Springer Nature Switzerland, Cham, 325–343. https://doi.org/10.1007/978-3-031-35634-6_23

[15] Fabrizio Maria Cau, Hanna Hauptmann, Lucio Davide Spano, and others. 2023. Effects of AI and Logic-Style Explanations on Users' Decisions Under Different Levels of Uncertainty. *ACM Transactions on Interactive Intelligent Systems* 13, 4 (2023), 3588320. https://doi.org/10.1145/3588320 Publisher: ACM.

[16] Shenghui Chen, Kayla Boggess, and Lu Feng. 2020. Towards Transparent Robotic Planning via Contrastive Explanations. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. 4942–4949. https://ieeexplore.ieee.org/document/9341773/

[17] Nicholas Conlon, Daniel Szafir, and Nisar Ahmed. 2022. "I'm Confident This Will End Poorly": Robot Proficiency Self-Assessment in Human-Robot Teaming. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2127–2134.

[18] Anamaria Crisan, Margaret Drouhard, Jesse Vig, and Nazneen Rajani. 2022. Interactive Model Cards: A Human-Centered Approach to Model Documentation. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 427–439. https://doi.org/10.1145/3531146.3533108

[19] Dalai Dos Santos Ribeiro, Gabriel Diniz Junqueira Barbosa, Marisa Do Carmo Silva, Hélio Lopes, and Simone Diniz Junqueira Barbosa. 2021. Exploring the impact of classification probabilities on users' trust in ambiguous instances. In *2021 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*. 1–9. https://doi.org/10.1109/VL/HCC51201.2021.9576291

[20] Stefan Eckhardt, Michael Knaeble, Andreas Bucher, Daniel Staehelin, and others. 2023. "Garbage In, Garbage Out": Mitigating Human Biases in Data Entry by Means of Artificial Intelligence. In *IFIP Conference on Human-Computer Interaction*. 29–44. https://doi.org/10.1007/978-3-031-42286-7_2

[21] Liam Eloy, Christina Spencer, Erin Doherty, and Lisa Hirshfield. 2023. Capturing the Dynamics of Trust and Team Processes in Human-Human-Agent Teams via Multidimensional Neural Recurrence Analyses. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1 (2023), 1–27. https://doi.org/10.1145/3579598

[22] Mohammed Abdul Aziz Fahim, Mahmudul Haque Khan, and Tobias Jensen. 2021. Do Integral Emotions Affect Trust? The Mediating Effect of Emotions on Trust in the Context of Human-Agent Interaction. *Proceedings of the 2021 ACM Conference on Human Factors in Computing Systems* (2021), 523–536. https://doi.org/10.1145/3461778.3461997

[23] Paula Ferronato and Masooda Bashir. 2020. Does the Propensity to Take Risks Influence Human Interactions with Autonomous Systems?. In *Advances in Human Factors in Cybersecurity: Proceedings of AHFE 2020 Virtual Conference*. Springer, 40–50. https://doi.org/10.1007/978-3-030-52581-1_4

[24] Julia Fleiß, Erik Bäck, and Stefan Thalmann. 2024. Mitigating Algorithm Aversion in Recruiting: A Study on Explainable AI for Conversational Agents. *ACM SIGMIS Database: the DATABASE for Advances in Information Systems* 55, 4 (2024), 67–82. https://doi.org/10.1145/3645057.3645062

[25] Ankit Gupta, Deblina Basu, Raghavendra Ghantasala, Shengkai Qiu, and others. 2022. To Trust or Not to Trust: How a Conversational Interface Affects Trust in a Decision Support System. In *Proceedings of the ACM on Human-Computer Interaction*, Vol. 6. ACM, 3485447.3512248. https://doi.org/10.1145/3485447.3512248 Issue: CSCW1.

[26] Glenda Hannibal, Astrid Weiss, and Vicky Charisi. 2021. "The robot may not notice my discomfort" – Examining the Experience of Vulnerability for Trust in Human-Robot Interaction. In *2021 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)*. 704–711. https://doi.org/10.1109/RO-MAN50785.2021.9515513

[27] Guanghui He, Abhinav Bharos, and Ujwal Gadiraju. 2024. To Err Is AI! Debugging as an Intervention to Facilitate Appropriate Reliance on AI Systems. In *Proceedings of the 35th ACM Conference on Hypertext and Social Media*. ACM, 3648188.3675130. https://doi.org/10.1145/3648188.3675130

[28] Guanghui He, Stefan Buijsman, and Ujwal Gadiraju. 2023. How Stated Accuracy of an AI System and Analogies to Explain Accuracy Affect Human Reliance on the System. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1 (2023), 3610067. https://doi.org/10.1145/3610067 Publisher: ACM.

