systems

MDPI

Systematic Review

X-AI Techniques for Human–AI Teams: The Implementation-Design Framework

John Turner *, Hoda Parvaneh Shirazi, Heesun Kim, Jiajia Du, Yeonji Jung and Xiaoyan Xu

Department of Educational Administration & Human Resource Development, College of Education and Human Development, Texas A&M University, College Station, TX 77843, USA; hoda.p87@tamu.edu (H.P.S.); heesun@tamu.edu (H.K.); dujiajia1231@tamu.edu (J.D.); yjjung@tamu.edu (Y.J.); xxu@tamu.edu (X.X.)
* Correspondence: j.turner@tamu.edu

Abstract

Explainable artificial intelligence (X-AI) techniques aim to make the actions and decisions of autonomous systems understandable to humans interacting with these systems. In human–AI teams, explainability supports individual understanding and coordination, shared mental models, and collective decision-making among humans and AI agents. Research has shown that X-AI enhances trust in autonomous systems, improves human–AI team performance, and supports collaboration across domains including aviation, finance, healthcare, hospitality, and sports. However, X-AI technologies face difficult challenges, including a lack of transparency and interpretability due to complex underlying models, also known as the "black-box" nature of AI systems. These technologies also lack any universally accepted evaluation metrics and have limited generalizability across applications. One deficit in the X-AI literature is that most frameworks focus on individual-level outcomes, with limited attention to team-level processes. The current study conducted a systematic literature review adhering to PRISMA guidelines and the SALSA framework. This study introduces the Implementation-Design (I-D) framework that organizes X-AI approaches along two dimensions: implementation, ranging from visual to interactive approaches, and design, ranging from isolated explanations to workflow-integrated systems. This framework captures lower-level engagement, involving individual users, to higher-level understanding that is necessary for teams and collectives. Findings indicate that visual explanation approaches support user engagement, while interactive workflow approaches promote deeper understanding, appropriate reliance, and distributed cognition within human–AI teams. Implications highlight the need for team-oriented explainability grounded in shared mental models, transactive memory systems, and collaborative X-AI artifacts. Practical guidelines are included to support researchers and practitioners in selecting appropriate X-AI techniques based on their context and level of analysis. The I-D framework is offered as a conceptual organizing model to guide research and practice, and empirical validation is identified as a priority for future work.

Keywords: explainable AI; human–AI teams; X-AI; systematic literature review

Check for updates

Academic Editor: Jinzhi Lu

Received: 21 May 2026
Revised: 3 July 2026
Accepted: 14 July 2026
Published: 20 July 2026

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

1. Introduction

Explainable artificial intelligence (X-AI) technologies aim to make the actions and decisions of autonomous systems understandable to humans interacting with these systems. In human–AI contexts, explainability is not only important for individual users' understanding but also for enabling coordination, shared understanding, and collective

decision-making among human and AI agents. Without X-AI technologies, trust in autonomous systems remains low [1,2] and varies depending on the level of expertise, task complexity, and the amount of risk involved [3,4]. In team-based environments, these challenges extend beyond individual trust to issues of trust calibration, coordination, and alignment between human and AI agents.

Recent research has identified several benefits of X-AI technologies, including increased trust in autonomous systems [2,5,6], improved performance [7,8], and enhanced understanding of how AI systems operate [9]. These benefits also contribute to improving human–AI collaboration [10,11]. Explainable AI is already being applied across a wide range of domains, including aviation [12,13], finance [7,14], healthcare [3,15], hospitality [2] and sports [16,17]. Recent domain-specific reviews further illustrate the rapid sectoral expansion of X-AI in clinical settings, including mortality prediction [18], dementia detection [19], depression assessment [20], and biomedical imaging [21]. Like much of the field, however, these reviews remain centered on individual clinical decision-making rather than team-level processes. Most of the X-AI research remains centered on individual-level outcomes such as user trust, interpretability, and decision support, with comparatively less attention given to how explainability supports team-level processes such as coordination, shared mental models, team effectiveness, and collective performance.

Despite the benefits and widespread applications, explainable AI comes with several challenges. First, X-AI systems often lack transparency and interpretability [22–25], partly due to the complex mathematical and statistical models underlying these systems [26]. This contributes to the "black-box" nature of AI [27,28] that can compromise trust [29]. Second, these limitations make it difficult to justify or explain how decisions are made, particularly in complex contexts [29–31]. Third, opaque systems may produce decisions that are fallible, incomplete, or misaligned with the problem context [32]. Fourth, there is no universally accepted metric for evaluating the quality of X-AI systems [30], prompting calls for standardized evaluation frameworks [26]. Finally, X-AI approaches often lack generalizability across diverse applications [31,33].

Researchers have called for improvements in the accuracy, applicability, and transparency of X-AI technologies [22,28,33]. However, these challenges are not solely rooted in model complexity. A growing body of research suggests that limitations in explainability also emerge from how AI outputs are communicated, interpreted, and integrated into human activity. This issue becomes especially critical in human–AI teams, where explainability must support not only understanding but also coordination, interaction, and collective sensemaking. In this regard, explainability can be understood as a sociotechnical design problem, enacted through interfaces, interactions, and workflows that shape how users engage with AI systems. Visual and interactive implementation approaches play a central role in this process by influencing how explanations are delivered, how users interact with them, and how they are embedded within decision-making processes. Rather than viewing these approaches as isolated techniques, they represent key mechanisms through which explainability is operationalized in practice. Accordingly, the current study focuses on how visual and interactive implementation approaches are designed and applied to improve explanations and workflow design in human–AI teams. By synthesizing these approaches, this study aims to address the limited integration of team-level explainability in the existing literature. The anticipated outcome is to identify how X-AI can better support human–AI engagement, coordination, and understanding within complex, team-based environments. The current review contributes to the X-AI and human–AI team literature by proposing the Implementation-Design framework that organizes X-AI approaches along two dimensions: implementation, ranging from visual to interactive approaches, and design, ranging from isolated explanations to workflow-integrated systems.

2. Methodology

The current study employed a systematic literature review to examine how explainable artificial intelligence is implemented and designed to support human–AI teams, with particular attention to visual, interactive, explanation-based, and workflow-integration approaches. This review adhered to the PRISMA 2020 guidelines [34,35] to ensure transparency and rigor in reporting and followed the SALSA framework (Search, Appraisal, Literature Review, Synthesis, and Analysis) as outlined by Booth et al. [36] to guide the process for this review.

2.1. Search and Appraisal

A comprehensive database search was conducted on 29 October 2025, to identify literature related to explainable artificial intelligence (X-AI) and human–AI teams. The search was designed in alignment with systematic review standards and followed PRISMA 2020 reporting procedures to ensure transparency and rigor. To capture interdisciplinary scholarship, multiple databases were searched, including ERIC, Education Source Ultimate, APA PsycInfo, Computer Source, Human Resources Abstracts, Business Source Ultimate, Academic Search Ultimate (EBSCO), Web of Science (Clarivate), and ProQuest Dissertations & Theses Global. In addition, gray literature sources were consulted, including conference proceedings indexed in the Web of Science, Organizational and Government Reports, and Policy Commons.

The search strategy was structured around key conceptual categories. In Academic Search Ultimate (EBSCO), for example, the search combined the following title/abstract terms:

- Concept 1 (Explainable AI):
  "Explainable AI technique*" OR "X-AI" OR "XAI" OR "explain* AI" OR "AI explanation" OR "explanation user interface" OR ("explainable AI" N2 technique*).
- Concept 2 (Human–AI Teams):
  "Human-AI team*" OR (Human* N2 AI).

These concepts were combined using Boolean operators (AND/OR) and proximity operators (e.g., N2) to maximize sensitivity while maintaining conceptual precision. Where available, subject headings were also incorporated to enhance retrieval. No date restrictions were imposed to capture both foundational and emerging research in this evolving field. The research deliberately centered on the two core constructs that define the review's scope (explainable AI and human–AI teams) rather than on the broader adjacent literature. Related terms such as "inte4rpretable SI," "interpretable machine learning," "human-autonomy teaming," and "human-agent collaboration" were not used as primary search terms. Although the truncated "explain*" operator and the proximity-based "Human* N2 AI" expression captured many conceptually adjacent records, the comparatively focused term set is acknowledged as a limitation that may have excluded studies indexed only under these alternative labels (Section 4.5).

All retrieved records were exported into a reference management system, and duplicates were removed prior to screening. The initial search yielded a substantial pool of studies across databases.

2.2. Screening Process

Following PRISMA 2020 guidelines, screening occurred in two stages.

2.2.1. Stage 1: Title and Abstract Screening

All identified studies were reviewed at the title and abstract level to assess preliminary relevance. Studies were excluded if they

- Did not focus on explainable AI or AI explanation mechanisms;
- Did not address human–AI collaboration, interaction, or teaming;
- Focused solely on technical algorithm development without human/team implications;
- Were unrelated to organizational, team, or collaborative contexts;
- Were non-scholarly materials (e.g., editorials, opinion pieces, book reviews).

2.2.2. Stage 2: Full-Text Screening

Studies that met initial relevance criteria were retrieved for full-text review. Explicit inclusion and exclusion criteria were developed prior to screening and consistently applied.

Inclusion Criteria

Studies were included if they

1. Examined explainable AI, AI transparency, or AI explanation techniques.
2. Addressed human–AI interaction, collaboration, or team-based contexts.
3. Included empirical, conceptual, or theoretical contributions advancing understanding of human–AI teamwork.
4. Were published in English.
5. Were peer-reviewed journal articles, conference papers, or relevant scholarly dissertations.
6. Were accessible (e.g., full text).

Exclusion Criteria

Studies were excluded if they

1. Focused exclusively on algorithmic performance without human/team implications.
2. Examined individual human–computer interaction without collaborative or team dimensions.
3. Were purely technical engineering papers unrelated to organizational, behavioral, or team outcomes.
4. Were non-peer-reviewed materials (unless included as relevant gray literature).
5. Represented duplicate publications (in which case the most complete version was retained).

Consistent with the English language inclusion criterion, non-English language publications were excluded during screening.

2.3. Data Extraction and Review Process

All included studies were analyzed using a structured data extraction matrix. The matrix captured key information such as

- Publication year and authors;
- Disciplinary field;
- Conceptualization of explainability;
- Team or collaboration context;
- Methodology (qualitative, quantitative, mixed methods, conceptual);
- Theoretical frameworks;
- Key findings related to human–AI collaboration (e.g., trust, decision-making, performance).

The review team met regularly throughout the screening and extraction phases to resolve discrepancies and ensure consistency in coding. Disagreements were discussed and resolved through consensus to maintain methodological rigor and transparency. To improve transparency and reproducibility, each record was independently assessed by at least two members of the review team using a shared, predefined coding schema (Table 1)

that operationalized the review's core categories (visual, interactive, explanation, workflow) with explicit definitions, classification cues, and exemplar studies. Coders applied these definitions independently and then reconciled their classifications in regular adjudication meetings, with the lead author resolving any residual disagreements. Because final classifications were established through this structured consensus procedure, rather than through independent parallel double-coding of the literature, formal inter-rater reliability statistics (e.g., Cohen's K, Krippendorff's alpha) were not computed. This is acknowledged as a limitation (Section 4.5), and future protocol-registered extensions of this review should report such indices. This multi-database, multi-stage screening approach strengthened the comprehensiveness and credibility of the review while ensuring alignment with PRISMA 2020 standards.

Table 1. Coding schema for the core review categories.

