mathematics

MDPI

Systematic Review

Integrating Explainable Artificial Intelligence in Extended Reality Environments: A Systematic Survey †

Clara Maathuis [1,*], Marina Anca Cidota [2,*], Dragoș Datcu [3] and Letiția Marin [2]

[1] Department of Computer Science, Open University of The Netherlands, 6419 AT Heerlen, The Netherlands
[2] Department of Computer Science, Faculty of Mathematics and Computer Science, University of Bucharest, 010014 Bucharest, Romania; letitia@fmi.unibuc.ro
[3] Independent Researcher, 2628 ZT Delft, The Netherlands; email@dragosdatcu.eu
* Correspondence: clara.maathuis@ou.nl (C.M.); cidota@fmi.unibuc.ro (M.A.C.)

† This paper is an extended version of our published paper: Explainable Artificial Intelligence Techniques for Extended Reality Systems: A Systematic Literature Review. In Proceedings of the International Conference on User-System Interaction, ICUSI, Constanţa, Romania, 19–20 September 2024.

Abstract: The integration of Artificial Intelligence (AI) within Extended Reality (XR) technologies has the potential to revolutionize user experiences by creating more immersive, interactive, and personalized environments. Nevertheless, the complexity and opacity of AI systems raise significant concerns regarding the transparency of data handling, reasoning processes, and decision-making mechanisms inherent in these technologies. To address these challenges, the implementation of explainable AI (XAI) methods and techniques becomes imperative, as they not only ensure compliance with prevailing ethical, social, and legal standards, norms, and principles, but also foster user trust and facilitate the broader adoption of AI solutions in XR applications. Despite the growing interest from both research and practitioner communities in this area, there is an important gap in the literature concerning a review of XAI methods specifically applied and tailored to XR systems. On this behalf, this research presents a systematic literature review that synthesizes current research on XAI approaches applied within the XR domain. Accordingly, this research aims to identify prevailing trends, assess the effectiveness of various XAI techniques, and highlight potential avenues for future research. It then contributes to the foundational understanding necessary for the development of transparent and trustworthy AI systems for XR systems using XAI technologies while enhancing the user experience and promoting responsible AI deployment.

Keywords: explainable AI; responsible AI; trustworthy AI; extended reality; augmented reality; virtual reality

MSC: 68-02; 68T99

Academic Editor: Michael Voskoglou

Received: 10 December 2024
Revised: 11 January 2025
Accepted: 14 January 2025
Published: 17 January 2025

Citation: Maathuis, C.; Cidota, M.A.; Datcu, D.; Marin, L. Integrating Explainable Artificial Intelligence in Extended Reality Environments: A Systematic Survey. *Mathematics* **2025**, *13*, 290. https://doi.org/10.3390/math13020290

1. Introduction

Recent scientific and technological advances in the fields of Artificial Intelligence (AI) and Extended Reality (XR) make it possible to use them in almost all areas of human activity. The level of maturity of these technologies allows them to perfectly intertwine, giving us opportunities to explore ways in which such intelligent systems can address specific challenges in areas like health, the environment, safety, (cyber) security, education, industry, and social media. XR is an umbrella term referring to virtual, augmented, and mixed (VR, AR, MR) reality, covering the whole spectrum of virtual and augmented experiences as described in Figure 1 by Milgram's Reality–Virtuality Continuum (RVC) [1]. It seamlessly

merges the physical and virtual worlds, creating environments where users can interact with computer-generated elements in real-time.

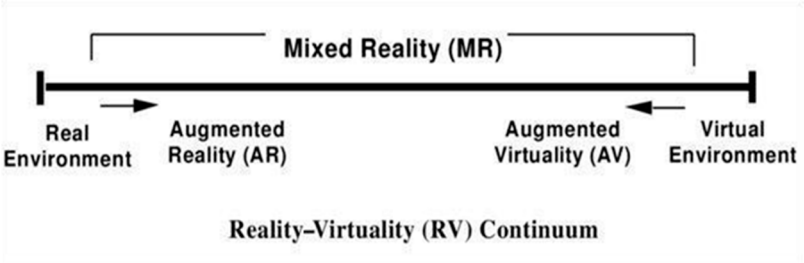

Figure 1. Milgram's Reality–Virtuality Continuum, taken from [1].

The research work on the integration of AI and XR has been widely explored [2–9], proving its capacity to transform many fields by offering user-customizable and engaging experiences, skill augmentation, and new perspectives for innovation. For example, content appropriate to the learning capacity of pupils can be created and presented in an attractive manner for a successful learning session [10]. At the same time, personalized exercises can help patients with motor dysfunction during rehabilitation; adaptive scenarios can be designed for training in various fields in a safe and cost effective environment, with real-time data analysis and automated feedback for evaluation; and virtual stores offer personalized and engaging shopping experiences. The entertainment industry uses XR for immersive gaming and movies, while AI can be used to create XR worlds, making the users' experiences more natural, intuitive, and personalized. Different fields benefit from collaborative XR systems that would allow a remote expert to assist a worker on the spot, while AI can provide customized support and real-time information.

Another important domain that integrates the XR and AI capabilities is digital twins [11], which are virtual representations of a real-world physical entity, system, or process. By becoming smart digital twins that can learn, adapt, and optimize using real-time data, a proper frame for better understanding and innovation in various fields is created.

Nevertheless, when it comes to good interaction between humans and intelligent systems, trust and understanding have an important role. And in more and more countries, there are regulations being introduced to remedy the potential problems that may derive from the opaque behavior of AI systems. For example, the EU has introduced the General Data Protection Regulation [12], which grants individuals the right to receive "...meaningful information of the logic involved" in automated decision-making processes that concern them. Developing techniques by which AI algorithms can explain how they made decisions would overcome such challenges.

In this context, the domain of Explainable Artificial Intelligence (XAI) has emerged with the objective of making AI systems transparent, accountable, and trustworthy.

More precisely, XAI refers to the use of AI models for the extraction of relevant knowledge about domain relationships contained in data. The insights provided in a specific domain problem by this relevant knowledge are often used to guide communication, actions, and discovery [13]. XAI aims to offer explanations in a form that can be easily and clearly understood, thereby closing the gap between AI technology and human comprehension and trust, ultimately ensuring that users can effectively interact with and make informed decisions about AI-driven systems.

The importance of and research interest in XAI is also reflected by a significant number of the existent surveys, as presented in [14]. There are surveys that review explainable methods in a broader context or in a narrower context, focusing on certain domains (e.g., on the healthcare domain), or focusing on topics (e.g., contrastive and counterfactual explanations, or explanations and their evaluation for recommender systems).

In this research, we present a systematic review of papers where XAI methods are used in various tasks in XR environments to acquire an insight into the up-to-date progress, and aim to offer a set of directions for future research to overcome the current limitations that we have identified.

We start by presenting the research methodology that we adopted, where four research questions have been formulated in order to clearly define the key aspects addressed in this research. The Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) framework [15–17] was applied to create a protocol that provides transparency and reproducibility in the selection process of the papers in our systematic review. Thus, we surveyed papers that were published in four scientific databases between the years 2013 and 2023. As Figure 2 shows, the use of XAI in XR environments is still at its beginning, but with an increasing trend.

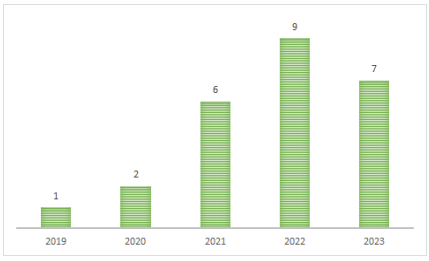

Figure 2. The number of papers with XAI in XR per year.

We identified that XAI was used in different XR applications that required the following tasks: classification [18–28]; regression/forecasting [20,21,27,29,30]; recommendation [31–34]; policy learning [35]; and survival analysis [36].

We describe the domains of application of XR systems that integrate XAI, presenting the benefits and opportunities offered by these technologies when applied in various fields.

We discuss the XR systems integrating XAI, with focus on the hardware, type of data involved, and data collection methods.

The XAI methods are identified within the context of the established taxonomy which comprises two primary distinctions with regard to model dependency (model-based vs. model-agnostic) and the type of explanation provided (local vs. global). Each category is examined, accompanied by examples from the reviewed papers.

The evaluation of XAI methods in XR environments is then presented, with a focus on assessing the explanations. We identify four approaches for evaluating these explanations and provide detailed information and references to additional resources to help researchers apply these methods in their own work. Additionally, we suggest five best practice recommendations for designing, implementing, and assessing XR systems that integrate XAI methods.

We close with a section of discussions and conclusions, where the current limitations are discussed and some directions for future research are proposed.

We can summarize our contribution as follows:

- We provide an overview that facilitates a clear understanding of the synergy between XAI and XR from different perspectives (i.e., domains of application with benefits and opportunities; the types of XR systems integrating XAI; the XAI methods and techniques involved and their evaluation).
- We provide a categorization of XAI methods and of the evaluation techniques, as well a list of five recommendations as a matter of best practice for the evaluation of the explanations.
- We identify the current limitations and propose a set of directions for future research.

2. Research Methodology

This research aims to conduct a systematic literature review and synthesis of existent literature pertaining to the transparency aspect of XR technologies, with a specific focus on the implementation of XAI methodologies and techniques. This exploration is motivated by the increasing prevalence and complexity of XR systems, which necessitates a deeper understanding of their decision-making processes, operational mechanisms, and results provided when embedding various AI mechanisms. By examining the intersection of XR and XAI, this study aims to reflect on current developments, identify potential gaps in the research landscape, and provide a foundation for future research in this rapidly evolving field. To guide this analytical process and further ensure a structured and transparent approach is taken, the PRISMA methodology [15–17] is followed and a series of research questions have been formulated. These questions serve as the framework directing our attention to key aspects of assuring AI transparency in XR technologies and the efficacy of various XAI approaches in addressing these challenges. Through this analytical approach, a contribution is being made by providing valuable insights to the scientific and practitioner communities on the current and potential future developments of more transparent and interpretable XR systems. To this end, the following research questions were formulated:

- RQ1: What are the limitations, benefits, opportunities, and domains of application that integrate XAI into XR systems?
- RQ2: Which XR systems integrate XAI methods and techniques?
- RQ3: What are the XAI methods and techniques used for XR systems?
- RQ4: How are the identified XAI techniques and applications evaluated for XR systems?