[29] Simon Hoesterey and Lea Onnasch. 2023. The Effect of Risk on Trust Attitude and Trust Behavior in Interaction with Information and Decision Automation. *Cognition, Technology & Work* 25 (2023), 301–317. https://doi.org/10.1007/s10111-022-00718-y

[30] Christian Humer, Andreas Hinterreiter, Benedikt Leichtmann, Philipp Tschandl, and Andreas Holzinger. 2024. Reassuring, Misleading, Debunking: Comparing Effects of XAI Methods on Human Decisions. *ACM Transactions on Interactive Intelligent Systems* 14, 1 (2024), 3665647. https://doi.org/10.1145/3665647 Publisher: ACM.

[31] Tobias Jahn, Philipp Hühn, and Maximilian Förster. 2024. Wasn't Expecting that – Using Abnormality as a Key to Design a Novel User-Centric Explainable AI Method. In *Design Science Research for a Resilient Future*, Munir Mandviwalla, Matthias Söllner, and Tuure Tuunanen (Eds.). Springer Nature Switzerland, Cham, 66–80. https://doi.org/10.1007/978-3-031-61175-9_5

[32] Ion Juvina, Michael G. Collins, Othalia Larue, William G. Kennedy, Ewart de Visser, and Celso de Melo. 2019. Toward a Unified Theory of Learned Trust in Interpersonal and Human-Machine Interactions. *ACM Transactions on Interactive Intelligent Systems* 9, 4 (2019), 1–33. https://doi.org/10.1145/3230735 Publisher: ACM.

[33] Peter K. Kahr, Gerrit Rooks, Martijn C. Willemsen, and others. 2024. Understanding Trust and Reliance Development in AI Advice: Assessing Model Accuracy, Model Explanations, and Experiences from Previous Interactions. *ACM Transactions on Interactive Intelligent Systems* 14, 1 (2024), 3686164. https://doi.org/10.1145/3686164 Publisher: ACM.

[34] Frank Kaptein, Joost Broekens, Koen Hindriks, and Mark Neerincx. 2019. Evaluating Cognitive and Affective Intelligent Agent Explanations in a Long-Term Health-Support Application for Children with Type 1 Diabetes. In *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*. 1–7. https://doi.org/10.1109/ACII.2019.8925526

[35] Frank Kaptein, Joost Broekens, Koen Hindriks, and Mark Neerincx. 2019. Evaluating cognitive and affective intelligent agent explanations in a long-term health-support application for children with type 1 diabetes. In *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 1–7.

[36] Michael Kelly, Anuj Kumar, Padhraic Smyth, and Mark Steyvers. 2023. Capturing Humans' Mental Models of AI: An Item Response Theory Approach. In *Proceedings of the 2023 ACM SIGCHI Conference on Human Factors in Computing Systems (CHI)*. https://doi.org/10.1145/3593013.3594111

[37] Anjali Khurana, Parsa Alamzadeh, and Parmit K. Chilana. 2021. ChatrEx: Designing Explainable Chatbot Interfaces for Enhancing Usefulness, Transparency, and Trust. In *2021 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*. 1–11. https://doi.org/10.1109/VL/HCC51201.2021.9576440

[38] Chanjun Kim, Xiang Lin, Christopher Collins, Graham W. Taylor, and others. 2021. Learn, Generate, Rank, Explain: A Case Study of Visual Explanation by Generative Machine Learning. *ACM Transactions on Interactive Intelligent Systems* 11, 3-4 (2021), 3465407. https://doi.org/10.1145/3465407 Publisher: ACM.

[39] Jiwoo Kim, Sangwoo Yu, Ryan Detrick, and Ning Li. 2024. Exploring Students' Perspectives on Generative AI-Assisted Academic Writing. *Education and Information Technologies* (2024). https://doi.org/10.1007/s10639-024-12878-7

[40] Sunnie S. Y. Kim, Elizabeth Anne Watkins, Olga Russakovsky, Ruth Fong, and Andrés Monroy-Hernández. 2023. Humans, AI, and Context: Understanding End-Users' Trust in a Real-World Computer Vision Application. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (Chicago, IL, USA) *(FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 77–88. https://doi.org/10.1145/3593013.3593978

[41] Helena Kopecka, Jason Such, and Michael Luck. 2024. Preferences for AI Explanations Based on Cognitive Style and Socio-Cultural Factors. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW2 (2024), 345–369. https://doi.org/10.1145/3637386

[42] Esther S. Kox, Johanna H. Kerstholt, Tjerk F. Hueting, and Peter W. de Vries. 2021. Trust Repair in Human-Agent Teams: The Effectiveness of Explanations and Expressing Regret. *Autonomous Agents and Multi-Agent Systems* 35, 2 (2021), 1–20. https://doi.org/10.1007/s10458-021-09515-9 Publisher: Springer.