Category (Dimension)	Operational Definition	Classification Cues	Exemplar Study(ies)
Visual (Implementation)	Explainability delivered through static graphical representations of model behavior that support perceptual access and interpretation.	Saliency/attention maps, SHAP or feature-importance plots, partial dependence plots, decision-tree diagrams; user views but cannot manipulate the explanation.	[22,37]
Interactive (Implementation)	Explainability enacted through user participation, allowing users to query, manipulate, or contest explanations during task execution.	Counterfactual re-runs, parameter adjustment, on-demand or dialogic explanation, what-if exploration.	[38,39]
Explanation (Design)	Design oriented toward discrete, isolated accounts of specific outputs or predictions for end users.	Post-hoc feature attribution or example-based rationales delivered as standalone artifacts.	[26,33]
Workflow (Design)	Design that embeds explainability within decision processes, coordination structures, and the AI lifecycle over time.	Task/requirements analysis, timing- and role-aligned delivery, governance or maturity processes.	[40,41]

The search and screening procedures resulted in an initial pool of 1250 records identified across the selected databases and gray literature sources. After removing 210 duplicate records, 1040 records were screened at the title and abstract level. Of these, 850 records were excluded because they did not meet the scope of the review. A total of 190 reports were sought for retrieval, and 20 reports were excluded because the full text was not available or could not be accessed. The remaining 170 full-text reports were assessed for eligibility. During the full-text screening, 100 reports were excluded for reasons including insufficient focus on human–AI teaming or collaboration, purely technical emphasis without human or team implications, individual-level human–computer interaction without a team dimension, lack of relevance to organizational or collective decision-making contexts, non-scholarly format, insufficient methodological detail, or duplicate/overlapping publication. Ultimately, 70 studies met all inclusion criteria and were retained for the final review and synthesis. This systematic selection process strengthened the rigor and transparency of the study while ensuring that the final evidence base was directly aligned with the research purpose. Figure 1 presents the PRISMA flow diagram summarizing the

article selection process for the current study (see Supplementary Materials: PRISMA 2020 Checklist, X-AI PRISMA Figure). The figure illustrates the progression of records from the initial identification of 1250 studies, through the screening and exclusion phases, to the final inclusion of 70 studies in the review. In line with the PRISMA 2020 guidelines, this visual representation provides a transparent account of the evidence selection process and enables readers to clearly trace how the final sample of studies was established.

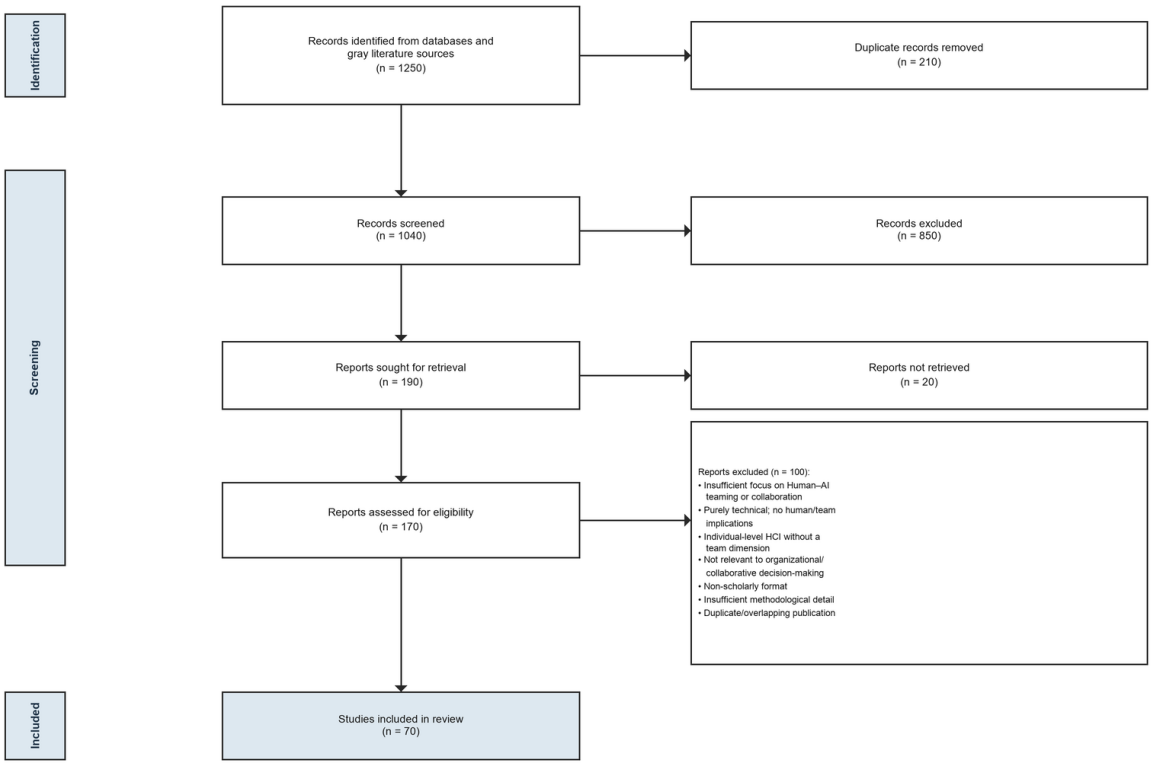


Figure 1. Search strategy.

3. Review of the Literature

The review of the literature synthesizes prior scholarship on explainable artificial intelligence within human–AI team contexts to establish the conceptual and empirical foundation for the current study. This section begins by clarifying key definitions related to explainable AI, human–AI collaboration, and team-based interaction, followed by an examination of major theoretical perspectives that have guided research in this area. It then reviews the documented benefits and limitations of explainability for trust, decision-making, coordination, and performance, while highlighting how these outcomes differ at the individual and team levels of analysis. Finally, the section identifies critical gaps in the existing literature, particularly the limited attention given to collective processes, organizational applications, and the broader sociotechnical conditions that shape effective human–AI teamwork.

3.1. Definitions

Explainable artificial intelligence (X-AI) is defined in multiple ways across the literature, with some definitions remaining broad and others tailored to specific contexts. In general, X-AI refers to the ability of AI systems to provide understandable explanations for their decisions and actions. As Kolajo and Daramola [4] noted, "XAI is a developing area of AI research that supports a collection of instruments, methods, and algorithms that may produce superior interpretable, intuitive, and human-comprehensible justifications for AI actions" (p. S120). Other general definitions included X-AI as a means of offering

explanations [41,42], making visible the "black-box" processes that often come with AI technologies [29] and enhancing the understanding of how AI knowledge is created [15].

Within the machine learning (ML) literature, some view X-AI as a subfield of ML [24,30,33], while others embed X-AI into current models that provide "instance-level explanations" [28] (p. 2). X-AI has also been identified as being a subfield of Safe AI that enhances human–AI capabilities through an "understanding of *automation's inner workings*" [12] (p. 869). Some include human-centered X-AI to highlight their focus on human–AI decision-makers [43,44] and to cultivate trust in AI [45]. Some differentiate between low-level and high-level X-AI. Low-level X-AI refers to information relating to a decision process, whereas high-level X-AI provides an explanation about the entire process that also includes a decision logic tree [3]. For the field of education, X-AI stressed the importance of enhancing "students' AI literacy" [11] (p. 1865).

Within this body of work, transparency and explainability are consistently emphasized as being as important as the accuracy of AI outcomes [4]. This perspective reinforces the idea that effective AI systems must not only perform well but also communicate their reasoning in ways that are meaningful to human users.

3.2. Explainable Approaches

Explainable AI techniques have been categorized into four primary dimensions: ante hoc (intrinsic) versus post hoc, model-specific versus model-agnostic, surrogate approaches, and global versus local explanations [4]. Ante hoc approaches rely on inherently interpretable models, where transparency is embedded directly into the system design. In contrast, post hoc approaches are used to explain models that are not inherently interpretable, attempting to provide insight into systems that lack intrinsic transparency. Within post hoc approaches, model-specific explanations focus on interpreting the internal structure and mechanics of a given model, whereas model-agnostic explanations operate independently of the model, providing explanations based on input–output relationships. Together, these approaches offer complementary perspectives, enabling a more comprehensive understanding of AI systems. Surrogate models, also categorized as post hoc approaches [4], are used to approximate and explain more complex models by replacing them with simpler, interpretable representations. Finally, global explanations aim to describe the overall behavior of a model, while local explanations focus on specific predictions or instances [4]. The techniques examined in this study are categorized according to these dimensions. The following sections synthesize the literature by identifying how different techniques align with these categories, while also considering how their application varies across contexts.

3.3. Explainable Techniques

Explainable AI operates across multiple levels of analysis, ranging from individual users (e.g., end users and user experience) to collective contexts such as groups and teams. Although individual-level X-AI is not the primary focus of this study, both levels were examined to clarify their distinct purposes, benefits, and limitations. This distinction is critical, as it demonstrates that the selection of X-AI techniques should be driven by the level of analysis and use context, rather than solely by technical sophistication or popularity in the broader literature.

3.3.1. X-AI at the Individual Level

Among the reviewed studies, a substantial portion (n = 14 out of 38) applied X-AI at the individual level, consistently emphasizing that explainability is inherently audience-dependent. This perspective is reflected both in how X-AI is defined [30,46] and in how it is applied in practice, such as supporting decision-makers in disaster assessment [37],

assisting forecasters and emergency managers [47], or aiding patients and clinicians in healthcare contexts [46,48].

Because these applications target specific users, they tend to prioritize local, instance-specific explanations over global representations of model behavior. Mohseni et al. [44] defined local explanations as those that "aim to explain the relationship between specific input–output pairs or the reasoning behind the results for an individual user query" (p. 24). Correspondingly, individual-level X-AI implementation strategies emphasize personalized interface design, visualization, and interaction modalities tailored to single users [25,37,49]. This emphasis aligns with the predominance of post hoc explanation methods at the individual level, which are applied after AI decisions are made. These methods support users in understanding past decisions, evaluating specific predictions, and learning about system behavior [30,50]. Within this context, trust in AI is conceptualized as a subjective, individual judgment, often measured through self-reported attitudes and Likert-scale assessments of perceived trustworthiness [30,44,47]. As a result, the effectiveness of individual-level X-AI is typically evaluated through outcomes such as task performance, cognitive load, and individual decision quality [30,37]. These findings highlight that, at the individual level, explainability primarily functions as a mechanism for supporting comprehension, justification, and personal decision-making.

3.3.2. X-AI Team/Group Level

While individual-level X-AI primarily focuses on user comprehension, higher levels of X-AI application shift attention toward teams, emphasizing shared mental models, collective understanding, and improvements in team performance (n = 18). Although there is not yet a fully established theoretical framework for shared mental models in human–AI contexts, Andrews et al. [51] conceptualized them as a shared conceptual and methodological foundation between human factors and AI explainability. However, empirical evidence remains limited regarding the extent to which shared mental models directly improve coordination and performance in human–AI teams [51]. At the team level, X-AI is designed to support communication, coordination, and workflow integration, rather than focusing solely on individual interface usability [3]. In contrast to individual-level X-AI, which emphasizes trust building, team-level X-AI prioritizes appropriate reliance and trust calibration, ensuring that human agents engage with AI systems in ways that reflect both their capabilities and limitations [3,26].

3.3.3. Techniques at the Individual Level

At the individual level, explainable AI frequently relies on techniques such as Shapley Additive Explanations (SHAP), which quantify the contribution of each feature to a model's output using Shapley values derived from cooperative game theory. SHAP provides both global and local feature importance, making it well-suited for individual decision-support contexts [22,52]. The prevalence of SHAP reflects its alignment with individual-level needs for interpretability and justification. For example, SHAP has been used to help researchers identify how factors such as age, placement, and nationality influence performance predictions in sports analytics [52], as well as to provide coaches with visual insights into areas requiring improvement [22]. These applications demonstrate how feature attribution techniques support localized understanding and individualized decision-making, reinforcing the broader pattern that individual-level X-AI prioritizes interpretability and explanation of specific outcomes. A brief overview of the different techniques and their associated approaches are provided in Table 2.