The systematic review methodology employed in this study encompasses a series of meticulously planned and executed stages. Initially, the research objectives were clearly defined, establishing the scope and focus of this research. This approach ensures a transparent and reproducible process, allowing for the clear documentation of the literature selection and exclusion criteria, thereby enhancing the validity of the review's findings. Furthermore, the relevant research studies were identified from various scientific databases to ensure a broad and representative sample of the literature. This was followed by an in-depth analysis of the identified studies, involving the extraction of valuable insights and the critical evaluation and synthesis of the findings presented in the selected studies. These steps are further presented and visually represented in Figure 3, which serves as a graphical roadmap of the research process, illustrating the logical progression and interconnectedness of each stage in the systematic review conducted [15–17].

In the first phase, Identification, three main search variables were considered when collecting relevant studies: (i) XR technologies, (ii) XAI methods and/or techniques, and (iii) AI techniques. For each of these search variables, various combinations of keywords depicted in Table 1 were considered for collection. In this process, search queries were formed and executed in the following four scientific databases: IEEE Xplore, ACM Digital Library, Web of Science, and Wiley. Therein, the search was executed between 1 January 2013 and 31 December 2023, and a total number of 6307 records were found, as presented in Table 2. Furthermore, after the removal of duplicates, 5681 unique records remained.

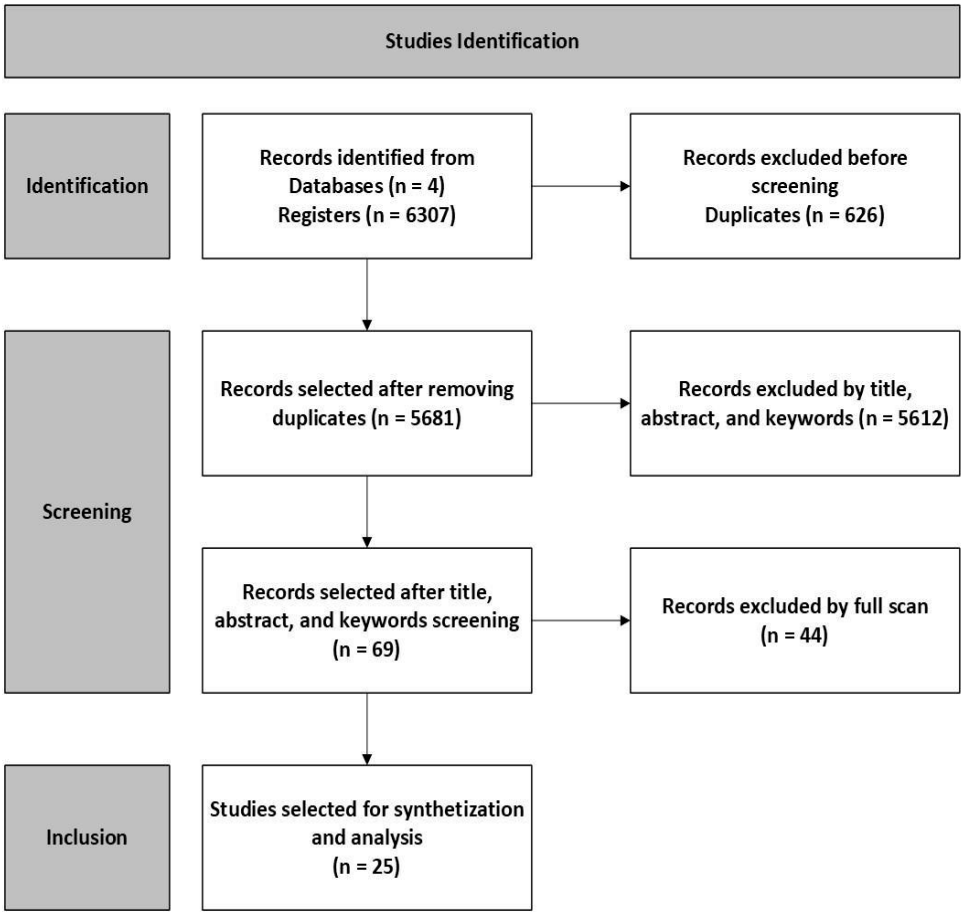

Figure 3. Research methodology diagram.

Table 1. Search variables and keywords.

Database	Results	AI
XR OR Extended Reality	Explainable OR Explainability	Artificial Intelligence OR AI
VR OR Virtual Reality AR OR Augmented Reality MR OR Mixed Reality	Interpretable OR Interpretability XAI	Machine Learning OR ML

Table 2. Database search results.

Database	Results
IEEE Xplore	2378
ACM Digital Library	1124
Web of Science	364
Wiley	2441

In the second phase, Screening, three steps were conducted. In the first step, the unique records were analyzed considering their relevance to the topic and research questions asked based on the screening of the title, abstract, and keywords used. In this step, the inclusion criteria included (i) studies tackling both XAI and XR topics, and (ii) studies being either review or application oriented. This refinement was conducted until there were 69 remaining studies that were further considered for full-study screening. In the second step, these studies were examined considering the same inclusion criteria as in the previous step, with the exclusion criteria of (i) irrelevance to the topic, (ii) a language other

than English, (iii) extended studies, (iv) inaccessibility, and (v) studies that were less than three pages. From applying these criteria, 25 studies remained as the final set of studies to be considered in the systematic literature review.

In the third phase, Inclusion, content analysis was conducted in order to systematically assess and synthesize both the approaches and findings proposed in the final set of studies considered. In this phase, a team of four researchers conducted an in-depth analysis through multiple iterative research meetings. In this process, four thematic groups that match the four research questions considered in these studies were tackled: (i) challenges and opportunities, (ii) XR systems integrating XAI, (iii) XAI methods and techniques, and (iv) evaluation mechanisms considered. Furthermore, cross-checking mechanisms were implemented at various stages in this phase in order to ensure the reliability and validity of the analytical process. The final set of studies included in the review process is presented in Table 3, with their corresponding scientific venues and the distribution of studies depicted in Figure 4 below.

Table 3. Overview of the research studies.

Reference	Research Aim	Venue
[18]	To build an intelligent AR training framework that provides a complete visualization of the Neonatal Endotracheal Intubation (ETI) procedure for the real-time guidance and assessment of trainees.	IEEE ISMAR (International Symposium on Mixed and Augmented Reality) Conference
[19]	To differentiate attention levels that are captured during a perceptual discrimination task presented on two different viewing platforms.	IEEE AIVR (International Conference on Artificial Intelligence and Virtual Reality) Conference
[21]	To develop an XAI-based framework for developing explainable cybersickness-detection machine learning (ML) models for VR systems.	ACM IUI (International Conference on Intelligent User Interfaces) Conference
[22]	To build a biometric identification system that considers the user's gaze behavior and head orientation while following a moving stimulus.	ACM VRST (Symposium on Virtual Reality Software and Technology) Conference
[23]	To predict pedestrian behavior and potential collision situations when crossing the street using ML classification techniques for supporting the design of Autonomous Emergency Braking (AEB) systems in commercial vehicles.	Applied Science Journal
[24]	To build an XAI framework for simulation-based training in surgery.	PloS One Journal
[25]	To investigate users' uni- and bi-manual finger behavior from their interaction with eight different universal interface elements.	Human–Computer Interaction Journal

Table 3. *Cont.*

Reference	Research Aim	Venue
[26]	To understand the classification accuracy of different motor tasks based on brain activity and artifacts in Electroencephalography (EEG) signals for VR neurorehabilitation paradigms.	Sensors Journal
[27]	To develop an XAI-based framework named LiveVR for cybersickness detection.	IEEE VR (International Conference on Virtual Reality and 3D User Interfaces) Conference
[28]	To develop a machine learning approach for predicting teachers' expertise based on off-the-shelf VR device data.	IEEE ISMAR Conference
[29]	To build a multimodal deep fusion system for cybersickness (i.e., motion sickness due to exposure to XR systems) forecasting based on users' physiological, head-tracking, and eye-tracking data.	IEEE ISMAR Conference
[30]	To develop an interpretable machine learning approach for forecasting personal learning performance in VR-based safety training using real-time biometric responses	Automation in Construction Journal
[31]	To explore the design space of mobile AR explanations for movie recommendation scores and provide recommendations for designing AR explanations.	IEEE AIVR Conference
[32]	To explore the design space of mobile AR explanations for movie recommendation scores.	IEEE ISMAR Conference
[33]	To assess the role of explanations for trust enhancement in VR assistance systems.	ACM MC (International Conference on Mensch and Computer) Conference
[34]	To investigate the impact of a smartphone-based AR shopping assistant application that uses personalized recommendations and XAI features related to customer shopping experiences.	Research on Interactive Marketing Journal
[35]	To investigate the use of AR interfaces for live communication between human and robotic teammates in collaborative tasks focusing on prescriptive and descriptive guidance.	AAMAS (International Conference on Autonomous Agents and MultiAgents Systems) Conference
[36]	To investigate pedestrian crossing behavior in the presence of automated vehicles and explore the factors influencing pedestrians' waiting time.	Transportation Research Part C: Emerging Technologies Journal
[37]	To propose ten open challenges, primarily focusing on challenges related to the visualization of medical imaging data, such as explainability and dealing with uncertainty.	IEEE Computer Graphics and Applications Journal
[38]	To develop a novel AR interface that enables the effective diagnosis of a robot's error behavior, enhances its skills, and allows it to upgrade its knowledge structure.	Applied AI Letters Journal