[43] Emilie S. Kox, L. B. Siegling, and J. H. Kerstholt. 2022. Trust Development in Military and Civilian Human–Agent Teams: The Effect of Social-Cognitive Recovery Strategies. *International Journal of Social Robotics* 14 (2022), 289–305. https://doi.org/10.1007/s12369-022-00871-4

[44] Michael Kraus, Natalie Wagner, Zoraida Callejas, and Wolfgang Minker. 2021. The Role of Trust in Proactive Conversational Assistants. *IEEE Access* 9 (2021), 25610–25623. https://doi.org/10.1109/ACCESS.2021.9509523

[45] Shan G. Lakhmani, Julia L. Wright, Michael Schwartz, and Daniel Barber. 2020. Exploring the Effect of Communication Patterns and Transparency on the Attitudes Towards Robots. In *Advances in Human Factors and Simulation*, Daniel N. Cassenti (Ed.). Springer International Publishing, Cham, 27–36. https://doi.org/10.1007/978-3-030-20148-7_3

[46] Cameron Lawless, Jochen Schoeffer, Loc Le, Kevin Rowan, and Sandip Sen. 2024. "I Want It That Way": Enabling Interactive Decision Support Using Large Language Models and Constraint Programming. *ACM Transactions on Interactive Intelligent Systems* (2024). https://doi.org/10.1145/3685053

[47] Huao Li, Yao Fan, Keyang Zheng, Michael Lewis, and Katia Sycara. 2023. Personalized Decision Supports based on Theory of Mind Modeling and Explainable Reinforcement Learning. In *2023 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. 4865–4870. https://doi.org/10.1109/SMC53992.2023.10394414

[48] Han Liu, Vivian Lai, and Chenhao Tan. 2021. Understanding the Effect of Out-of-Distribution Examples and Interactive Explanations on Human-AI Decision Making. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 3479552. https://doi.org/10.1145/3479552 Publisher: ACM.

[49] Michael Lochner and Daniel Smilek. 2023. The Uncertain Advisor: Trust, Accuracy, and Self-Correction in an Automated Decision Support System. *Cognitive Processing* 24 (2023), 123–137. https://doi.org/10.1007/s10339-022-01113-1

[50] Joseph Lopez, Cameron Textor, Craig Lancaster, Ben Schelble, and Gary Freeman. 2023. The Complex Relationship of AI Ethics and Trust in Human–AI Teaming: Insights from Advanced Real-World Subject Matter Experts. *AI and Ethics* (2023). https://doi.org/10.1007/s43681-023-00303-7 Publisher: Springer.

[51] Yu Lu, Deliang Wang, Penghe Chen, and Zhi Zhang. 2024. Design and Evaluation of Trustworthy Knowledge Tracing Model for Intelligent Tutoring System. *IEEE Transactions on Learning Technologies* 17, 1 (2024), 1–10. https://doi.org/10.1109/TLT.2024.10535209 Publisher: IEEE.

[52] Jeffrey B. Lyons and Karen T. Wynne. 2021. Human-Machine Teaming: Evaluating Dimensions Using Narratives. *Human-Intelligent Systems Integration* (2021). https://doi.org/10.1007/s42454-020-00019-7

[53] Akihiro Maehigashi, Yosuke Fukuchi, and Seiji Yamada. 2023. Empirical investigation of how robot's pointing gesture influences trust in and acceptance of heatmap-based XAI. In *2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. 2134–2139. https://doi.org/10.1109/RO-MAN57019.2023.10309507

[54] Dong Hae Mangalindan, Ericka Rovira, and Vaibhav Srivastava. 2023. On Trust-aware Assistance-seeking in Human-Supervised Autonomy. In *2023 American Control Conference (ACC)*. 3901–3906. https://doi.org/10.23919/ACC55779.2023.10156103

[55] John L. Marble, Abigail M. Greenberg, James W. Bonny, and Samuel M. Kain. 2021. Platforms for Assessing Relationships: Trust with Near Ecologically-Valid Risk, and Team Interaction. In *Intelligent Systems: A Comparative Analysis*. 113–128. https://doi.org/10.1007/978-3-030-89385-9_13