Table 2. X-AI techniques and approaches.

Technique	Context	Approach	Source
Case Based, Counterfactual, Decision Tree, Feature Importance	Virtual agents	Post hoc	[26]
IKE-XAI	Child development	Post hoc	[53]
SHAP, ICE, LIME	Decision-making	Post hoc	[42]
SHAP, LIME, ASTRID, G-Rex	Health education	Post hoc	[12]
Nonspecific	Hospitality	Post hoc	[2]
User-oriented visualizations, interfaces, and toolkits	Software development	Post hoc	[25]
Saliency Maps	Event detection	Model-specific	[4]
ICE, PDP	Event detection	Model-agnostic	[4]
Example-based Explanations	Energy sector	Local	[33]
LIME	Software engineering	Local	[28]
LIME, SHAP	Event detection	Local	[4]
The Distillation Technique; Propagation, Gradient, & Occlusion	Event detection	Global	[28]
SHAP	Decision support, aviation safety	Post hoc, model-agnostic	[54]
LIME	Team	Post hoc, local	[55]
Feature Importance Analysis	Framework development	Global or local	[41]
Attention Visualization	N/A	Local, model-specific	[41]
Decision Tree	Human–machine team	Ante hoc, local, global	[39]
SPEAR	Decision-making	Post hoc, model-specific, local	[38]
Counterfactual Analysis, Model Distillation	N/A	Post hoc, model-agnostic, local	[41]
SHAP	Sports	Post hoc, local, global	[22]
LIME, SHAP	N/A	Post hoc, local, global	[41]
IML	Sports	Post hoc, local, global	[23]

Note: Automatic STRucture Identification method (ASTRID); Generic Rule Extraction (G-Rex); Implicit Knowledge Extraction with eXplainable Artificial Intelligence (IKE-XAI); Individual Conditional Expectations (ICE); interpretable machine learning (IML); Local Interpretable Model-agnostic Explanations (LIME); partial dependence plot (PDP); Shapley Additive Explanations (SHAP); Single-shot Policy Elicitation for Augmenting Rewards (SPEAR).

3.4. The Implementation-Design (I-D) Framework

The current study is structured around the Implementation-Design (I-D) framework (see Figure 2), which conceptualizes explainable AI along two dimensions: implementations and design. The findings of this review indicate that implementation techniques range along a continuum from primarily visual approaches to more interactive approaches. At the same time, design approaches range from focusing on isolated explanations for end users to structuring broader workflows and processes in which explanations are embedded. These dimensions form a continuum in which visual explanation techniques primarily support engagement at lower levels of the framework, while interactive workflow approaches support deeper understanding at higher levels.

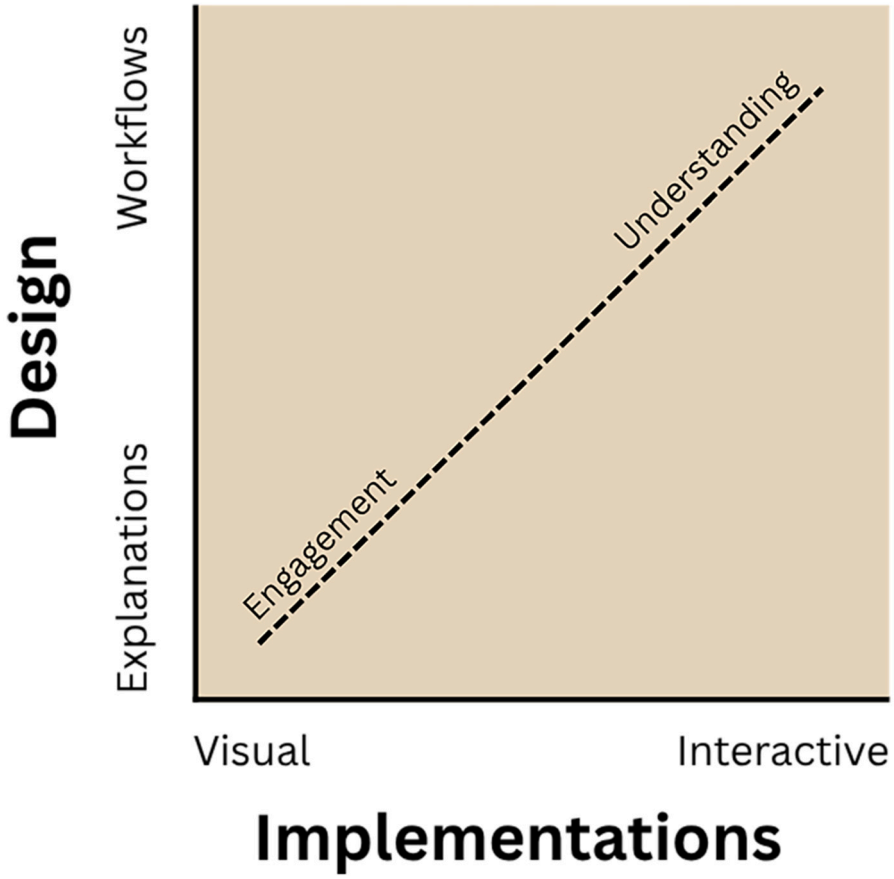


Figure 2. Implementation-Design framework.

The two dimensions of the I-D framework were derived inductively from the coding of the reviewed studies (Table 1) and were retained because they organize a question that existing X-AI taxonomies leave largely unaddressed. Established classifications (ante hoc versus post hoc, local versus global, and model-specific versus model-agnostic [30]) describe how an explanation is computed relative to the underlying model. The I-D framework is orthogonal to these schemes: rather than classifying the computational provenance of an explanation, it characterizes how explainability is the operational provenance of an explanation, and it characterizes how explainability is operationalized for human use. The implementation dimension corresponds to an established distinction in human–computer interaction between representation and participation, between presenting model behavior and enabling users to act on it. The design dimension corresponds to the sociotechnical distinction between an isolated artifact and an embedded process. Positioning the framework along these two previously unlinked axes provides a theoretical rationale for the dimensions that extends beyond inductive categorization and clarifies that the I-D framework complements, rather than replaces, computation-centric taxonomies of explanation.

The I-D framework captures progressive growth that expands the utility of AI outputs from lower-level engagement toward higher-level understanding. Lower-level engagement is best when dealing with individual employees, whereas higher-level understanding is necessary when meeting the needs of larger teams and collectives. The primary goal of X-AI implementation, per the I-D framework, is not only to explain decisions for engagement but to also design interactive workflows that allow users to develop a clear and actionable understanding of the AI systems with which they interact.

The current literature review is organized around the I-D framework. First, studies emphasizing visual implementation approaches are examined, followed by those emphasizing

visual interactive approaches. Next, the review captures the design explanations literature, followed by the design workflow literature. Finally, the findings are synthesized to demonstrate how visual explanations support engagement and how interactive workflows contribute to higher levels of understanding. The analysis also considers the conditions under which engagement or understanding may be prioritized and when different X-AI techniques are most appropriately applied.

3.4.1. Implementations

A key distinction identified in this review is between visual and interactive implementations of explainable AI. This distinction aligns with the Implementation-Design (I-D) framework, where visual approaches represent lower levels of implementation focused on presenting information, while interactive approaches represent higher levels that enable user engagement, manipulation, and integration within workflows. Accordingly, clearly differentiating what constitutes "visual" versus "interactive" X-AI is critical for understanding how explainability is operationalized across contexts.

Visual

In the context of X-AI, "visualization" refers to the use of graphical representations, such as charts, plots, and diagrams, to make AI outputs more understandable and accessible to human users. Visualizations can be viewed as a means of connecting humans and machines, allowing humans to better identify patterns, interpret, and diagnose outcomes [56]. Because complex algorithms often produce results that are difficult to interpret, visualization plays a central role in translating these outputs into forms that support human intuition and sensemaking. As Barredo Arrieta et al. [30] noted, visualization techniques are particularly effective in communicating statistical and computational results to users who may not have advanced technical expertise. Visual explanations enable users to interpret abstract computational processes by representing elements such as gradients [44], attention weights [37], probability scores [26], and feature interactions [52] in recognizable patterns. Across the reviewed studies, common visualization techniques include saliency maps, SHAP-based visualizations, LIME-based visualizations, decision trees, feature importance graphs, uncertainty displays, and partial dependence plots (PDPs). Within the I-D framework, visual approaches primarily support perceptual access and initial engagement with AI outputs. They enhance transparency by making model behavior visible, but they typically remain descriptive and retrospective, providing insight into what the model has done rather than enabling users to actively influence or interrogate the system.

Saliency Maps

Saliency maps and attention-based visualizations are widely used in deep learning contexts to provide spatially grounded explanations. These techniques highlight the regions of input data that most strongly influence a model's prediction, allowing users to identify which features or areas are most relevant to the output. Recent studies demonstrate that saliency maps function as part of broader multimodal explanation systems. Andrews et al. [51], for example, showed that these visualizations can be combined with nonverbal cues, additional visual elements, and iterative dialogue to support users in constructing mental models of underlying AI processes. Technically, these visualizations are implemented through methods such as gradient-based saliency extraction, activation mapping, and attention mechanisms within neural network architectures [30]. While saliency maps enhance interpretability by making model reasoning more visible, they remain primarily visual artifacts. As such, they align with lower levels of the I-D framework, where explainability is delivered through representation rather than interaction. Their effectiveness lies

in supporting recognition and interpretation, but they do not, on their own, enable users to engage dynamically with AI systems or influence decision-making processes.

Interactive

Visual explainable AI (X-AI) methods primarily aim to represent model behavior through graphical outputs, whereas interactive X-AI extends this approach by actively involving users in the sensemaking process. Across the reviewed literature, interactivity is not defined by the mere presence of visual elements but by the extent to which users are enabled to participate in, influence, and interrogate the explanation process. In this sense, interactive X-AI shifts explainability from descriptive output toward participatory sensemaking, where understanding emerges through engagement rather than passive reception. This distinction is central to the Implementation-Design (I-D) framework. While visual approaches provide static representations of model behavior, interactive approaches embed explainability within user action, enabling dynamic exploration, adaptation, and decision-making. Thus, interactivity represents a higher level of implementation, where explainability becomes an active process rather than a delivered artifact.

Across the literature, interactive X-AI consistently demonstrates three defining characteristics. First, it enables user manipulation and exploration of the system, including modifying inputs, testing alternative scenarios, and examining the implications of different decisions [31,39]. Second, interactive explanations are temporally situated, meaning they are provided on demand or during task execution, rather than retrospectively after decisions are made [3,57]. Third, interactive X-AI supports iterative sensemaking, allowing users to refine their mental models of AI behavior through repeated interaction and feedback [10,51]. Unlike visual explanations, such as SHAP plots, saliency maps, or partial dependence plots, interactive explanations are not defined by their representational form but by their functional affordances. A system may include visual elements yet remain non-interactive if users cannot query, manipulate, or challenge the explanation. Conversely, interactive X-AI may incorporate visual, textual, or dialogic modalities, as long as users are actively engaged in the explanation process.