Table 3. *Cont.*

Reference	Research Aim	Venue
[39]	To reveal the potential of VR to overcome most of the desktop eye-tracking limitations and become a disruptive technology for clinical practice in assessing X-ray images.	IEEE VR Conference
[40]	To provide an XAI, blockchain (BC), and immersive technology-based framework for Metaverse healthcare.	IEEE MetaCom (International Conference on Metaverse Computing) Conference
[41]	To discuss the research agenda for guiding the development of the next generation of Discrete Event Simulation (DES) systems for real-time interactive animations used in XR systems while taking a human-centric stance and including XAI methods.	Industrial Information Integration Journal
[42]	To build a lung cancer patient-assistant service named MetaLung on the Metaverse together with an AI-based Decision Support System for treatment selection.	IEEE MetaCom Conference

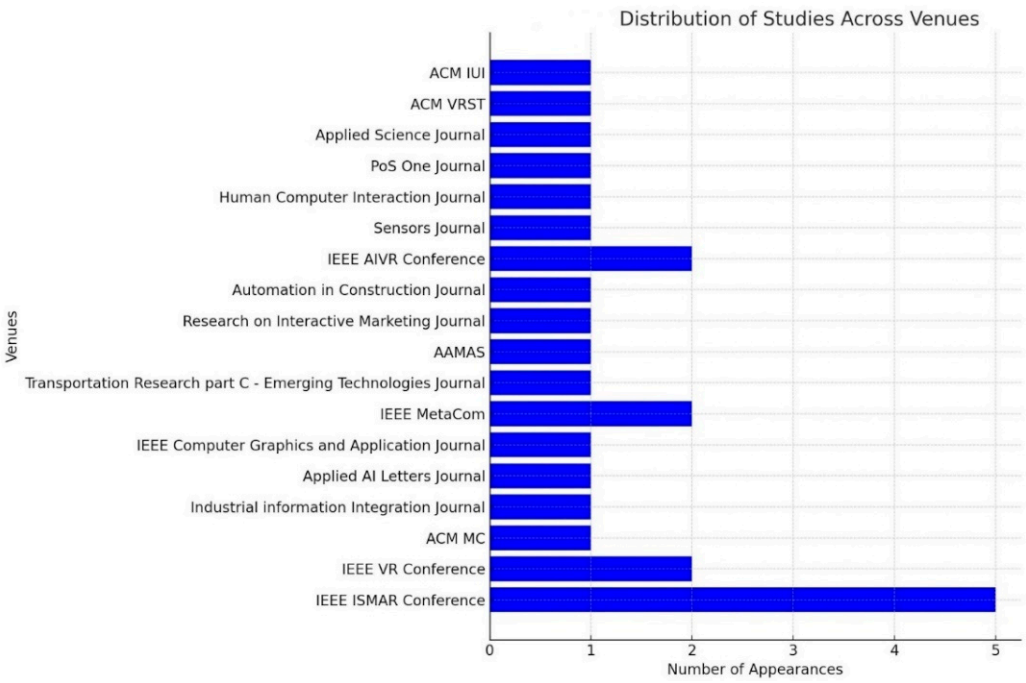

Figure 4. Distribution of studies and venues.

To better understand the current research landscape at the intersection of XR and XAI and identify potential synergies, we have conducted a co-occurrence network analysis of the 30 most relevant technical keywords and their connections, based on the frequencies of occurrence in the title and the abstract of the research studies included in this review process. The resulting network in Figure 5 reveals several distinct clusters of interconnected terms, highlighting the multifaceted nature of research in this area.

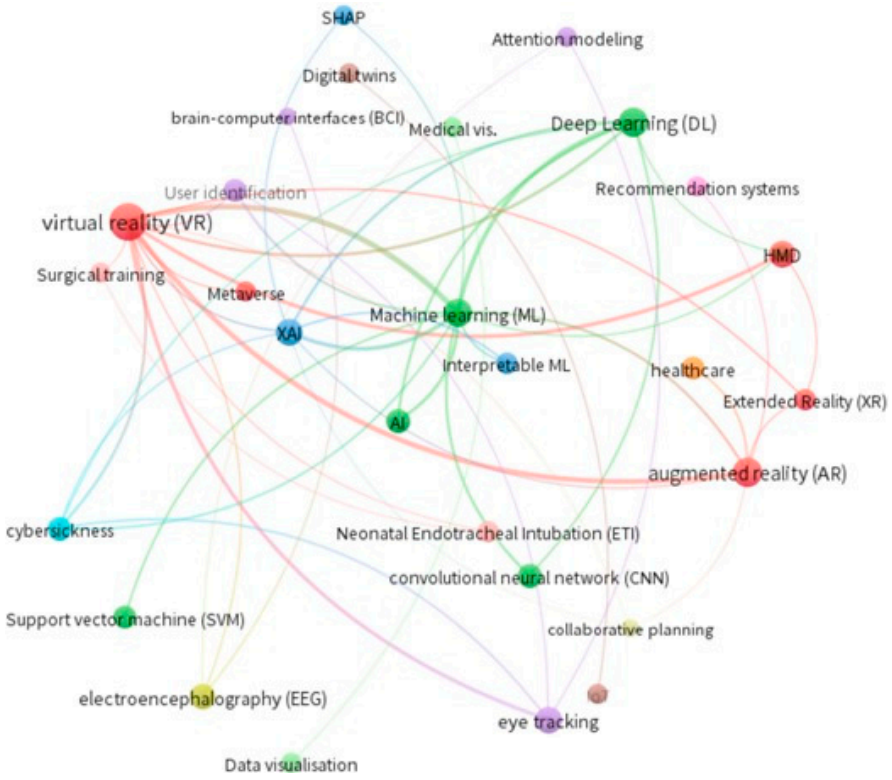

Figure 5. A co-occurrence network highlighting the first 30 most relevant technical keywords from the titles and the abstracts of the reviewed research studies.

3. Domains of Application, Benefits, and Opportunities

The explainability of AI methods significantly enhances their usability and trust across various domains where AI tools are applied. This systematic review identified multiple areas of XAI application, as depicted in Figure 6. The distribution of the application domains shows clearly that healthcare, recommendation systems, and safety capabilities are emerging as the most prominent, while other domains like education and security face existing knowledge gaps.

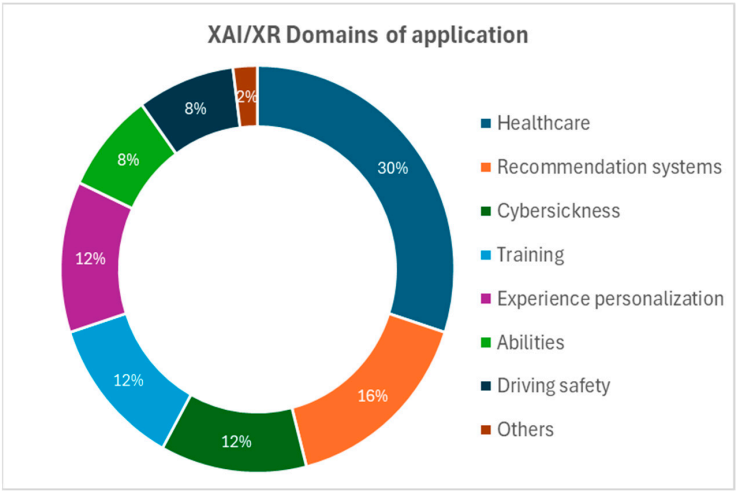

Figure 6. The domains of application of XAI/XR.

Healthcare. As the most extensively studied domain, healthcare constitutes nearly 30% of the reviewed references. This category encompasses a diverse range of applications, including the following: AR- or VR-based training for health practitioners, such as training

for Neonatal Endotracheal Intubation [18]; diagnostic support, like X-ray readings and diagnoses [39] or Metaverse integration for lung cancer detection [42]; surgical educational teaching [24]; the classification of neurological signals, such as motor tasks from EEG signals [26]; innovative medical technologies like Metaverse-integrated clinics and hospitals [40]; and medical data visualization [40].

Recommendation systems. Representing 16% of the reviewed literature, recommendation systems demonstrate XAI's versatility in personalized guidance. Applications span multiple contexts: entertainment-based recommendations, like movie recommendations [31]; action recommendations [35]; and consumer decision support, such as shopping assistance [32,34].

Cybersickness, training, and experience personalization. Each of these domains of application appear in 12% of the articles, highlighting XAI's potential in specialized areas. The cybersickness research topic is focused on detection methods [20,21,27,29]. Training benefits, overlapping with healthcare, include ETI training [18] and training doctors in the Metaverse [40]. VR-based safety training evaluation was studied in the construction field [30]. Experience personalization includes biometrics [22], AR shopping assistance [34], and hand-tracking data for user identification in AR and VR [25].

Abilities and driving safety. Comprising 8% of the papers each, these domains reveal XAI's critical applications in attention assessment [19], expertise evaluation [28], and driving safety innovations, such as collision detection [23] and pedestrian crossing behavior for automated driving [36].

Additional application domains we identified include the following: prescriptive and descriptive guidance [35], human and robot teams [35], and robot improvement through an AR interface [28].

We present the domain of application and the brief benefits for each paper included in our survey in Table 4.

Table 4. Overview of the application domains and benefits.