[56] Gerald Matthews, Jinchao Lin, April Rose Panganiban, and Michael D. Long. 2020. Individual Differences in Trust in Autonomous Robots: Implications for Transparency. *IEEE Transactions on Human-Machine Systems* 50, 3 (June 2020), 234–244. https://doi.org/10.1109/THMS.2019.2947592

[57] Srishti Mehrotra, Catholijn M. Jonker, and Myrthe L. Tielman. 2021. More Similar Values, More Trust? The Effect of Value Similarity on Trust in Human-Agent Interaction. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. ACM, 872–881. https://doi.org/10.1145/3461702.3462576

[58] Meera Natarajan and Matthew Gombolay. 2020. Effects of Anthropomorphism and Accountability on Trust in Human-Robot Interaction. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*. ACM, 33–42. https://doi.org/10.1145/3319502.3374839

[59] Brian Nesset, Gopalan Rajendran, Joao D. Almeida Lopes, and Kerstin Dautenhahn. 2022. Sensitivity of Trust Scales in the Face of Errors. In *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. https://doi.org/10.1109/HRI53445.2022.9889427

[60] Xuan Phong Nguyen, Thanh Hai Tran, Nhat Bao Pham, Dinh-Nguyen Do, and Takehisa Yairi. 2022. Human Language Explanation for a Decision Making Agent via Automated Rationale Generation. *IEEE Access* 10 (2022), 122457–122470. https://ieeexplore.ieee.org/document/9918019/

[61] X Phong Nguyen, Tho H Tran, Nguyen B Pham, Dung N Do, and Takehisa Yairi. 2022. Human language explanation for a decision making agent via automated rationale generation. *IEEE Access* 10 (2022), 110727–110741.

[62] Chuka T. Okolo, Divyansh Agarwal, Nicola Dell, and Aditya Vashistha. 2024. "If it is easy to understand then it will have value": Examining Perceptions of Explainable AI with Community Health Workers in Rural India. *Proceedings of the ACM on Human-Computer Interaction* 8, CSCW1 (2024), 3637348. https://doi.org/10.1145/3637348 Publisher: ACM.

[63] Alessio Pipitone, Antonio Geraci, Alessandro D'Amico, and Valeria Seidita. 2024. Robot's Inner Speech Effects on Human Trust and Anthropomorphism. *International Journal of Social Robotics* (2024). https://doi.org/10.1007/s12369-023-01002-3

[64] Peng Qian and Vaibhav V. Unhelkar. 2024. Interactively Explaining Robot Policies to Humans in Integrated Virtual and Physical Training Environments. In *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*. 523–531. https://doi.org/10.1145/3610978.3640656

[65] Ashish Raikwar, David Mifsud, and Christopher D. Wickens. 2024. Beyond the Wizard of Oz: Negative Effects of Imperfect Machine Learning to Examine the Impact of Reliability of Augmented Reality Cues on Visual Search. *IEEE Transactions on Human-Machine Systems* (2024). https://doi.org/10.1109/THMS.2024.10458346

[66] Marco Romeo, Patrick E. McKenna, Daniel A. Robb, and others. 2022. Exploring Theory of Mind for Human-Robot Collaboration. In *2022 IEEE International Conference on Robotics and Automation (ICRA)*. 789–797. https://doi.org/10.1109/ICRA.2022.9900550

[67] Kenneth A. Roundtree, Jack R. Cody, Joshua Leaf, and Hakan O. Demirel. 2021. Human-Collective Visualization Transparency. *Swarm Intelligence* (2021). https://doi.org/10.1007/s11721-021-00194-6

[68] Sina Salimzadeh and Ujwal Gadiraju. 2024. When in Doubt! Understanding the Role of Task Characteristics on Peer Decision-Making with AI Assistance. In *Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization (UMAP)*. https://doi.org/10.1145/3627043.3659567

[69] Nina Scharowski, Markus Benk, Sarah J. Kühne, and Lukas Wettstein. 2023. Certification Labels for Trustworthy AI: Insights from an Empirical Mixed-Method Study. *Proceedings of the 2023 ACM Conference on Human Factors in Computing Systems* 6, CSCW2 (2023), 172–190. https://doi.org/10.1145/3593013.3593994

[70] Jordan R. Schoenherr and Robyn Thomson. 2024. When AI Fails, Who Do We Blame? Attributing Responsibility in Human-AI Interactions. *IEEE Transactions on Technology and Society* 5, 1 (2024), 10457538. https://doi.org/10.1109/TTS.2024.10457538 Publisher: IEEE.