Several forms of interactivity emerge across the reviewed studies. One of the most prevalent is interactive decision support, where users directly engage with AI outputs to improve task performance. For example, Bansal et al. [55] demonstrated that enabling users to question AI predictions and compare them with human judgments improved complementary team performance. Similarly, Paleja and colleagues [39] found that embedding interpretable decision trees within human–AI systems allowed users to adapt their strategies based on AI feedback, fostering adaptive collaboration rather than static reliance. A second form involves policy- and reward-based interaction, where explanations are embedded within decision-making workflows. Tabrez et al. [31] introduced an approach in which users receive semantic feedback about AI policies and adjust their behavior accordingly. In these contexts, explainability functions not as retrospective justification but as an interactive coaching mechanism that shapes ongoing decision processes.

A third form is particularly prominent at the team and group level, where interactive X-AI supports shared understanding and coordination among multiple human and AI agents. Andrews et al. [51] and Hauptman et al. [3] emphasized that interactive explainability enables shared situation awareness and facilitates appropriate reliance in dynamic or high-risk environments. Here, interactivity extends beyond individual understanding to support collective reasoning, coordination, and error recovery. Additionally, collaborative interaction strategies highlight how humans and AI jointly contribute to decision-making processes. Gomez et al. [10] showed that engaging users in the explanation process helps mitigate knowledge asymmetries between human and AI agents, thereby improving both decision

quality and user engagement. These findings reinforce that interactivity is fundamentally rooted in human agency and participation rather than in the format of explanation delivery.

An important insight from the literature is that interactive X-AI often extends beyond explanation into workflow design. Unlike visual explanations, which support the interpretation of completed outputs, interactive explanations are embedded within task execution and influence decisions as they unfold. Studies by Hemmer and colleagues [57] and Hauptman et al. [3] demonstrated that aligning interactive explanations with task timing, decision criticality, and user roles is essential for achieving calibrated trust and effective human–AI collaboration. Within the I-D framework, this shift represents a movement from transparency to engagement, from making AI behavior visible to enabling users to actively work with, test, and adapt to AI systems. This transition is particularly critical in team-based contexts, where performance depends on coordination, adaptation, and appropriate reliance rather than individual interpretation alone [51]. Interactive X-AI should be understood not as a specific technique but as a design orientation. It emphasizes human agency, iterative sensemaking, and integration within workflows, positioning explainability as an ongoing, participatory process. Compared to purely visual approaches, interactive X-AI represents a more advanced level of implementation, particularly suited for complex, dynamic, and collaborative environments where static explanations are insufficient to support effective human–AI teaming.

3.4.2. Design

To enhance user trust in AI systems, Value Sensitive Design (VSD) has been introduced as an approach that ensures human values are systematically and comprehensively integrated throughout the design process [4]. Across empirical studies in explainable AI (X-AI) and human–AI teaming, "design" is most consistently conceptualized as a sociotechnical process. It extends beyond model development to include interface design, interaction structures, and the environmental conditions that support shared understanding and responsible use. Within the Implementation-Design (I-D) framework, design represents the highest level of explainability maturity, where explainability is not treated as an output or feature but as an embedded property of the entire human–AI system. In this sense, design integrates implementation choices (e.g., visual and interactive mechanisms) into coherent sociotechnical systems aligned with human goals, tasks, and values. A recurring pattern across the literature is team-oriented design, where AI is conceptualized as a teammate rather than a tool. In these systems, explainability is designed to support coordination, mutual predictability, and the development of shared mental models. Empirical and conceptual studies emphasize that effective human–AI team performance depends on aligned representations of tasks, roles, and expectations, positioning X-AI design as a mechanism for building and sustaining these shared understandings [40,51]. Supporting frameworks from human factor research, such as situation awareness, further inform design choices by emphasizing the need to make AI behavior observable, understandable, and anticipatable [58]. Accordingly, "good design" in X-AI moves beyond post hoc explanation and embeds explainability directly into teamwork processes.

Another prominent theme is the centrality of human-facing interface design. Explanations are frequently integrated into existing work environments, such as decision-support systems and collaborative platforms, and are presented in ways that align with human cognitive processes and communication norms [40,44]. This includes multimodal explanation strategies (e.g., visual panels, highlights, example-based explanations, uncertainty displays) as well as autonomy-aware design, where the depth, timing, and auditability of explanations adapt to the level of AI autonomy [3]. Across studies, design decisions are consistently framed as a balance between model-side transparency (e.g., interpretable archi-

tectures, uncertainty-aware systems) and user-side needs (e.g., comprehension, workload, expertise) [37,44].

Importantly, recent empirical work positions X-AI design within broader commitments to responsible and inclusive design. Rather than treating ethics as an add-on, considerations such as fairness, accountability, transparency, and accessibility are increasingly embedded as core design constraints [25,30]. This shift expands the meaning of design to include stakeholder-centered processes, such as participatory design discussions and early-stage co-reasoning among interdisciplinary teams [59]. The literature suggests that trustworthy X-AI design is best understood as human–environment integration. It involves aligning model capabilities, explanation mechanisms (visual and interactive), interface delivery, autonomy policies, and value commitments into a unified system. Within the I-D framework, this positions design as the level at which explainability becomes fully operationalized, transitioning from isolated explanations to embedded, value-driven, and context-aware human–AI systems.

Explanations

Across empirical studies, "explanations" refer to the specific representational techniques used to make an AI system's outputs, reasoning, uncertainty, or learned patterns more understandable and actionable for people. In practice, studies most frequently implement explanations through post hoc, model-agnostic methods (especially feature attribution) and example-based strategies, with delivery shaped by user expertise, task stakes, and interface constraints [30,44]. A dominant empirical approach is featuring attribution, particularly SHAP, used to quantify feature contributions and provide ranked importance, dependence, and interaction visualizations. This is common in applied performance contexts (e.g., team performance analytics), where explanations aim to translate complex prediction surfaces into interpretable drivers that users can discuss and act upon [22,23].

In parallel, LIME-style local explanations appear often in text classification and requirement-oriented contexts, offering highlighted tokens, local rationales, and stability/fidelity checks to support interpretability claims [28]. These choices reflect a pragmatic tendency: when predictive performance is prioritized, explanation is frequently layered on rather than built into inherently interpretable models [30].

Another prominent line of empirical work uses example-based explanations, such as showing similar cases or prototypes, which can support user trust and understanding by grounding model behavior in concrete instances [33]. Closely related are designs that combine visual explanation and natural language explanation, sometimes manipulating assertiveness or detail to examine how explanation style affects decision-making [26]. Across these studies, the explanation format is treated as a design variable that is visual, textual, or hybrid or interactive or non-interactive because comprehension and use depend heavily on how explanations are encountered in real tasks [44].

Importantly, empirical evidence also cautions that explanations do not uniformly improve outcomes. Some work shows that explanation quality and imperfections matter for downstream judgments and behavior and that human–AI teams can still underperform compared with AI alone in certain conditions, highlighting the risk of over-relying on explanations as a guarantee of appropriate trust or superior performance [10]. As a result, several empirical and review-based contributions recommend tailoring explanations to user expertise, case context, and decision stakes and combining system-level and prediction-level information to support calibrated reliance [29].

Workflow

In empirical studies, "workflow" captures how explainable AI (X-AI) is operationalized within human activity over time. More than a sequence of interactions, workflow represents a temporal and organizational construct that integrates decision processes, coordination structures, and system use across the AI lifecycle. It encompasses how informational needs are identified, how explanations are selected and delivered at different stages, and how their effectiveness is evaluated and refined. Across the literature, a consistent finding is that trustworthy X-AI depends not on isolated explanation outputs but on how well explainability is integrated into workflows that support decision-making and coordination [60]. In this sense, explainability effectiveness emerges from its alignment with when, where, and how decisions are made, rather than from the quality of standalone explanations alone. Within the Implementation-Design (I-D) framework, workflow represents the dynamic layer that connects implementation (visual and interactive mechanisms) with design (system-level integration), ensuring that explainability functions as an ongoing process rather than a static feature.

A key workflow contribution identified in the literature is task- and team-based requirements analysis. Studies emphasize defining goals, subgoals, decision points, and situation awareness needs before determining what explanations should be provided and in what form [53]. In human–AI team contexts, workflow design also involves clarifying interdependence and information-sharing structures, ensuring that explanations enhance coordination and mutual predictability rather than introduce redundancy or cognitive overload [51]. This aligns with human-centered design principles, which highlight that explanation timing, frequency, and level of detail must be carefully calibrated to support real-time performance.

A second recurring pattern is the use of iterative design–evaluation cycles. Explanation strategies are continuously refined based on empirical assessment of trust calibration, mental model development, task performance, and user experience. Multidisciplinary frameworks position evaluation not as an endpoint but as an integral stage of the workflow that must account for task context, user characteristics, and explanation goals [44,61]. This iterative approach is particularly important given evidence that explanations can sometimes mislead users or have variable effects depending on context [43]. A third dimension of workflow extends beyond user interaction into organizational governance. In this broader view, workflow includes structured processes for managing trust, risk, and accountability across the AI lifecycle. Approaches such as maturity models and governance frameworks prioritize transparency, privacy, and control mechanisms, while also identifying gaps and guiding continuous improvement [30,41]. Complementary research on early lifecycle co-reasoning further supports this perspective by demonstrating how interdisciplinary collaboration can embed explainability and ethical considerations into system design before deployment decisions constrain future options [59]. The literature positions workflow as the mechanism through which explainability becomes actionable and effective. It integrates temporal sequencing, organizational structures, and iterative learning processes, ensuring that explainability is not delivered as a one-time output but enacted as part of ongoing human–AI interaction and decision-making. Within the I-D framework, workflow thus plays a critical bridging role, connecting technical implementation and sociotechnical design into a coherent, adaptive system.

3.5. The Four Quadrants

The I-D Framework, shown in Figure 2, is divided into four quadrants. This framework is built along two orthogonal dimensions: the implementation dimension (horizontal) and the design dimension (vertical). Within these two dimensions are four quadrants.

The first quadrant (bottom-left) consists of the "Visual Implementations + Explanatory Design" components. This is the "Visual Explanations" quadrant and provides engagement through visual decision expectations. The second quadrant (top-left) consists of the "Visual Implementation + Workflows Design" components. This is the "Visual Workflows" quadrant and provides system-level understanding through visual design. The third quadrant (bottom-right) consists of the "Interactive Implementation + Explanatory Design" components. This is the "Interactive Explanations" quadrant and provides deep understanding through interactive exploration of decisions. The fourth quadrant (top-right) consists of the "Interactive Implementation + Workflows Design" components. This is the "Interactive Workflows" quadrant dedicated to comprehensive understanding through interactive engagement with the entire system's design. The four quadrants, along with their position, purpose, and primary goal are highlighted in Figure 3.