Reference	Application Domain	Benefits
[18]	Healthcare training	Responding to the increased demand of healthcare providers through the developed framework
[19]	Abilities	Improving attention levels in young adults
[20]	Cybersickness	Allowing the better analysis and design of cybersickness-detection and -reduction models
[21]	Cybersickness	Supporting the development of cybersickness-detection ML models through an (XAI)-based framework for VR-LENS
[22]	Experience personalization/user identification	Introducing a biometric identification system based on the user's gaze behavior
[23]	Driving safety	Using machine learning techniques (classification models) to predict collisions based on several features concerning both the driver and the pedestrians, thus supporting decision-making in AEB (Autonomous Emergency Braking) systems

Table 4. *Cont.*

Reference	Application Domain	Benefits
[24]	Healthcare	Introducing a framework using XAI for simulation-based training in surgery; the proposed framework is validated through an automated platform for educational feedback
[25]	User identification	Distinguishing users of AR and VR devices through finger behavior when interacting with different interface elements
[26]	Healthcare	Comparing the classification accuracy of decoding certain motor tasks based on data in an EEG signal
[27]	Cybersickness	Proposing an XAI-based framework for cybersickness detection, explaining the model outcome and determining the eye-tracking features as being the most important
[28]	Expertise evaluation/robot improvement through AR interface	Predicting teacher expertise based on collected data related to eye-tracking and controller-tracking
[29]	Cybersickness	Forecasting the apparition of cybersickness 60 seconds in advance, through the use of eye-tracking, heart rate, and galvanic skin response data
[30]	VR-based safety training evaluation	Introducing, through an interpretable ML approach, the real-time monitoring and diagnosis of the learning performance of construction workers
[31]	Recommendation systems	Explaining AR models for movie recommendation scores
[32]	Recommendation systems	Explaining the recommendationsprovided by a context-aware AR shopping assistant system
[34]	Recommendation systems/shopping experience	Showing that explainable recommendations provided by a shopping assistant artifact can support the decision-making of customers, thus improving the shopping experience
[35]	Recommendation systems/decision-making	Using prescriptive and descriptive visual guidance in human and robot teams via AR interfaces
[36]	Driving safety	Proposing an interpretable ML framework which allows the exploration of the factors that influence pedestrian behavior (waiting time before crossing) in the presence of automated vehicles
[38]	Robot behavior	Presenting an AR system which allows the diagnosis of erroneous behavior in robots and the improving of their skills, based on interpretable and/or-graph knowledge representation

Table 4. *Cont.*

Reference	Application Domain	Benefits
[39]	Healthcare	Supporting decision models for radiologists
[42]	Healthcare	Proposing a privacy- and integrity-preserving architecture supporting a Metaverse application for lung cancer detection, monitoring, and treatment

Furthermore, next to the benefits provided in relation to the application domain considered, a set of research agenda perspectives was identified and is presented in Table 5 in order to support the formulation of existing knowledge gaps, enhance domain relevance through real-world applications, and encourage cross-domain collaboration. These perspectives include trust in assistance systems [33], Discrete Event Simulation (DES) systems [41], challenges in visualization research [37], and a brief state-of-the-art review of Metaverse healthcare [40].

Table 5. Overview of the studies and research agenda perspectives.

Reference	Application Domain	Benefits/Aim
[33]	Trust in assistance systems	Shows that, even if the explanations can increase trust, users may still not rely on an assistance system. The relationship between trust and reliance requires further research.
[37]	Challenges in visualization research	Presents ten open challenges in the visualization of medical data research, including challenges focusing on explainable AI.
[40]	State-of-the-art review of Metaverse healthcare	Briefly presents the Metaverse technologies for healthcare, also proposing an architecture of Metaverse healthcare based on XAI, blockchain, and immersive technologies.
[41]	Discrete Event Simulation (DES) systems	The discussed agenda aims to guide the development of future DES systems in the context of the human-centric technology approaches which are central to the Industry 5.0 paradigm.

This systematic review reveals diverse and promising opportunities and future perspectives for XAI across multiple domains. These opportunities encompass innovative methodological developments, targeted framework design, and potential research trajectories, with significant implications for various sectors.

The reviewed literature highlights a range of key opportunities, including the development of customized frameworks to address specific industry needs, such as in healthcare [18], and the exploration of innovative methodologies like cybersickness detection using explainable machine learning (xML) [20] or predicting cybersickness severity 90 seconds in advance through deep temporal convolutional forecasting models [29]. Future directions often focus on expanding research to larger populations [20], leveraging AR to enhance

user experiences in areas like movie recommendations [31], and utilizing AR for designing explanation artifacts and visualizations in recommendation systems [32].

Additional opportunities involve exploring diverse applications, such as behavioral biometrics for user authentication and access control [22], or studying human–robot team dynamics in rescue missions, where visual guidance algorithms play a critical role [35]. While some findings may not yet present strong results, they still highlight promising directions for further study, including the relationship between trust and reliance in AI systems [33]. Improving user decision-making through explainable recommendations, as demonstrated in shopping experiences [34], is another notable area of potential. Research addressing sensitive issues like braking assistance and critical concerns such as pedestrian safety also holds significant promise, as shown in [23] and [36]. In the medical field, opportunities include enhancing teaching methods [24]. Medical visualization remains a valuable tool for advancing healthcare [37], while interpretability in diagnosing robot behaviors can improve human trust through generated explanations [28]. Other opportunities include identifying the most relevant feature class for cybersickness, which has been shown to be related to eye-tracking [27], applying XAI predictive models in Metaverse medical applications [42], predicting teacher expertise [28], and delivering personalized safety training for construction workers [30]. Also, AR and VR can be applied in identification and authorization [25], offering significant benefits across various information systems.

From the point of view of the practical approach of the papers, we note that, out of 25 papers, 21 were addressed either theoretically, practically, or both. Specifically, 76% of the total papers were addressed theoretically, and 90% were addressed practically.

4. XR Systems Integrating XAI

Various XR systems integrate methods and techniques related to XAI. They differ on several aspects, e.g., the type of visualization device targeted, whether or not they take extra body measurements, or their application domain.

From the publications studied, six publications highlight AR approaches, fourteen publications present VR-based approaches, while five publications discuss mixed AR/VR systems (Figure 7).

Figure 7. AR, VR, and mixed AR and VR publications.

Considering the subset of publications describing AR systems, four of the systems detailed make use of optical see-through devices, while three of the systems relate to video see-through equipment.

The remaining two AR-related publications do not provide specifications with regard to the type of AR viewer. With regard to the type of equipment, 14 of the research publications target HMD-oriented systems (Head Mounted Display). From the collection of publications focusing on HMDs, ten publications specifically target VR HMDs and three publications tackle AR HMDs, while one publication covers both AR and VR HMDs (Figure 8).

Figure 8. From the publications targeting HMD systems, ten specifically focus on VR HMDs and three focus on AR HMDs, while one publication covers both AR and VR HMDs.

The most used AR device is the Hololens (including versions), appearing in four of the AR-related publications (Figure 9). One publication focuses on the use of monitors and one publication focuses on the use of tablets. Moreover, three publications focus on the use of smartphone devices. From the three publications which focus on smartphones, the most used operating system was Android (two publications). The Oculus headset appears in one publication. At the same time, the HTC Vive (including multiple versions) is the most-employed VR device, used to support the research approaches in seven of the VR publications.

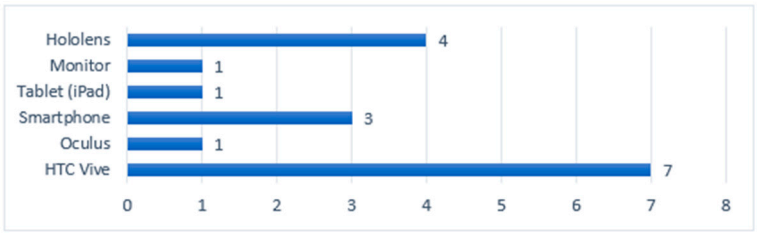

Figure 9. The distribution of device types and device brands in the selected papers. The most used AR device is the Hololens, while the most used VR device is the HTC Vive.

Some of the publications on explainable AI-enhanced techniques for XR in our review study make use of extra sensors (13 research studies). The integration of the extra sensors allows for the analyzing and tracking of user behavior at specific time moments, as well as dynamically. Additional sensors are employed in 3 AR-oriented publications and by 10 VR-oriented publications. The most integrated type of extra sensor is the eye-tracker, in six VR publications (Figure 9). Other sensors, like head-trackers, brain-related sensors (i.e., EEG), or physiological trackers (i.e., related to heart rate, galvanic skin response, breath-rate) each show up in three VR publications on explainable AI techniques for XR systems. Regarding AR systems, three publications employ hand-oriented sensors (Figure 10).

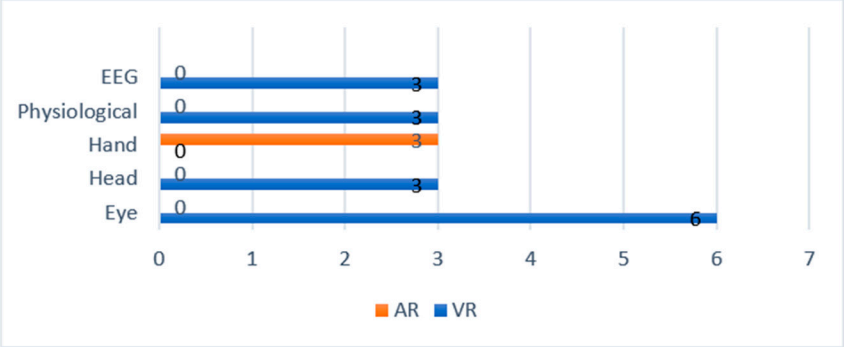

Figure 10. Additional sensors in AR and VR systems.