[71] Tim Schrills and Thomas Franke. 2020. Color for Characters - Effects of Visual Explanations of AI on Trust and Observability. In *Artificial Intelligence in HCI*, Helmut Degen and Lauren Reinerman-Jones (Eds.). Springer International Publishing, Cham, 121–135. https://doi.org/10.1007/978-3-030-50334-5_8

[72] Larissa Shamseer, David Moher, Mike Clarke, Davina Ghersi, Alessandro Liberati, Mark Petticrew, Paul Shekelle, and Lesley A Stewart. 2015. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015: elaboration and explanation. *BMJ* 349 (2015). https://doi.org/10.1136/bmj.g7647

[73] Divya Srivastava, J. Mason Lilly, and Karen M. Feigh. 2024. Exploring the Role of Judgement and Shared Situation Awareness When Working with AI Recommender Systems. *Cognition, Technology & Work* (2024). https://link.springer.com/article/10.1007/s10111-024-00771-9

[74] Hugo Vasconcelos, Maximilian Jörke, Rafael Lins, and Ujwal Gadiraju. 2023. Explanations Can Reduce Overreliance on AI Systems During Decision-Making. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW2 (2023), 3579605. https://doi.org/10.1145/3579605 Publisher: ACM.

[75] Xinru Wang and Ming Yin. 2022. Effects of explanations in ai-assisted decision making: Principles and comparisons. *ACM Transactions on Interactive Intelligent Systems* 12, 4 (2022), 1–36.

[76] Xiang Wang and Ming Yin. 2022. Effects of Explanations in AI-Assisted Decision Making: Principles and Comparisons. *ACM Transactions on Interactive Intelligent Systems* 12, 1 (2022), 3519266. https://doi.org/10.1145/3519266 Publisher: ACM.

[77] Katharina Weitz, Ruben Schlagowski, and Elisabeth André. 2021. Demystifying artificial intelligence for end-users: findings from a participatory machine learning show. In *German Conference on Artificial Intelligence (Künstliche Intelligenz)*. Springer, 257–270.

[78] Katharina Weitz, Ralf Schlagowski, and Elisabeth André. 2021. Demystifying Artificial Intelligence for End-Users: Findings from a Participatory Machine Learning Show. In *German Conference on Artificial Intelligence (KI 2021)*. Springer, 247–253. https://doi.org/10.1007/978-3-030-87626-5_19

[79] Andrea Wentzel, Sherine Attia, Xiaoyu Zhang, and others. 2024. DITTO: A Visual Digital Twin for Interventions and Temporal Treatment Outcomes in Head and Neck Cancer. *IEEE Transactions on Medical Imaging* (2024). https://doi.org/10.1109/TMI.2024.10670532

[80] Maximilian Wittmann. 2024. Exploring the Effect of Anthropomorphic Design on Trust in Industrial Robots: Insights from a Metaverse Cobot Experiment. In *2024 21st International Conference on Ubiquitous Robots (UR)*. IEEE, 118–124.

[81] Yang Ye, Hengxu You, and Jing Du. 2023. Improved Trust in Human-Robot Collaboration with ChatGPT. *IEEE Access* 11 (2023), 54678–54689. https://doi.org/10.1109/ACCESS.2023.10141597

[82] Yang Ye, Hengxu You, and Jing Du. 2023. Improved trust in human-robot collaboration with ChatGPT. *IEEE Access* 11 (2023), 55748–55754.

[83] Pian Yu, Shuyang Dong, Shili Sheng, Lu Feng, and Marta Kwiatkowska. 2024. Trust-aware motion planning for human-robot collaboration under distribution temporal logic specifications. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 12949–12955.

[84] Rui Zhang, Wen Duan, Christopher Flathmann, Nathan McNeese, Guo Freeman, and Alyssa Williams. 2023. Investigating AI Teammate Communication Strategies and Their Impact in Human-AI Teams for Effective Teamwork. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW2 (Oct. 2023), 281:1–281:31. https://doi.org/10.1145/3610072

[85] Ruijia Zhang, Christopher Flathmann, Gregory Musick, Benjamin Schelble, and others. 2024. I Know This Looks Bad, But I Can Explain: Understanding When AI Should Explain Actions in Human-AI Teams. *ACM Transactions on Interactive Intelligent Systems* 14, 1 (2024), 3635474. https://doi.org/10.1145/3635474 Publisher: ACM.

[86] Xiaotong Zou, Pei-Luen Patrick Rau, and Yixuan Zhao. 2024. Investigating the Impact of Different Stressors on Trust in Intelligent Decision Support Systems. In *International Conference on Human-Computer Interaction*. 315–326. https://link.springer.com/chapter/10.1007/978-3-031-60901-5_22