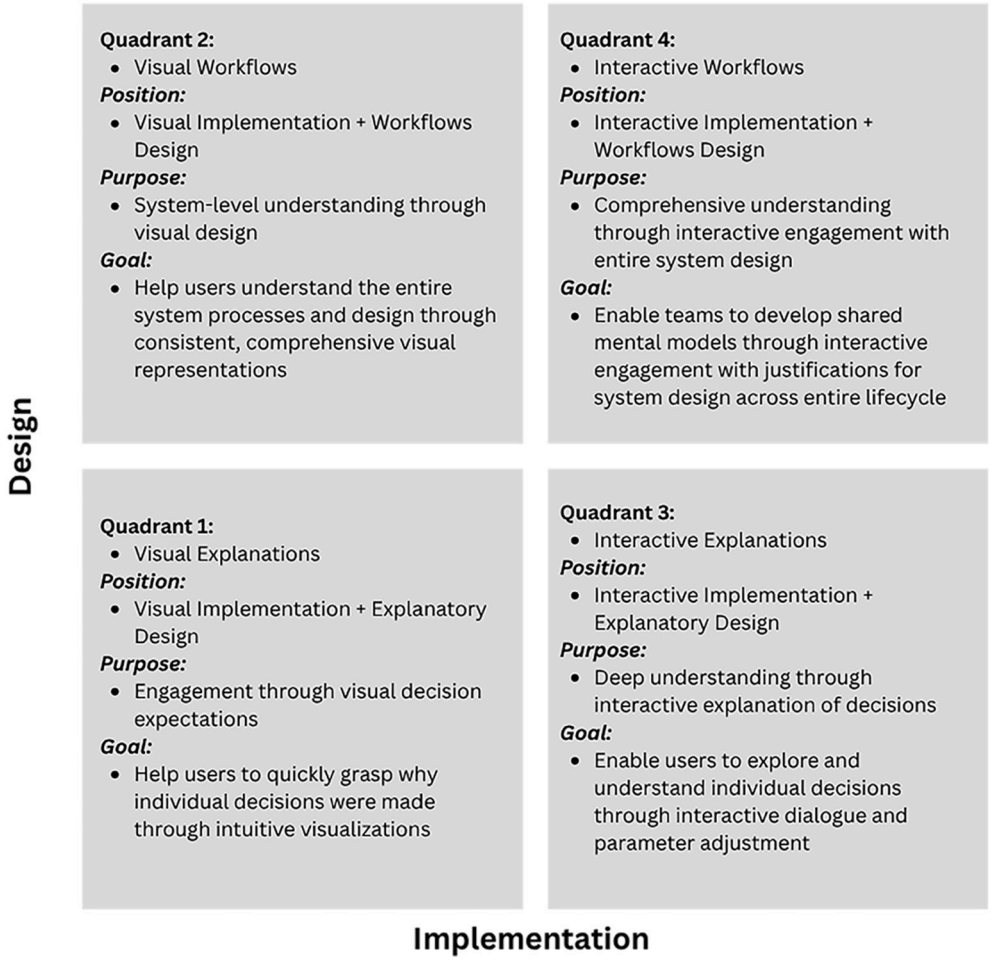


Figure 3. The four quadrants.

Although the framework is presented as four quadrants for clarity, the implementation and design dimensions are continuous axes rather than discrete bins. The quadrants are therefore best read as regions within two-dimensional space rather than as mutually exclusive categories. Many contemporary systems span more than one region. A SHAP dashboard, for example, combines a static visual representation (lower implementation), interactive exploration such as filtering and what-if adjustment (higher implementation), and embedding within a decision process (higher design). SHAP is therefore located not at a single point but along a trajectory that moves from the visual explanations region toward the interactive workflows region.

To classify such hybrid systems consistently, we apply two rules. First, a system is positioned according to its dominant explanatory function relative to the task, what the end user relies on to understand or act on the model. This is in contrast to just presenting visual or interactive elements. Second, systems that genuinely operate across regions are represented as vectors spanning the relevant quadrants, making multi-functionality an explicit property of the system. This helps to reduce ambiguity across the whole system. In summary, the non-exclusivity of the quadrants is a designed feature of a dimensional model. The axes describe where a system places its explanatory effort, and a single system can legitimately occupy a region, and edge, or a path between regions.

3.5.1. X-AI Strategies for the Four Quadrants

Part of the initial plan for the current study was to provide a list of X-AI techniques for each of the four quadrants. However, as it became apparent that this field is still new and technologies are constantly being designed, merged, or modified to meet new demands in the workplace, we decided to address specific strategies to help guide researchers and practitioners when deciding on which techniques work best for their situation. Table 3 provides a brief outline of recommended strategies that could be applied to each of the four quadrants.

Table 3. Strategies for the four quadrants.

Quadrant	Strategy
Q1: Visual Explanations	Add visualization layer to existing model
	Generate saliency maps, feature importance plots
	Create attention mechanism visualizations
	Design intuitive visual representations
Q2: Visual Workflows	Design unified visual language for system
	Create visualization ecosystem
	Document system processes visually
	Establish design standards for transparency
	Integrate visualizations throughout user experience
Q3: Interactive Explanations	Implement interactive explanatory tools
	Enable parameter adjustment and re-runs
	Support question-answering about decisions
	Create dialogue interfaces
	Adapt explanations to user inputs
Q4: Interactive Workflows	Build justificatory explanation systems
	Enable team dialogue about design assumptions
	Create shared visual representations
	Implement shared mental model development processes
	Support team learning behaviors

The decision to offer strategic guidance rather than a fixed catalogue of techniques is deliberate and reflects a design choice for sustainability. Because specific tools are continually being created, merged, and deprecated, a framework tied to named techniques would date quickly, whereas the I-D dimensions remain stable as the technique landscape changes. The framework is nonetheless operational. As a use case, a practitioner first identifies the

unit of analysis (e.g., individual, team) and whether the immediate need is engagement or deeper understanding. Then the practitioner selects the corresponding quadrant strategy from Table 3. Applied to systems already represented in this review, the framework consistently classifies: feature-importance visualizations for individual sports-performance feedback fall in visual explanations (Q1) [22]; uncertainty-aware visual displays used to brief multidisciplinary teams fall in visual workflows (Q2) [37,60]; interactive decision trees and counterfactual coaching that let users probe and adapt to model behavior fall in interactive explanations (Q3) [38,39]; and team-oriented, lifecycle-embedded explanation processes that support shared mental models fall in interactive workflows (Q4) [40,51]. These worked examples illustrate that the framework can locate existing systems even though it does not enumerate every technique.

3.5.2. Positioning the I-D Framework Among Existing Frameworks

To clarify the framework's contribution relative to established work, Table 4 compares the I-D framework with three influential frameworks in the X-AI and human–AI teaming literature: the Situation Awareness Framework for Explainable AI (SAFE-AI) [60]; Endsley's situation-awareness-based account of transparency and explainability for human–AI teams [40]; and Mohseni et al.'s [44] multidisciplinary design-and-evaluation framework along with the classic computation-centric taxonomy of explanation types [30]. The comparison shows that prior frameworks are oriented primarily toward the individual user, focusing on either the cognitive requirements of awareness or the provenance of explanation methods. The I-D framework organizes explainability by how it is implemented and designed for engagement versus understanding across individual and team levels. The I-D framework is therefore complementary rather than competing. It can be paired with situation-awareness requirements analysis and with method-level taxonomies, while adding an explicit team-oriented, operationalized-focused lens that these frameworks do not provide. This positioning indicates that the framework is a genuinely new organizing scheme rather than a relabeling of existing distinctions.

Table 4. Comparison of the I-D framework with existing frameworks.

Framework	Primary Focus	Unit of Analysis	What It Organizes	Relation to I-D Framework
SAFE-AI [60]	Aligning XAI with situation-awareness requirements	Individual operator	SA-based information requirements for explanations	SA requirements map onto the I-D workflow (design) dimension
Endsley [40]	Transparency, explainability, and SA for human–AI teams	Individual and team	Cognitive requirements for observable, predictable AI	I-D operationalizes these requirements as implementation and design choices
Mohseni et al. [44]	Multidisciplinary design and evaluation of XAI	Individual (audience-specific)	Design goals and evaluation measures by audience	Shares audience sensitivity; I-D adds the implementation × design axes
XAI taxonomy [30]	Classifying explanation methods	Model-centric	Ante/post hoc, local/global, model-specific/-agnostic	Orthogonal: provenance of explanations vs. their operationalization
I-D framework (this study)	Operationalizing explainability for engagement and understanding	Individual to team	Implementation (visual↔interactive) × design (explanation↔workflow)	Integrating scheme proposed in this review

4. Discussion

The findings of this systematic review suggest that explainable artificial intelligence (X-AI) should be understood not merely as a mechanism for improving transparency but

as a sociotechnical design problem that shapes how humans and AI coordinate, interpret, and act within team-based environments. Across the reviewed literature, a consistent pattern emerged. Much of the X-AI research remains focused on individual-level outcomes such as interpretability, trust, and decision support, whereas the demands of human–AI teams require explainability that is more interactional, workflow-integrated, and explicitly team-oriented. This shift matters because explanations in team settings are not simply retrospective accounts of model behavior. They function as resources for aligning expectations, coordinating action, and supporting collective sensemaking over time [51].

4.1. The I-D Framework to Support the Team Level of Analysis

A central implication of this review implies that explainability requirements shift as the level of analysis moves from individuals to teams. At the individual level, visual and post hoc explanations often serve interpretive and justificatory output. At the team level, however, explainability must do more than support interpretation. It must also facilitate communication, coordination, and mutual predictability among actors with different roles, expertise, information needs, and developing situation awareness [40,51,60].

The Implementation-Design (I-D) framework captures this transition. Lower-level visual techniques, such as feature importance plots, saliency maps, partial dependence plots, and other local explanations support perceptual access to model outputs. These approaches are often effective for individual engagement because they translate otherwise opaque model behavior into accessible signals. Unfortunately, these same techniques may be insufficient in team contexts, where success depends less on whether one person understands an output and more on whether multiple actors can align their interpretations and coordinate next steps [51].

This distinction aligns with the human–AI teaming literature, which emphasizes shared situation awareness, mutual understanding, and complementary performance rather than just individual task accuracy [5,51,54,62]. Shared mental models are especially relevant here [17,40,46,51,63]. Their development depends on team learning behaviors such as co-construction, collaborative construction, clarification, and constructive conflict through which team members negotiate common ground and reconcile differing interpretations [51]. Explanation in this sense is not external to teamwork, rather it is one of the mechanisms by which shared understanding is formed and maintained [3,51]. This becomes important because evidence suggests that complementary team performance is a key outcome in human–AI systems, outperforming both the human or the AI systems operating independently [55].

This notion of complementary performance reflects the importance of leveraging the distinct strengths of both human and AI agents. Effective X-AI design at the team level enables this synergy by supporting adaptive collaboration, where human agents adjust their strategies in response to AI outputs and feedback [39]. Moreover, because AI systems are not infallible, several studies examine how teams respond to imperfect explanations and recover from AI errors through collective reasoning processes [64]. These collaborative outcomes are typically evaluated using team-level performance metrics such as end-to-end accuracy [55], mean absolute error [65], and coordination efficiency, including time-based measures of task completion [39]. Together, these findings highlight that, at the team level, explainability functions as a mechanism for enabling coordination, adaptive decision-making, and collective performance rather than solely supporting individual understanding.

4.2. From Individual Interpretability to Team Coordination

A central insight from the current study is that explainability operates differently across levels of analysis. At the individual level, X-AI primarily supports comprehension, trust formation, and post hoc justification of decisions. These functions are commonly achieved through techniques such as feature attribution, local explanations, and visualization-based outputs that help users interpret model behavior. SHAP is especially prominent in this context because it provides both local and global feature attributes in a comparatively accessible way. For example, SHAP has been used to identify how variables such as age, nationality, and placement shape performance predictions in sports analytics, providing coaches with interpretable insights into what areas may require attention [52]. In healthcare and decision-support settings, SHAP is similarly used to quantify which factors drive a particular prediction, helping users understand why a case was flagged, prioritized, or classified in a certain way [24].