The integration of eye and head trackers in XR enables user identification through xML models [22,25]. Moreover, eye trackers and physiological sensors used for measuring

and monitoring human physical and emotional states are found to provide the most reliable predictions for cybersickness, based on the XAI analysis [20,21,27,29].

The research findings on explainable AI for XR systems are facilitated through the analysis of data collected during the researchers' own experiments or by using open-source datasets. Data from the researchers' own experiments is employed in 5 publications specifically targeting AR systems, and in 12 publications targeting VR systems (Figure 11). Researchers' own datasets are used in one publication targeting both AR and VR systems.

Figure 11. AR, VR, and AR/VR-focused publications that make use of their own datasets.

The data collection primarily involved (1) conducting user studies and experiments in VR/AR environments; (2) collecting sensor data like eye-tracking, head-tracking, and physiological signals; (3) gathering user interaction and behavior data during specific VR/AR tasks; and (4) using publicly available datasets, in some cases. Sample sizes ranged from 16 to 252 participants (see Table 6). The data were used for various purposes such as cybersickness detection, user identification, pedestrian behavior modeling, expertise assessment, and evaluating AR/VR interfaces and applications.

Questionnaires are typically used in user studies to collect feedback from the user during or after experiments. In our selection of scientific studies, the questionnaire is used as a tool to validate the findings regarding explainable AI techniques for XR systems in three AR-related publications and in five VR-related publications. One publication which reports on AR and VR systems makes use of questionnaires as well (Figure 12). The rest of the publications do not mention using questionnaires in their studies.

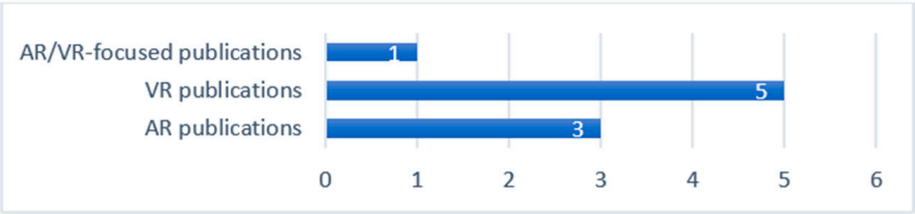

Figure 12. AR, VR, and AR/VR publications including questionnaires.

From our review study, it appears that research on XR systems integrating XAI focuses more on VR technology and HMD setups. The HTC Vive is the most used HMD for VR, while the Hololens is the HMD which empowers most of the AR-oriented platforms. The subset of research studies which employ additional sensors tracking user behavior and the subset which does not employ additional sensors are balanced. However, more than double (in relative percentage) the number of VR systems use extra body sensors, as compared to AR systems. Ideally, more research publications on AR systems integrating tracking sensors to support XAI will be available in the future.

Table 6. A summary of the datasets and data collection methods used in the relevant papers.

Reference	Dataset and Data Collection Method Used
[19]	EEG data collected from 20 healthy young adults performing perceptual discrimination tasks on 2D (PC monitor) and VR (HMD) viewing platforms.
[20]	Physiological and gameplay data collected from publicly available datasets for cybersickness detection and prediction.
[21]	Integrated sensor, gameplay, and bio-physiological datasets used for evaluating cybersickness-detection models.
[22]	Hand-tracking data collected from 16 participants across 2 sessions as they interacted with 8 universal interface elements in AR/VR.
[23]	VR-based experiments with 50 participants to collect motion-tracking data for collision avoidance behavior modeling.
[24]	EEG data collected from 28 skilled and 22 novice participants performing VR subpial brain tumor resection task.
[25]	Tracking data of 16 participants interacting with 8 interface elements in AR/VR collected across 2 sessions for user identification.
[26]	EEG data collected from 17 chronic stroke patients performing 6 motor tasks in VR-based neurorehabilitation setting.
[27]	Integrated sensor data (eye-tracking, head-tracking, physiological signals) used for evaluating cybersickness-detection deep learning models.
[28]	Eye-tracking and controller-tracking data collected from participants in VR-based teacher-expertise assessment study.
[29]	Eye-tracking, head-tracking, heart rate, and galvanic skin response data collected from 30 participants immersed in 5 VR simulations. Fast-motion scale used as ground truth.
[30]	Eye-tracking and EEG data collected in real-time from 30 construction workers during VR-based safety training on fall accidents.
[33]	Empirical study in VR environment with 40 participants interacting with public transport assistance system. Trust measured via questionnaire and behavior.
[34]	Online experiment with 252 participants evaluating the impact of AR shopping-assistant application with explainable AI features.
[35]	Data collected from 180 participants in 4 locations through dynamic and immersive VR experiment to explore pedestrian crossing behavior when encountering automated vehicles.
[36]	Dynamic VR experiment with 180 participants to collect behavioral data on pedestrian wait times at crosswalks with automated vehicles present.

Figure 13 illustrates the diagram summarizing the system limitations mentioned in the relevant papers. One aspect to mention is the limited deployment and testing on actual VR/AR devices. Many studies developed models and systems but did not deploy them on real VR/AR headsets [21,27]. Moreover, the lack of diversity in participant demographics seems significant. Several studies acknowledged limitations in participant diversity in terms of age, gender, and background, which could affect generalizability [20,21,36]. The limited scope of application scenarios represents another limitation. Some studies focused

on specific application domains or tasks, and further research is needed to explore the effectiveness of the proposed approaches in broader contexts [24,28,34].

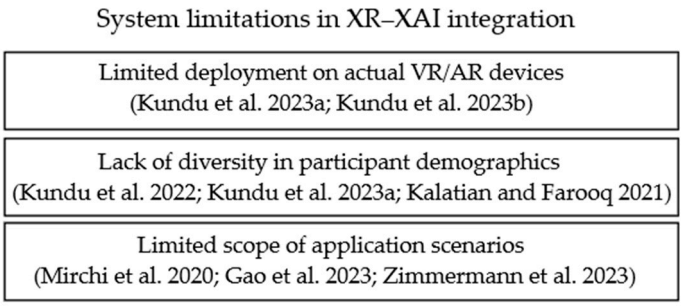

Figure 13. System limitations mentioned in the relevant papers [20,21,24,27,28,34,36].

Equally relevant is the fact that more publications will appropriately be made available on XAI-enhanced XR systems targeting even more applications domains. At the moment, the medical domain is the most explored among the few domains tackled in the existing literature. Surprisingly, the questionnaire-driven method has relatively low incidence for assessing the approaches presented in the selected publications.

5. XAI Methods and Techniques Used

In the context of AI, explainability refers to the capacity to elucidate the internal mechanisms, data used, and decision-making processes of AI models in a manner that is comprehensible to humans. This is crucial for embedding and building trust in AI systems, as it allows users to understand, validate, and effectively manage AI-based decisions. Moreover, explainability involves making the (often) complex and opaque internal mechanisms of AI models transparent and interpretable, which is essential for identifying and mitigating potential issues, errors, or biases within these models. To tackle the explainability of AI models, the field of XAI considers a suite of methods and techniques designed to make the behavior and outputs of AI models more transparent while maintaining high performance standards. This approach aims to bridge the gap between complex AI models and human understanding, facilitating compliance with regulatory requirements and enhancing overall system transparency [43,44].

To reflect on the application of XAI in the XR domain, it is essential to examine the diverse approaches using specific XAI mechanisms within the context of established taxonomic categories. This taxonomy encompasses two primary distinctions considering the model dependency and type of explanation provided: model-based versus model-agnostic methods, where global, local, or a combination of global and local explanations are provided [43]. Each of these categories is further discussed.

Model-based XAI techniques are inherently interpretable due to their specific architectural design or internal structure. These techniques are integrated into the model during its development or training phase, and are tailored to particular AI models by providing insights into the models' inner components and decision-making process. Examples of such techniques include decision trees, Bayesian Belief Networks, and linear regression models [45]. Alternatively, model-agnostic explainable AI methods are distinguished by their ability to be applied to any AI model regardless of the underlying model architecture, providing flexibility. Nevertheless, this implies potentially sacrificing some degree of insight into the model's inner functional mechanisms. Accordingly, an overview of the methods used in the studies analyzed in this research is provided in Table 7, as follows:

Table 7. Overview of the model-based and model-agnostic methods used.

Reference	Explanation Method (Model-Based Versus Model-Agnostic)	Explanation Technique Used
[18]	Model-based and model-agnostic	Attention-based CNN and GRAD-CAM++
[19]	Model-based	EEGNet
[20]	Model-based and model-agnostic	Decision tree, logistic regression, explainable boosting machine
[21]	Model-agnostic	SHAP, MSA, LIME, and PDP
[22]	Model-based	k-NN
[23]	Model-based	Decision tree, random forests, Chi-Square test
[24]	Model-based	SVM
[25]	Model-based	Random forests
[26]	Model-based	TFA, PCA, LDA, SVM
[27,28,30]	Model-agnostic	SHAP
[29]	Model-based	NBEATs
[31,34]	Not mentioned	Not mentioned
[32]	Model-based	FIA
[33]	Manually generated	Manually generated
[35]	Model-based	Markov Decision Process
[36]	Model-agnostic	RRELIEF and SHAP
[37,39–42]	Research agenda/review study	Research agenda/review study
[38]	Model-based	AOG

In the context of model-based explanations, the majority of studies consider the development of methods that relate to baseline ML techniques like SVM and explainable boosting machines, followed by various statistical and visualization techniques like the Chi-Square test and AOG, and a small representation of deep learning methods like attention-based CNN and EFGNet. At the same time, in the context of model-agnostic explanation techniques based on decision trees, LIME (Local Interpretable Model-agnostic Explanations), and SHAP (Shapley Additive exPlanations) are mainly used. Coupled with these, global and local explanations have complementary roles in enhancing the interpretability of AI models. Global explanations offer a comprehensive view of the behavior of the AI model across the entire input space, while local explanations focus on individual predictions from specific instances in order to explain why the model makes a particular decision. Accordingly, LIME is seen as a prime example of an XAI method that provides local explanations, while SHAP is considered an XAI method that is able to provide both local and global explanations. Specifically, LIME operates by creating a simplified, interpretable model that approximates the complex AI model's behavior in the vicinity of a specific input, thereby showing the feature importance for that particular instance. SHAP employs game theory principles, specifically Shapley values, to quantify feature contributions. It does that by considering all possible feature combinations to assign importance values, providing a more rigorous approach for providing explanations [46,47]. The distribution of the types of explanations used is depicted in Figure 14.

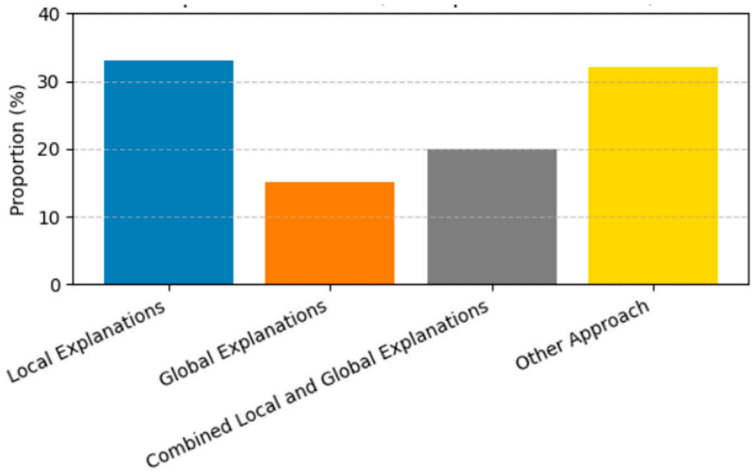

Figure 14. Types of explanations used.

For model-based approaches, as shown in the word cloud captured in Figure 15, [20] considers logistic regression and EBM approaches for model interpretation in the context of cybersickness classification. For logistic regression, global explanations are derived from feature importance, highlighting the significance of physiological measures in predicting cybersickness. However, the model's reliance on similar features for both positive and negative classifications suggests potential accuracy issues. EBM provides a more nuanced global explanation using SHAP values to rank feature importance, with user-specific attributes like vision problems and demographics emerging as key predictors. Ref. [25] presents a random forests-based approach where explanations are provided in the context of EEG signal classification for movement detection through permutation feature importance. This approach assesses the impact of each feature on model performance by randomly shuffling feature values and observing the resulting change in prediction accuracy. To enhance interpretability, a feature-wise mean value is calculated across aggregate functions and user interface elements, resulting in a ranked list of the ten most important features. This is particularly relevant in the context of Brain–Computer Interface (BCI) applications, where distinguishing between genuine brain signals and movement-related artifacts is crucial. For educational purposes, [24] proposes a SVM model for a Virtual Operative Assistant developed for surgical skill assessment. The model provides transparency into its decision-making process by revealing the contribution of individual metrics to the final classification. Herein, the model's output is determined by a series of inputs and their corresponding weights, with a positive outcome optimized when each metric–weight combination is positive. This feature allows for a granular analysis of performance, making the decision boundary interpretable, and enhancing the model's utility as a teaching tool. Furthermore, [22] implements a kNN explainable classifier (with $k = 1$) for participant identity classification which demonstrates a balance between model simplicity and interpretability in the context of gaze- and head-orientation data analysis. This approach aggregates complex time-series data into summary statistics, creating a unified feature set of 21 features across 11 participants. The use of arithmetic means as a summary statistic for each gaze feature allows for the standardization of varying amounts of gaze behaviors across participants, enabling direct comparisons between recordings. This explainability is particularly valuable in understanding the model's decision-making process, as it allows for the examination of nearest neighbors, feature contributions, and individual instance analysis. The transparency of the 1NN model provides insights into which gaze- and head-orientation features are most influential in distinguishing between participants by providing a clear trade-off between model complexity and interpretability.

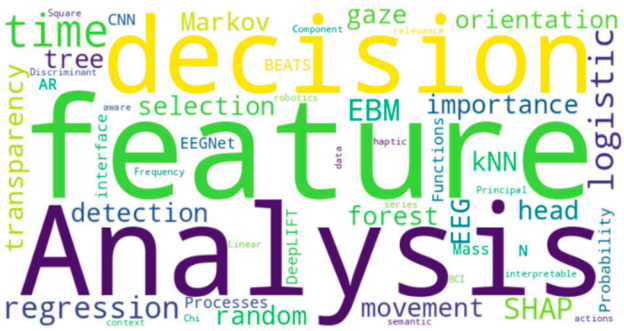

Figure 15. Model-based approaches word cloud.

In the field of human–robot collaboration, [35] builds a model based on Markov Decision Processes (MDPs) and Probability Mass Functions (PMFs) to create a transparent and interpretable decision-making framework. The system, known as MARS (Min-entropy Algorithm for Robot-supplied Suggestions), utilizes a two-tiered MDP structure: one for the robot's actions and another for human recommendations. The robot's MDP is parametrized by a continuously updated PMF, which represents the probabilistic distribution of potential mine locations. This PMF is dynamically revised based on new observations, ensuring that the model's understanding of the environment evolves in real-time. The human recommendation MDP, solved using the updated PMF, generates actionable guidance for the human operator. Then, the system's explainability is enhanced through an AR interface, which visually communicates both the PMF and action recommendations to the human agent. This multifaceted perspective not only facilitates effective decision-making under uncertainty but also provides a clear, interpretable representation of the AI's reasoning process.

Based on deep learning techniques, [19] builds a model for EEG feature extraction and visualization using EEGNet, a lightweight 2D CNN, to extract features from preprocessed EEG time-series data. This model is then applied to classify the extracted features into either two classes (2D/VR-induced attention) or four classes (2D-Target, VR-Target, 2D-Distractor, and VR-Distractor). To provide interpretability to the model proposed, the DeepLIFT (Deep Learning Important FeaTures) method is implemented for feature visualization. Hence, the combination of EEGNet for feature extraction and DeepLIFT for visualization allows the understanding of the neural correlates of attention in different platforms. Also based on CNNs, [18] builds an attention-based dilated CNN framework for the real-time performance evaluation and localization of motion regions in Endotracheal Intubation (ETI) procedures. This technique integrates the model's ability to focus on discriminative motion patterns within kinematic multi-dimensional time-series (MTS) data. The network architecture, comprising five convolutional modules and utilizing Global Average Pooling (GAP), is designed to process input data and output a three-level score label. The attention mechanism, a key feature of this model, learns a weight vector that amplifies the contribution of highly correlated regions, thereby creating a more representative feature space. This approach allows the model to differentiate between relevant and irrelevant movements in the ETI procedure, which is important for accurate skill assessment. Then, it provides local explanations by highlighting important regions of the input data that contribute to specific predictions. This capability is valuable in the context of AR feedback, as it enables the localization of undesirable motion patterns and facilitates skill improvement in medical training scenarios. In the context of time-series forecasting, [29] provides a Neural Basis Expansion Analysis for Interpretable Time Series (N-BEATS) model based on residual blocks with 128 hidden layers with ReLU activation functions, contributing to the model's ability to capture complex non-linear patterns. The final output is derived from the sum of all

the residual block outputs, allowing for a decomposition of the forecast into interpretable components. This architecture enables N-BEATS to supply local explanations by analyzing the contributions of individual basis functions to specific predictions. Unlike traditional black-box models, N-BEATS achieves superior accuracy while maintaining interpretability, making it particularly valuable in domains where understanding the underlying factors driving predictions is important.

Ref. [32] proposes a context-aware AR-based explainable system for implementing context-aware explanations tailored to specific AR scenarios and the utilization of feature-importance analysis visualizations. The research explores various visualization strategies, including color-coding and size-scaling, to effectively communicate recommendations to users. Although these presentation methods do not definitively indicate the underlying XAI approach, the emphasis on context-specific explanations and feature importance suggests a close integration with the model's internal mechanisms. This context-aware XAI approach in AR systems represents a significant step towards more transparent and user-friendly decision support tools in AR environments. For predicting vehicle-to-pedestrian collisions in VR scenarios, the Chi-Square test is applied during the data preprocessing phase. This method evaluates the independence and relationships between categorical explanatory variables and the response variable, providing valuable insights into feature relevance. By assessing the significance of variables such as reaction type and pedestrian behavior, the Chi-Square test aids in informed feature selection, ensuring that critical predictors are retained in the model in a clear and interpretable way [23].

In robotics, the AOG technique is used to create a comprehensive knowledge representation by combining semantic action information with haptic data. This facilitates providing real-time explanations of robot behaviors, enhancing human–robot interaction and trust. The AOG's hierarchical structure, integrating logic and graphics, captures complex relationships between entities and concepts, offering a more nuanced understanding of robot decision-making processes. This reflects the importance of detailed, real-time explanations in human–robot collaboration. Furthermore, the integration of an AR interface with the AOG allows users to visualize the robot's knowledge structure, diagnose issues, and interactively teach new actions. This combination demonstrates the interpretability of the knowledge representation and the importance of enabling users to understand why and how robots behave and act. In the context of BCIs (Brain–Computer Interfaces) used for neurorehabilitation, [26] considers the use of various interpretable methods based on the interpretation of weight vectors from Time-Frequency Analysis (TFA), visualization of task-based topographies, and projection of binary weight vectors back to sensor space that allow an understanding of which features are most crucial for task classification and the visualization of the sensor space for different tasks. Moreover, interpretable ML methods for dimensionality reduction and classification such as Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) are used.