At the team level, however, the role of explainability expands significantly. A SHAP-based system offers three layers of situation awareness, XAI-1 looks at perception (what?), XAI-2 looks at comprehension (why?), and XAI-3 looks at counterfactuals (what if?) [12]. XAI-1 and XAI-2 functions are more appropriate for individual feedback, while XAI-3 is better served at a team level, providing cross-disciplinary dialog and multilayered explanations [12]. In a care team, a SHAP-based explanation for risk alert may support one clinician's understanding, but its real team value emerges when it helps the group deliberate whether the alert reflects a transient anomaly, a meaningful deterioration, or a case that warrants escalation [51]. In sports, a coach or analyst might examine SHAP outputs to understand why a model predicts low team performance. At the team level, the outputs can support a more distributed conversation among coaches, analysts, and players [22]. This information is valuable to coaches and their staff, allowing them to "target interventions and adjustments to improve the performance of teams" [22] (p. 14). These examples suggest that team-oriented SHAP interfaces may need to support layered access, shared output, and comparability across cases rather than only providing dense, expert-facing feature plots.

Interactive decision support offers a similar contrast. At the individual level, users often benefit from systems that allow them to them test "what if" scenarios, manipulate inputs, or review local explanations on demand [66,67]. At the team level, however, interactive decision support also helps mitigate knowledge imbalance and support collaborative user involvement, allowing different members to interrogate the system from different perspectives [59,68,69]. In team tasks, the explanation process is distributed. One person may compare model outputs to human judgment, another may focus on uncertainty, and another may relate the explanation back to operational constraints. This is especially important because collaborative reasoning has been shown to outperform simple human–AI dyadic reliance in some settings, and teams may use AI more effectively when it serves as a trigger for discussion rather than as an unquestioned item [64].

This shift from individual interpretability to team coordination also has important implications for evaluation. Explanations should not be judged solely by whether they increase subjective trust. The literature repeatedly warns that explanations can also create illusions of understanding, inflate confidence, or encourage overreliance [46,52]. In human–AI teams, one actor's misplaced confidence may propagate to others, amplifying downstream coordination failures. Evaluating X-AI over the long term should involve team member experience factors, such as "over-trust and under-trust on the system" [44] (p. 33). In a study that evaluated published research across multiple disciplines, Mohseni and colleagues [44] evaluated X-AI across five levels: mental model, usefulness and satisfaction, user trust and reliance, human–AI task performance, and computational measures. They

highlighted the need for providing an interdisciplinary effort for evaluating X-AI systems. Other studies identified that the purpose of evaluating X-AI should be to determine whether or not it helped to develop mental models and the identification of discrepancies in knowledge structures [46].

4.3. Explainability as an Interactional and Temporal Process

The findings further suggested that explainability should be conceptualized as an interactional and temporally situated process rather than a static output delivered after model inference. This conclusion directly extends the findings from the interactive and workflow dimensions. It shows that the effectiveness of explainability depends not only on what information is presented but when, where, and how it is encountered within ongoing human activity domains such as management, science, and technology [32].

Interactive approaches shift explainability from a one-time informational artifact to a mechanism for iterative sensemaking. Rather than simply displaying an explanation, such systems allow users to query model outputs, manipulate inputs, compare alternatives, explore counterfactuals, and observe how model recommendations change under different assumptions [61]. This makes explainability functional as opposed to just being representational. Findings also suggested that explanations are collaborative and iterative processes, often involving repeated question–answer exchanges in which participants can refine their understanding [51]. By making it a dynamic experience, users do not simply receive an explanation once; they reflect, reinterpret, and test the outcome as time progresses.

The workflow findings capture how explainability is operationalized across decisions, coordination structures, and broader systems use over time. This trustworthiness is built through an explanation's alignment with decisions to ensure successful transitions of work practices and roles [70]. Explanations must therefore be temporally aligned with the workflow itself. Some are most useful before action, when users are identifying the problem or deciding which decision-support system works best for the stated problem [29]. Some have utility during action to facilitate team discussions [71]. Others are more useful after action, where AI inferences are provided after initial task assessments have been made [72] or as an after-action debriefing [71].

This strengthens our point about iterative sensemaking. Literature related to workflow highlights iterative design–evaluation cycles in which explanation strategies are continuously being refined for effectiveness, trust calibration, and team performance [45]. At the team level, team members interpret AI outputs and compare them against unfolding events, updating the team's understanding through successive cycles of interaction [62,73]. From this perspective, explainability functions as a reoccurring resource embedded in a loop of interpretations and actions rather than as a one-time explanation delivered at the point of prediction [40].

Temporal dimensions of explainability are especially relevant when linked to shared mental models. Shared mental models are not static cognitive artifacts. Instead, they are maintained through perception, communication, inference, and synchronization [51]. Shared mental models can also degrade when they are not refreshed or coordinated through action [51]. This suggests that explainability in human–AI teams should support not only initial understanding but should also provide continuous updating and synchronization across the lifecycle of tasks. Technically accurate but poorly timed explanations can hamper team performance, whereas explanations that are both accurate and timely improve team effectiveness [40,51].

Interaction decision-support examples illustrate this well. For example, users may first receive a recommendation, then decide to explore counterfactuals, and later revisit explanation maps during an after-action review [61,74]. Other studies provided a parallel

processing technique where one team member queries the AI, another contributes contextual knowledge, and the remaining team member determines whether or not to act [51]. Similarly, workflow-oriented studies show evidence that explanations have utility when they are embedded into ongoing work rather than added as an alternative artifact or added after the fact [64].

This interactional and temporal perspective can also clarify why static explanations often fail. A feature ranking, saliency map, or confidence score may be informative in principle. However, if it is delivered at a point when the user cannot meaningfully act on it, it may be ignored or may impose unnecessary cognitive burden [13,45]. Explanations that are delivered too frequently or provide too much information can also be ignored, as shown in studies that looked at information overload during real-time operations [40]. Recommendations are made to provide information that is critical and timely, while preventing cognitive overload of agents.

4.4. Designing for Human–AI Team Performance and Appropriate Reliance

Another key implication of this review is that explainability should be evaluated according to its contribution to team-level performance rather than solely individual understanding. The reviewed literature suggests that effective human–AI teams are characterized by complementary performance, in which the combined capabilities of humans and AI exceed what either could achieve independently. In this context, explainability supports coordination, adaptive collaboration, collective performance, and ethical practice rather than only providing retroactive analyses [51,62,67].

This reinforces the distinction between trust and appropriate reliance. Explanations may increase confidence, but they can also increase overreliance or automation bias if they are overly persuasive or insufficiently transparent about uncertainty [1,64,72,73]. Studies of human–AI decision support show that collaborative user involvement can mitigate knowledge asymmetries and improve decision quality, while imperfect or misleading X-AI can impair judgment and performance [42,55,75]. In some cases, collaborative reasoning among people supported by AI appears to outperform settings where AI is treated as an authoritative teammate whose outputs are simply followed [64]. These findings suggest that the design goal should not be to maximize trust but to support informed and context-sensitive judgments about when and how to rely on AI outputs. In these instances, trust becomes an emergent property dependent on how well the AI supports human agents or team members.

Interactive and workflow-integration approaches appear especially promising. Allowing users to interrogate outputs, compare alternatives, and test assumptions help to promote active engagement with AI recommendations rather than treating them passively [1,76]. In team contexts, this also enables distributed cognition, allowing different team members to contribute complementary capabilities, question the AI from different perspectives, and jointly determine when to defer and when to intervene [7,43,51,63]. This aligns with arguments calling for AI to be treated as a collaborative member embedded within the team rather than as an autonomous actor [64].

4.5. Implications for Theory and Future Research

Findings suggest that explainability should be theorized less as a static property of AI systems and more as a situated feature of human–AI interaction embedded in collaborative teams [5,13]. The literature reviewed for the current study indicated that several X-AI techniques were developed primarily to make model behavior visible to individual users, often through post hoc visualization or local explanations. However, the requirements of human–AI teams introduce a broader set of concerns involving shared mental mod-

els, situation awareness, transactive memory systems, workflow timing, and calibrated reliance [15,40,46,51,77]. Building on these findings, future research should advance theory and empirical work in the following areas.

Before turning to specific directions, it is important to distinguish what this review establishes from what it proposes. The literature-supported findings concern how explainability functions are distributed across individual and team levels, the prevalence of visual and post hoc methods at the individual level, the role of interactive and workflow-integrated approaches in team settings, and documented risks of overreliance on explanations. In contrast, the constructs of shared mental models, distributed cognition, and transactive memory systems were not coded categories in the current review. These constructs were introduced for future research rather than as findings for the current review. The studies we draw on acknowledge that the supporting evidence in human–AI teams is still limited [15,51]. We are stating this clearly so that the framework's team-level claims are read as hypothetical, requiring further testing. The research directions that follow are offered in that spirit.

4.5.1. Team-Oriented Explainability

One theoretical implication of the current study is that current X-AI frameworks remain underdeveloped at the team level. Existing work has made substantial progress in classifying explanation methods, types, and purposes. Unfortunately, most of this progress has been primarily focused on individual comprehension, trust, and prediction capabilities. In contrast, the human–AI teaming literature emphasizes shared mental models, mutual predictability, and complementary performance as defining features of successful teaming [29,55]. This suggests a need for theory that explicitly explains how X-AI functions as a coordination resource within teams rather than as an interpretive aid for individual users.

Future research should concentrate on collaborative X-AI (C-XAI) artifacts that support team cognition (distributed cognition). Andrews et al. [51] argued that shared mental models in human–AI teams remain theoretical, lacking consistent conceptualization and measurement. Similarly, Merry and colleagues [46] suggested that explainability must be defined relative to audience, language, and purpose, making team-level explainability significantly different compared to individual-level explainability. Endsley's [40] work further supports this position by arguing that transparency and explainability are valuable if they support situation awareness and effective human–AI coordination dynamically. Together, these sources support the need for a more explicit theory of team-level explainability grounded in team cognition, role interdependence, and collective action.

4.5.2. Distributed Cognition

Future research should be directed toward explanations of distributed cognition in team and collaborative settings. The literature reviewed repeatedly highlighted shared mental models, transactive memory systems, and distributed cognition as relevant constructs. However, these concepts and models lack empirical testing. This creates a gap between conceptualized and realized explainability models for teams. Future studies should investigate whether explanations help teams build shared cognitive knowledge structures that are both accurate and similar. Other research should look at how knowledge and understanding are distributed across team members. Andrews et al. [51] noted that shared mental models provide a strong conceptual bridge between human factors and X-AI. However, they also emphasized that empirical evidence is limited to fully support such models. Bienefeld et al. [15] added that human–AI teaming can be strengthened through transactive memory systems and speaking up; however, they indicated that research on

X-AI and human–AI teams is lacking. They recommended further research around transactive memory systems to better assess team members' interpretations of AI content and its impact at the team level. The literature suggests further research around distributed cognition relating to explainability and its impact at the team level.

A related direction concerns the cognitive capacity of the non-human members of the team. This review treated explainability as a property that makes AI understandable to its human teammates, but a distributed cognition account also implies the reverse: the AI and its supporting tools need some working model of the task, the context, and the human collaborator. This is the focus of the emerging literature on cognitive digital twins, which extend conventional digital twins with reasoning, learning, and semantic knowledge so that they can adapt and support decision-making rather than only mirror the systems they represent [78,79]. In human-aware applications, a cognitive digital twin of the human collaborator anticipates the person's situational understanding and intentions, complementing the twin of the technical system to form a more integrated human–machine team [78,79]. This work maps onto the interactive workflow quadrant of the I-D framework: explanation makes the AI understandable to the human, while the AI's own modeling makes the human and the task understandable to the AI, and effective team cognition likely depends on both. Integrating cognitive digital twin research with X-AI is therefore a promising direction for extending the team-level account developed here.