For model-agnostic approaches, as illustrated in the word cloud captured in Figure 16, [20] develops a decision tree model for cybersickness predictions that benefits from the inherent interpretability through its own hierarchical structure, where each node represents a feature and each branch a decision rule. This is exemplified in the global feature-ranking visualization described, where the auto-movement of camera feature is identified as the root node for 11,088 observations, equally split between cybersickness and no-cybersickness cases. The tree's structure allows for the easy inference of feature importance, as features leading to better splits are naturally positioned higher in the hierarchy. Simultaneously, the model provides explanations by providing a clear logical flow of decision-making that can be understood without deep knowledge of the underlying model. In [18], a Grad-CAM++ visualization explanation technique is proposed for the

CNN model in the context of Endotracheal Intubation (ETI) training. This method implements a weighted function for gradient components, allowing for improved sensitivity in detecting multiple discriminative patterns within kinematic multi-dimensional time-series (MTS) data. This generates heat maps highlighting the regions of the input data most influential to the CNN's predictions. Nevertheless, the interpretation of these discriminative regions requires careful consideration, as they may not directly correspond to positive or negative performance indicators. To tackle this issue, expert-based ETI scoring rubrics are considered, focusing on motion trajectory proficiency. This combination allows for the translation of the statistical patterns identified by Grad-CAM++ into meaningful feedback on desirable (smooth and stable) and undesirable (repositioning and rocking) movements. By providing both global insights into the model's decision-making process and local explanations for specific predictions, Grad-CAM++ enables trainees to focus on critical aspects of their ETI procedures, thereby facilitating targeted improvement and enhancing the overall effectiveness of the training process.

Figure 16. Model-agnostic approaches word cloud.

Ref. [27] builds a deep SHAP method for computing values from deep learning-based cybersickness models. Specifically, it is used for feature reduction where features are ranked based on their importance scores. This approach enables the selection of the most influential features (e.g., top one-third) for model retraining, resulting in more efficient models with fewer parameters and faster training and inference times. By bridging the gap between complex deep learning models and interpretable predictions, SHAP enhances the understanding and potential refinement of cybersickness-detection systems. In an educational setting, [28] implements SHAP for distinguishing teacher expertise through eye-movement analysis. This approach provides local explanations by computing Shapley values for individual features, revealing their contributions to the model's predictions. Specifically, the model illustrates the significance of specific eye-movement patterns, such as fixations on identified disruptive students, in differentiating between novice and expert teachers. By providing interpretable insights into subconscious human behaviors during teaching tasks, SHAP not only enhances confidence in the model's validity, but also provides valuable clues for model improvement. The local nature of these explanations allows for a nuanced understanding of how particular features influence individual predictions, a fact that facilitates having a detailed view of the decision-making process. Furthermore, [30] proposes a SHAP approach for interpreting and enhancing forecast models in VR-based construction safety training. Notably, SHAP's interpretability proves crucial in analyzing complex biometric data, such as eye-tracking and EEG measurements, collected during VR training. This comprehensive approach not only enhances model performance but also provides actionable insights into the relationships between learner characteristics and training outcomes, facilitating informed decision-making in educational contexts. Using both LIME and SHAP methods, [40] provides explanations for allowing healthcare professionals to discern how specific features, such as patient symptoms and

test results, influence outcomes. This transparency is crucial in clinical settings, where understanding the rationale behind AI-driven decisions can significantly impact patient care and treatment strategies. Furthermore, the iterative application of these techniques allows for continuous feedback on and the refinement of AI models, ensuring they remain aligned with the evolving needs of healthcare practitioners and patients in the Metaverse.

Remaining in the healthcare domain, [36] focuses on deploying RRELIEF and SHAP methods for high-dimensional healthcare data analysis in order to enhance both feature selection and interpretability. By evaluating the ability of covariates to distinguish between instances, RRELIEF ranks features based on their relevance, effectively reducing dimensionality before training models such as the Cox proportional hazards model. This preprocessing step is crucial for identifying the most impactful variables that influence patient outcomes. Further, SHAP provides an interpretability framework that quantifies the contribution of each feature to the model's predictions, particularly in complex neural network-based survival models. This allows the understanding of how specific covariates affect survival probabilities, thereby compensating for the loss of transparency that often accompanies the transition from traditional regression models to more complex architectures. Together, RRELIEF and SHAP enable a comprehensive transparent approach to model development in the healthcare domain. For cybersickness detection, [21] proposes a comprehensive XAI-VR-LENS framework that implements several XAI methods, i.e., LIME, SHAP, Morris Sensitivity Analysis (MSA), and PDPs (Partial Dependence Plots), for assuring model interpretability and feature selection. Specifically, SHAP helps identify the most significant features contributing to cybersickness outcomes. Complementing this, MSA conducts a global sensitivity analysis by adjusting one feature at a time, thereby quantifying the impact of individual features on the model's predictions, which aids in ranking feature importance. For local explanations, LIME approximates the model's decision boundary around specific predictions, offering insights into how individual features affect outcomes for particular instances. In addition, PDPs allow the visualization of the marginal effects of one or two features on the predicted outcomes, illustrating the nature of relationships between features and predictions. Together, these methods provide an in-depth understanding of the model's behavior, enabling the identification and ranking of dominant features that contribute to cybersickness, which are subsequently used to retrain a more efficient model with reduced complexity, ultimately enhancing the model's deployment in practical applications.

6. Evaluation of the XAI Methods Used

Evaluating XAI methods is crucial to ensure that these AI systems not only perform accurately but also provide correct, clear, and comprehensible explanations for their decisions. By analyzing the papers selected for this survey, we identified three directions for evaluation:

(1) The **evaluation of the performance** of the proposed models for different tasks, using quantitative measures like accuracy, F-1 score, confusion matrices, Mean Square Error (MSE), the total time/number of moves to solve a puzzle, Concordance Index, etc.
(2) The **evaluation of the user experience** in the AR/VR environments, using both quantitative and qualitative measures from open/semi-structured interviews and/or validated questionnaires like SUS, NASA-TLX, Simulator Sickness Questionnaire (SSQ), etc.
(3) The **evaluation of the explanations** of the models, for which there is no standard evaluation strategy.

While for the performance of the models and for the user experience there are widely accepted and validated metrics, evaluating explainable AI presents several complexities due to its unique characteristics:

1. Subjectivity: Interpretability is often context-dependent and varies based on the user's perspective. What one person considers interpretable might not align with another's viewpoint. Human-centric metrics like "trust" or "comprehensibility" are subjective and challenging to quantify objectively.
2. Diverse Models: Explainable AI encompasses a wide range of models, from simple linear regressions to complex neural networks. Each model type requires tailored evaluation techniques with different levels of difficulty (e.g., evaluating the impact of deep learning layers on overall interpretability is intricate).
3. Trade-offs: Balancing interpretability with performance (accuracy, speed, etc.) is challenging. More interpretable models may sacrifice predictive power, while highly accurate models might lack transparency.
4. Lack of Ground Truth: Unlike traditional metrics (e.g., accuracy), there is no universally accepted "ground truth" for interpretability. Quantifying how well an explanation aligns with human intuition remains elusive.
5. Dynamic Context: Interpretability requirements change based on the application. What is interpretable for medical diagnoses might differ from financial fraud detection.

The lack of a standardized evaluation framework, coupled with the multifaceted nature of interpretability, makes assessing explainable AI a complex endeavor. Therefore, we focused on how the evaluation of the explanations was conducted in the reviewed articles.

First, we identified the two main types of XAI evaluations of explanations, as proposed in [14]: with users and without users. By users we mean participants in controlled human experiments, who are usually lay persons (i.e., non-experts in the field of application).

For each main type of evaluation, we distinguished two subcategories, briefly described in Table 8 and discussed in more details below.

Table 8. Types of XAI evaluation.

Types of XAI Evaluations of Explanations		Ref
with users	subjective	[31–35,38]
	objective	[32,33,35]
without users	in relation to the knowledge in the field of application	[18,19,22,24]
	in relation to model performance	[18,20–30,36]

6.1. The Subjective Evaluation with Users

Subjective evaluation with users is carried out by asking for user-perceived quality in the presence of explanations. The data used for this type of assessment were collected from participants in user studies, who were asked to fill in questionnaires containing items illustrated in Figure 17. For the explanations to be assessed as good, the positive aspects in the questionnaires (like trust, transparency, usefulness, etc.) should have a high ranking, and the negative aspects (like irritation, mental load) should have a low ranking.

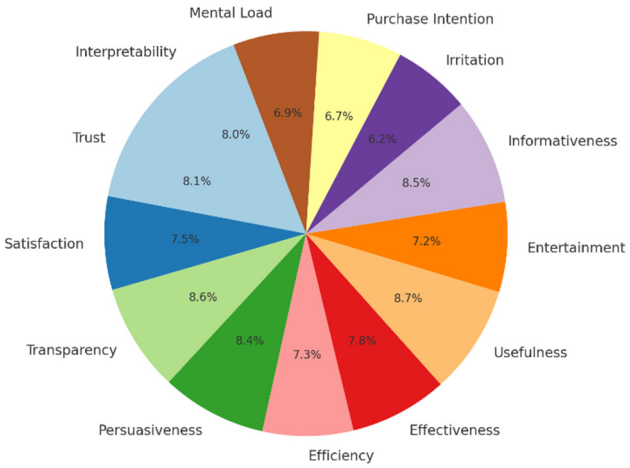

Figure 17. Subjective evaluation with users.

In [31,32], the explanations of the recommender systems were evaluated based on questionnaires containing questions for trust, satisfaction and transparency measurements as suggested in [48]; questions for persuasiveness, efficiency, and effectiveness collected from [49]; and questions regarding the overall transparency, perceived ease of use, and perceived usefulness from the ResQue questionnaire [50].