4.5.3. Limitations and Status of the Framework

Several limitations could be noted, and together they define the current status of the I-D framework as a conceptual, organizing model rather than as empirically validated theory. First, the framework was derived inductively from a synthesis of the literature and has not yet been tested through case studies, expert evaluation, Delphi panels, survey data, or controlled experiments. The quadrants and worked examples (Section 3.5) demonstrate that the framework can organize and classify existing systems, but this does not constitute validation. Future research should evaluate the framework empirically. For example, this could be achieved through an expert Delphi study assessing the completeness and discriminant validity of the two dimensions or through experiments testing whether quadrant-aligned design choices improve team-level outcomes.

Second, as noted in the methodology section, classifications were established through structured consensus rather than independent double-coding, so formal inter-rater reliability coefficients were not computed. Third, the search was scoped to two core constructs and did not incorporate adjacent terms such as "interpretable SI" or "human-autonomy teaming," which may have excluded some relevant studies. Finally, several team-level implications rest on constructs (e.g., shared mental models, distributed cognition) that were used as interpretive lenses and remain empirically underdetermined in human–AI settings. We therefore present the I-D framework as a foundation for theory-building and a guide for practice, while explicitly highlighting that further validation and testing are required.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/systems14070862/s1, "PRISMA_2000_checklist" and "X-AI PRISMA Figure".

Author Contributions: Contributions by the authors of the current article were composed in the following: conceptualization, J.T., H.P.S. and Y.J.; methodology, H.P.S.; software, J.T. and H.P.S.; validation, J.T., H.P.S. and X.X.; data curation, H.P.S., H.K., J.D. and X.X.; writing—original draft preparation, J.T., H.P.S., H.K., J.D. and X.X.; writing—review and editing, J.T., H.P.S. and Y.J.; supervision, J.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Any data collected can be requested from the corresponding author by email (John Turner).

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

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial Intelligence
H-AI	Human Artificial Intelligence
X-AI	Explainable Artificial Intelligence

References

1. Nourani, M.; Roy, C.; Block, J.E.; Honeycutt, D.R.; Rahman, T.; Ragan, E.D.; Gogate, V. On the importance of user backgrounds and impressions: Lessons learned from interactive AI applications. *ACM Trans. Interact. Intell. Syst.* **2022**, *12*, 28. [CrossRef]
2. Vössing, M.; Kühl, N.; Lind, M.; Satzger, G. Designing transparency for effective human-AI collaboration. *Inf. Syst. Front.* **2022**, *24*, 877–895. [CrossRef]
3. Hauptman, A.I.; Schelble, B.G.; Duan, W.; Flathmann, C.; McNeese, N.J. Understanding the influence of AI autonomy on AI explainability levels in human-AI teams using a mixed methods approach. *Cogn. Technol. Work* **2024**, *26*, 435–455. [CrossRef]
4. Kolajo, T.; Daramola, O. Human-centric and semantics-based explainable event detection: A survey. *Artif. Intell. Rev.* **2023**, *56*, 119–158. [CrossRef]
5. McNeese, N.J.; Schelble, B.G.; Canonico, L.B.; Demir, M. Who/What Is my teammate? Team composition considerations in human-AI teaming. *IEEE Trans. Hum. Mach. Syst.* **2021**, *51*, 288–299. [CrossRef]
6. Riedl, M.O. Human-centered artificial intelligence and machine learning. *Hum. Behav. Emerg. Technol.* **2019**, *1*, 33–36. [CrossRef]
7. Spitzer, P.; Holstein, J.; Hemmer, P.; Vössing, M.; Kühl, N.; Martin, D.; Satzger, G. Human delegation behavior in human-AI collaboration: The effect of contextual information. *Proc. ACM Hum. Comput. Interact.* **2025**, *9*, CSCW101. [CrossRef]
8. Westphal, M.; Hemmer, P.; Vössing, M.; Schemmer, M.; Vetter, S.; Satzger, G. Towards Understanding AI Delegation: The Role of Self-Efficacy and Visual Processing Ability. *ACM Trans. Interact. Intell. Syst.* **2025**, *15*, 5. [CrossRef]
9. Amir, O.; Doshi-Velez, F.; Sarne, D. Summarizing agent strategies. *Auton. Agents Multi-Agent Syst.* **2019**, *33*, 628–644. [CrossRef]
10. Gomez, C.; Unberath, M.; Huang, C.M. Mitigating knowledge imbalance in AI-advised decision-making through collaborative user involvement. *Int. J. Hum.-Comput. Stud.* **2023**, *172*, 102977. [CrossRef]
11. Marrone, R.; Zamecnik, A.; Joksimovic, S.; Johnson, J.; De Laat, M. Understanding student perceptions of artificial intelligence as a teammate. *Technol. Knowl. Learn.* **2025**, *30*, 1847–1869. [CrossRef]
12. Cabour, G.; Morales-Forero, A.; Ledoux, É.; Bassetto, S. An explanation space to align user studies with the technical development of Explainable AI. *AI Soc.* **2023**, *38*, 869–887. [CrossRef]
13. Duan, W.; Zhou, S.W.; Scalia, M.J.; Yin, X.Y.; Weng, N.; Zhang, R.H.; Freeman, G.; McNeese, N.; Gorman, J.; Tolston, M. Understanding the evolvement of trust over time within human-AI teams. *Proc. ACM Hum. Comput. Interact.* **2024**, *8*, 521. [CrossRef]
14. Hagras, H. Toward Human-Understandable, Explainable AI. *Computer* **2018**, *51*, 28–36. [CrossRef]
15. Bienefeld, N.; Kolbe, M.; Camen, G.; Huser, D.; Buehler, P.K. Human-AI teaming: Leveraging transactive memory and speaking up for enhanced team effectiveness. *Front. Psychol.* **2023**, *14*, 1208019. [CrossRef] [PubMed]
16. McNeese, N.J.; Demir, M.; Cooke, N.J.; She, M. Team situation awareness and conflict: A study of human–machine teaming. *J. Cogn. Eng. Decis. Mak.* **2021**, *15*, 83–96. [CrossRef]
17. Schelble, B.G.; Flathmann, C.; Macdonald, J.P.; Knijnenburg, B.; Brady, C.; McNeese, N.J. Modeling perceived information needs in human-AI teams: Improving AI teammate utility and driving team cognition. *Behav. Inf. Technol.* **2025**, *44*, 2069–2092. [CrossRef]
18. Shafiabady, N.; Akume, D.; Haghighat, M.; Ud Din, F.; Kabir, S.; Karim, A.; Zhou, J.; Alsharaydeh, E. Explainable AI for mortality prediction: A comparative study using the MIMIC-III dataset. *BMJ Health Care Inform.* **2026**, *33*, e101406. [CrossRef] [PubMed]
19. Nguyen, P.A.; Din, F.U.; Krug, M.; Jones, R. Use of explainable AI (xAI) in dementia detection and prognosis: A scoping review. *BMC Med. Inform. Decis. Mak.* **2026**, *26*, 43. [CrossRef] [PubMed]
20. Macaulay, A.; Din, F.U.; Nguyen, P.A.; Chiong, R.; Tully, P.J. Using explainable AI for assessment of depression: A systematic literature review. In *Next-Gen Healthcare*; Khalifa, N.E.M., Taha, M.H.N., Eds.; Springer: Berlin/Heidelberg, Germany, 2026.
21. Hettikankanamage, N.; Shafiabady, N.; Chatteur, F.; Wu, R.M.X.; Ud Din, F.; Zhou, J. Explainable Artificial Intelligence (XAI): A Systematic Review for Unveiling the Black Box Models and Their Relevance to Biomedical Imaging and Sensing. *Sensors* **2025**, *25*, 6649. [CrossRef] [PubMed]

22. Moustakidis, S.; Plakias, S.; Kokkotis, C.; Tsatalas, T.; Tsaopoulos, D. Predicting Football Team Performance with Explainable AI: Leveraging SHAP to Identify Key Team-Level Performance Metrics. *Future Internet* **2023**, *15*, 174. [CrossRef]
23. Puram, P.; Roy, S.; Srivastav, D.; Gurumurthy, A. Understanding the effect of contextual factors and decision making on team performance in Twenty20 cricket: An interpretable machine learning approach. *Ann. Oper. Res.* **2023**, *325*, 261–288. [CrossRef]
24. Silva-Aravena, F.; Núñez Delafuente, H.; Gutiérrez-Bahamondes, J.H.; Morales, J. A hybrid algorithm of ML and XAI to prevent breast cancer: A strategy to support decision making. *Cancers* **2023**, *15*, 2443. [CrossRef] [PubMed]
25. Wulff, K.; Finnestrand, H. Creating meaningful work in the age of AI: Explainable AI, explainability, and why it matters to organizational designers. *AI Soc.* **2024**, *39*, 1843–1856. [CrossRef]
26. Silva, A.; Schrum, M.; Hedlund-Botti, E.; Gopalan, N.; Gombolay, M. Explainable Artificial Intelligence: Evaluating the Objective and Subjective Impacts of xAI on Human-Agent Interaction. *Int. J. Hum. Comput. Interact.* **2023**, *39*, 1390–1404. [CrossRef]
27. Senoner, J.; Schallmoser, S.; Kratzwald, B.; Feuerriegel, S.; Netland, T. Explainable AI improves task performance in human–AI collaboration. *Sci. Rep.* **2024**, *14*, 31150. [CrossRef] [PubMed]
28. Taj, S.; Daudpota, S.M.; Imran, A.S.; Kastrati, Z. Aspect-based sentiment analysis for software requirements elicitation using fine-tuned bidirectional encoder representations from transformers and explainable artificial intelligence. *Eng. Appl. Artif. Intell.* **2025**, *151*, 110632. [CrossRef]
29. Subramanian, H.V.; Canfield, C.; Shank, D.B. Designing explainable AI to improve human-AI team performance: A medical stakeholder-driven scoping review. *Artif. Intell. Med.* **2024**, *149*, 102780. [CrossRef] [PubMed]
30. Barredo Arrieta, A.; Díaz-Rodríguez, N.; Del Ser, J.; Bennetot, A.; Tabik, S.; Barbado, A.; Garcia, S.; Gil-Lopez, S.; Molina, D.; Benjamins, R.; et al. Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf. Fusion* **2020**, *58*, 82–115. [CrossRef]
31. Tabrez, A.; Leonard, R.; Hayes, B. Single-shot policy explanation to improve task performance via semantic reward coaching. *Neural Comput. Appl.* **2025**, *37*, 22315–22337. [CrossRef]
32. Garibay, O.O.; Winslow, B.; Andolina, S.; Antona, M.; Bodenschatz, A.; Coursaris, C.; Falco, G.; Fiore, S.M.; Garibay, I.; Grieman, K.; et al. Six Human-Centered Artificial Intelligence Grand Challenges. *Int. J. Hum. Comput. Interact.* **2023**, *39*, 391–437. [CrossRef]
33. Perlmutter, M.; Gifford, R.; Krening, S. Impact of example-based XAI for neural networks on trust, understanding, and performance. *Int. J. Hum. Comput. Stud.* **2024**, *188*, 103277. [CrossRef]
34. Liberati, A.; Altman, D.G.; Tetzlaff, J.; Mulrow, C.; Gotzsche, P.C.; Ioannidis, J.P.A.; Clarke, M.; Devereaux, P.J.; Kleijnen, J.; Moher, D. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: Explanation and elaboration. *PLoS Med.* **2009**, *339*, b2700. [CrossRef]
35. Page, M.J.; McKenzie, J.E.; Bossuyt, P.M.; Boutron, I.; Hoffmann, T.C.; Mulrow, C.D.; Shamseer, L.; Tetzlaff, J.M.; Akl, E.A.; Brennan, S.E.; et al. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *Syst. Rev.* **2021**, *10*, 89. [CrossRef] [PubMed]
36. Booth, A.; Martyn-St James, M.; Clowes, M.; Sutton, A. *Systematic Approaches to a Successful Literature Review*; Sage: Newcastle upon Tyne, UK, 2021.
37. Cheng, C.S.; Behzadan, A.H.; Noshadravan, A. Uncertainty-aware convolutional neural network for explainable artificial intelligence-assisted disaster damage assessment. *Struct. Control Health Monit.* **2022**, *29*, e3019. [CrossRef]
38. Tabrez, A.; Agrawal, S.; Hayes, B. Explanation-based reward coaching to improve human performance via reinforcement learning. In Proceedings of the 2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI), Daegu, Republic of Korea, 11–14 March 2019; pp. 249–257. [CrossRef]
39. Paleja, R.; Ghuy, M.; Arachchige, N.R.; Jensen, R.; BGombolay, M. The utility of explainable AI in ad hoc human-machine teaming. In Proceedings of the 35th Conference on Neural Information Processing Systems (NeurIPS 2021), Atlanta, GA, USA, 6–14 December 2021.
40. Endsley, M.R. Supporting human-AI teams: Transparency, explainability, and situation awareness. *Comput. Hum. Behav.* **2023**, *140*, 107574. [CrossRef]
41. Mylrea, M.; Robinson, N. Artificial intelligence (AI) trust framework and maturity model: Applying an entropy lens to improve security, privacy, and ethical AI. *Entropy* **2023**, *25*, 1429. [CrossRef] [PubMed]
42. Coussement, K.; Abedin, M.Z.; Kraus, M.; Maldonado, S.; Topuz, K. Explainable AI for enhanced decision-making. *Decis. Support Syst.* **2024**, *184*, 114276. [CrossRef]
43. Morrison, K.; Spitzer, P.; Turri, V.; Feng, M.; Kühl, N.; Perer, A. The impact of imperfect XAI on human-AI decision-making. *Proc. ACM Hum. Comput. Interact.* **2024**, *8*, 183. [CrossRef]
44. Mohseni, S.; Zarei, N.; Ragan, E.D. A multidisciplinary survey and framework for design and evaluation of explainable AI systems. *ACM Trans. Interact. Intell. Syst.* **2021**, *11*, 24. [CrossRef]
45. Fernando, N.; Nakisa, B.; Ahmad, A.; Rastgoo, M.N. Adaptive XAI in high stakes environments: Modeling swift trust with multimodal feedback in human AI teams. *arXiv* **2025**, arXiv:2507.21158.

46. Merry, M.; Riddle, P.; Warren, J. A mental models approach for defining explainable artificial intelligence. *BMC Med. Inform. Decis. Mak.* **2021**, *21*, 344. [CrossRef] [PubMed]
47. McGovern, A.; Gagne Ii, D.J.; Wirz, C.D.; Ebert-Uphoff, I.; Bostrom, A.; Rao, Y.; Schumacher, A.; Flora, M.; Chase, R.; Mamalakis, A.; et al. Trustworthy Artificial Intelligence for Environmental Sciences. *Bull. Am. Meteorol. Soc.* **2023**, *104*, E1222–E1231. [CrossRef]
48. Sadeghi, Z.; Alizadehsani, R.; Cifci, M.A.; Kausar, S.; Rehman, R.; Mahanta, P.; Bora, P.K.; Almasri, A.; Alkhawaldeh, R.S.; Hussain, S.; et al. A review of Explainable Artificial Intelligence in healthcare. *Comput. Electr. Eng.* **2024**, *118*, 109370. [CrossRef]
49. Zhang, R.; Duan, W.; Flathmann, C.; McNeese, N.; Freeman, G.; Williams, A. Investigating AI Teammate Communication Strategies and Their Impact in Human-AI Teams for Effective Teamwork. *Proc. ACM Hum. Comput. Interact.* **2023**, *7*, 281. [CrossRef]
50. Zhang, G.L.; Chong, L.; Kotovsky, K.; Cagan, J. Trust in an AI versus a Human teammate: The effects of teammate identity and performance on Human-AI cooperation. *Comput. Hum. Behav.* **2023**, *139*, 107536. [CrossRef]
51. Andrews, R.W.; Lilly, J.M.; Srivastava, D.; Feigh, K.M. The role of shared mental models in human-AI teams: A theoretical review. *Theor. Issues Ergon. Sci.* **2022**, *24*, 129–175. [CrossRef]
52. Eisbach, S.; Mai, O.; Hertel, G. Combining theoretical modelling and machine learning approaches: The case of teamwork effects on individual effort expenditure. *New Ideas Psychol.* **2024**, *73*, 101077. [CrossRef]
53. Chraibi Kaadoud, I.; Bennetot, A.; Mawhin, B.; Charisi, V.; Díaz-Rodríguez, N. Explaining Aha! moments in artificial agents through IKE-XAI: Implicit Knowledge Extraction for eXplainable AI. *Neural Netw.* **2022**, *155*, 95–118. [CrossRef] [PubMed]
54. Yiu, C.Y.; Ng, K.K.H.; Li, X.; Zhang, X.; Li, Q.; Lam, H.S.; Chong, M.H. Towards safe and collaborative aerodrome operations: Assessing shared situational awareness for adverse weather detection with EEG-enabled Bayesian neural networks. *Adv. Eng. Inform.* **2022**, *53*, 101698. [CrossRef]
55. Bansal, G.; Wu, T.; Zhou, J.; Fok, R.; Nushi, B.; Kamar, E.; Ribeiro, M.T.; Weld, D. Does the whole exceed its parts? The effect of AI explanations on complementary team performance. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, Yokohama, Japan, 8–13 May 2021; pp. 1–16. [CrossRef]
56. Jia, S.C.; Li, Z.Y.; Chen, N.; Zhang, J.W. Towards visual explainable active learning for zero-shot classification. *IEEE Trans. Vis. Comput. Graph.* **2022**, *28*, 791–801. [CrossRef] [PubMed]
57. Hemmer, P.; Schemmer, M.; Kuhl, N.; Vossing, M.; Satzger, G. On the effect of information asymmetry in human-AI teams. In Proceedings of the Conference on Human Factors in Computing Systems (CHI 2022), New Orleans, LA, USA, 29 April–5 May 2022.
58. Sanneman, L.; Shah, J.A. Validating metrics for reward alignment in human-autonomy teaming. *Comput. Hum. Behav.* **2023**, *146*, 107809. [CrossRef]
59. Pacia, D.M.; Ravitsky, V.; Hansen, J.N.; Lundberg, E.; Schulz, W.; Bélisle-Pipon, J.-C. Early AI lifecycle co-reasoning: Ethics through integrated and diverse team science. *Am. J. Bioeth.* **2024**, *24*, 86–88. [CrossRef] [PubMed]
60. Sanneman, L.; Shah, J.A. The situation awareness framework for explainable AI (SAFE-AI) and human factors considerations for XAI systems. *Int. J. Hum. Comput. Interact.* **2022**, *38*, 1772–1788. [CrossRef]
61. Gunning, D.; Aha, D.W. DARPA's explainable artificial intelligence program. *AI Mag.* **2019**, *40*, 44–58. [CrossRef]
62. Richter, A.; Schwabe, G. "There is No 'AI' in 'TEAM'! or is there?"—Towards meaningful human-AI collaboration. *Australas. J. Inf. Syst.* **2025**, *29*, 1–11. [CrossRef]
63. Le Guillou, M.; Prevot, L.; Berberian, B. Bringing together ergonomic concepts and cognitive mechanisms for human-AI agents cooperation. *Int. J. Hum. Comput. Interact.* **2023**, *39*, 1827–1840. [CrossRef]
64. Federico, C.; Campagner, A.; Simone, C. The need to move away from agential-AI: Empirical investigations, useful concepts and open issues. *Int. J. Hum.-Comput. Stud.* **2021**, *155*, 102696. [CrossRef]
65. Hemmer, P.; Schemmer, M.; Vössing, M.; Kühl, N. Human-AI Complementarity in Hybrid Intelligence Systems: A Structured Literature Review. In Proceedings of the Pacific Asia Conference on Information Systems (PACIS), Dubai, United Arab Emirates, 12–14 July 2021; pp. 1–14.
66. Herrmann, T.; Pfeiffer, S. Keeping the organization in the loop: A socio-technical extension of human-centered artificial intelligence. *AI Soc.* **2023**, *38*, 1523–1542. [CrossRef]
67. Mallick, R.; Flathmann, C.; Duan, W.; Schelble, B.G.; McNeese, N.J. What you say vs what you do: Utilizing positive emotional expressions to relay AI teammate intent within human-AI teams. *Int. J. Hum.-Comput. Stud.* **2024**, *192*, 103355. [CrossRef]
68. Liang, Q.Y.; Gou, J.Q.; Wang, Z.; Dabic, M. Affordances and constraints of automation and augmentation: Lessons learned from development of a human-AI collaboration business simulation platform. *J. Glob. Inf. Manag.* **2024**, *32*, 1–27. [CrossRef]
69. Umme, H.; Habib, M.K.; Bogner, J.; Fritzsch, J.; Wagner, S. How do ML practitioners perceive explainability? an interview study of practices and challenges. *Empir. Softw. Eng.* **2025**, *30*, 18. [CrossRef]
70. Li, J.; Yeo, R.K. Artificial intelligence and human integration: A conceptual exploration of its influence on work processes and workplace learning. *Hum. Resour. Dev. Int.* **2024**, *27*, 367–387. [CrossRef]

71. Jiehuang, Z.; Han, Y. EID: Facilitating explainable AI design discussions in team-based settings. *Int. J. Crowd Sci.* **2023**, *7*, 47–54. [CrossRef]
72. Gomez, C.; Wang, R.; Breininger, K.; Casey, C.; Bradley, C.; Pavliak, M.; Pham, A.; Yohannan, J.; Unberath, M. Explainable AI enhances glaucoma referrals, yet the human-AI team still falls short of the AI alone. *arXiv* **2024**, arXiv:1904.09829. [CrossRef]
73. Tummala, V.S.; Burris-Melville, T.S.; Eskridge, T.C. AI as a team member: Redefining collaboration. *J. Leadersh. Stud.* **2025**, *18*, 67–80. [CrossRef]
74. Gajcin, J.; Dusparic, I. Redefining Counterfactual Explanations for Reinforcement Learning: Overview, Challenges and Opportunities. *ACM Comput. Surv.* **2024**, *56*, 219. [CrossRef]
75. Montealegre-López, N. Exploring the role of trust in AI-driven decision-making: A systematic literature review. *Manag. Rev. Q.* **2025**. [CrossRef]
76. Ren, Y.Q.; Deng, X.F.; Joshi, K.D. Unpacking human and AI complementarity: Insights from recent works. *Data Base Adv. Inf. Syst.* **2023**, *54*, 6–10. [CrossRef]
77. Chen, A.H.; Lyu, A.R.; Lu, Y.B. Member's performance in human-AI hybrid teams: A perspective of adaptability theory. *Inf. Technol. People* **2024**, *39*, 157–177. [CrossRef]
78. Zheng, X.; Lu, J.; Kiritsis, D. The emergence of cognitive digital twin: Vision, challenges and opportunities. *Int. J. Prod. Res.* **2022**, *60*, 7610–7632. [CrossRef]
79. Liu, Y.; Ji, T.; Guo, X.; Xu, X.; Polzer, J. A survey of cognitive digital twin and the potential use of LLMs. *Manuf. Lett.* **2025**, *44*, 1242–1253. [CrossRef]

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