In [33], the explicit trust and the propensity to trust in an assistance system were assessed in the presence/absence of the (manual) explanations. Explicit trust is the user's subjectively perceived level of trust, measured via the Trust in Automation Questionnaire developed in [51]. The propensity to trust was measured through the Propensity to Trust Questionnaire [52]. The participants in the study were also asked to rate the explanations on a scale from 1 (very bad) to 5 (very good).

In [34], the explanations of an AR shopping assistant were evaluated through a questionnaire that included a series of validated constructs like "Usefulness", "Entertainment", "Informativeness", "Irritation", "Purchase Intention", and "Trust" [53,54]. In order to ensure the high quality of the questionnaire (and, implicitly, the quality of the collected data), the validity and reliability of the constructs were assessed by using confirmatory factor analysis to test the measurement models. Consequently, some of the items from the questionnaire were excluded from further analysis. The participants also answered six open-ended questions to obtain their general opinion about the presented AR shopping assistant.

In [35], the explanations of the Markov Decision Process (MDP), presented as a heatmap of the state space information, were evaluated based on the ratings of different aspects of three concepts: trust, interpretability, and mental load. The survey contained questions from established questionnaires in the robotics and explainable AI community [53,55–57].

In [38], the explanations of for the actions of a robot were evaluated based on data collected from participants in a user study, who rated 'trust' as prediction accuracy [58], by answering a question adapted from [58,59], To what extent do you trust/believe this robot possesses the ability to open a medicine bottle?, on a scale between 0 and 100.

6.2. The Objective Evaluation with Users

Objective evaluation with users is realized by comparing the performance of participants on specific tasks with/without explanations available, as depicted in Figure 18. If the performance improves in the presence of explanations, then the explanations are assessed as good.

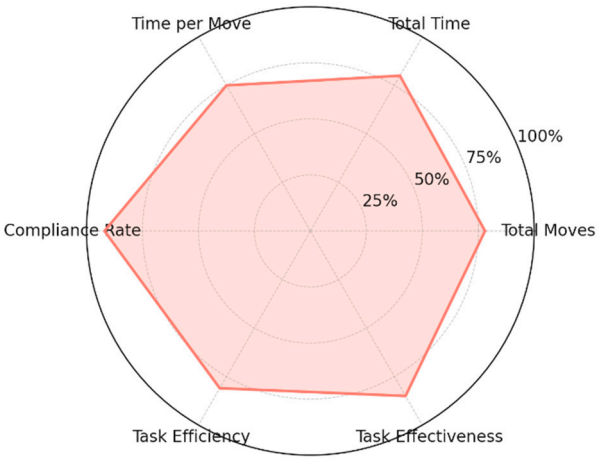

Figure 18. Objective evaluation with users.

In [32], three objective dependent variables (i.e., efficiency, effectiveness, and direction of persuasiveness, as suggested in [60,61]) were automatically measured during the tasks, in the absence/presence of the explanations.

In [33], the implicit trust in an assistance system was assessed in the presence/absence of the (manual) explanations. Implicit trust refers to measures of trust based on user behavior, e.g., how frequently did they use the specific technology or how often did they follow the technology's suggestions.

In [35], the objective metrics measured were as follows: Total Moves (the total number of moves needed to solve the task), Total Time (the total time needed to solve the task), Time per Move (the average time per move), and Compliance Rate (the percentage of moves taken matching the recommendation provided by the system).

6.3. The Evaluation of the Explanations Without Users, Realized in Relation to the Knowledge in the Field of Application

Here, by knowledge in the field of application we refer either to the direct opinion of an expert in that field or to a well-known result from the literature in that field. Thus, good explanations are considered as those who are in line with the knowledge in the field of application.

In [18], the discriminative movements for all score classes, visually interpreted based on the heatmap Grad-CAM++, were correlated with the experts' movement classification in the ETI scoring rubric.

In [19], the features' importance, measured by a gradient-based relevance attribution method DeepLIFT, were in accordance with a well-known conclusion in the neuroscience field.

In [22], the conclusion that the head-orientation data plays an important role in user identification was confirmed by other results obtained in previous work in VR.

In [24], the selected visual features from the VR simulator correlate with the visual rating scales of OSATS [62,63], which are considered to be the gold standard for the assessment of simulated surgical tasks.

6.4. The Evaluation of the Explanations Without Users, Realized in Relation to Model Performance

The models with the best predictive accuracy are considered the ones to provide the best explanations. Various architectures are proposed and/or different feature selection techniques are applied when searching for the best accuracy models. We identified the following approaches, as shown in Figure 19.

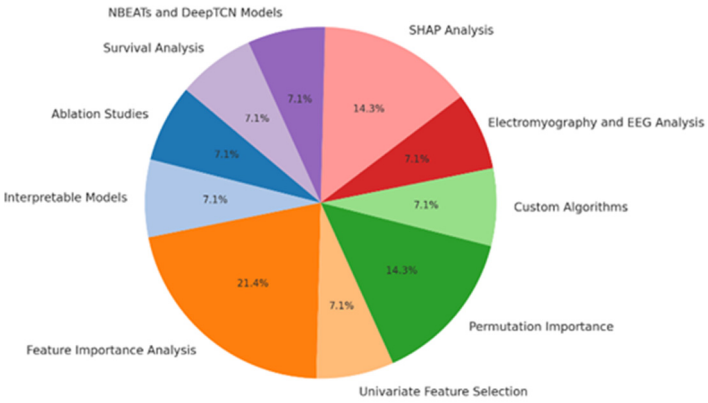

Figure 19. Evaluation without users, in relation to model performance.

In [18], several ablation studies were conducted to demonstrate the meaningful pattern localization (i.e., the discriminative regions identified by using heatmaps of Grad-CAM++) of the attention-based dilated CNN, which was the model with the best predictive accuracy.

In [20], the models utilized are inherently interpretable. The 'glass box' models with low accuracy provided 'poor' explanations. Local explanations were analyzed only for the most performant model (i.e., EBM) through visual interpretation based on the feature importance in terms of the Mean Absolute Score (MAS).

In [21,27,30], the top most dominant features for the predictive models were identified (using SHAP or Morris Sensitivity Analysis (MSA)), followed by a model reduction using these most important features; the resulting reduced models had a higher performance compared to the initial full models.

In [22], a univariate feature selection technique was used to create two feature sets. The 'quality' of the feature set that led to the best prediction accuracy was confirmed visually, by plotting the clusters for the 1-NN model, which were less overlapping (i.e., more discriminant) than for the feature set with a poor prediction accuracy.

In [23], 'glass box' models decision tree and random forest had a good prediction accuracy; for the random forest model, the Mean Decrease in Impurity (MDI) and Mean Decrease Accuracy (MDA) were utilized to determine the importance of the features [64]. The MDI describes how each individual feature contributes to the decrease in the Gini index across trees, while MDA is defined as the decrease in model accuracy when a single feature value is randomly shuffled (permutation). The evaluation of permutation importance is preferred because the Gini-importance measurement may be conditioned by the low cardinality of the features or by categorical variables with a small number of classes.

In [24], a custom algorithm which combines forward and backward feature selection was used [65], the optimal set of features being the one where the highest accuracy of the SVM model was obtained. The importance of the features was ranked based on the magnitude of the corresponding weights in a SVM model.

In [25], the permutation feature importance method was applied to rank the importance of the features in two feature sets with the highest accuracy in a random forest classifier.

In [26], different feature selection techniques for electromyography (EMG) and EEG signals were applied to create different feature sets that were analyzed based on the predictive accuracy of the classification models SVM and LDA; to further understand which features were utilized by the classification algorithm to maximize class separability between classes, weight vectors were re-projected into sensor space and topographies were created and visually analyzed.

In [28], the SHAP approach was applied for the best-performing classification model, the resulting most-informative features being consistent with the statistical results per-

formed (Mann–Whitney U test for testing differences in features between the two classes: novices and experts).

In [29], from different combinations of data modalities and forecasting horizons, the feature set resulting in the best predictive accuracy for the interpretable NBEATs model and the DeepTCN model was selected, but the interpretable outputs of the NBEATs model (i.e., trend and seasonality) were not discussed.

In [36], different feature selection techniques (e.g., VIF to remove highly correlated features; RReliefF to associate importance weights for features [66]; or statistical tests to find the significant features for a model) were applied to create different sets of features to train four models for survival analysis. For the best model according to the Concordance Index (C-index), visual analysis of SHAP values for the effects of the features was carried out and the results were compared to the effects of the features of the other models.

So far, we have identified four approaches for the evaluation of the explanations in XAI systems used in XR environments, and we have provided details and references to other resources when needed that would allow a researcher to successfully apply the presented methods in their own work.

Although these approaches are indeed able to provide a measure of the quality of the explanations, quantifying how well an explanation aligns with human understanding and intuition still remains a challenge, mainly due to the subjective nature of metrics like trust, transparency, and comprehensibility, and the lack of ground truth for interpretability.

With this in mind, we suggest a list of five recommendations as a matter of best practice as part of the contribution of this research. These recommendations would ensure the development, deployment, and utilization of XAI mechanisms in the XR domain by incorporating reliable, interpretable, and good-quality explanations. Additionally, they would facilitate the users' comprehension and trust when interacting with the underlying AI models used.

1. **The quality of data** is very important when assessing the quality of the explanations. The well-known expression "garbage in, garbage out" is well suited here: when an AI model is trained on flawed data, it learns nonsensical relations, which are revealed by the explanation. The explanations might then be perceived as being wrong, although they are truthfully reflecting the model's reasoning. This is considered the main shortcoming when evaluating explainable AI [14].
   Therefore, whenever the problem permits it, we should invest in having high-quality data. For example, in [18], the motion data in some specific medical procedures was evaluated by three experts whose scoring was assessed by the concordance correlation coefficients, which confirmed the raters' agreements on scoring. In [34], the quality of the collected data was ensured by using confirmatory factor analysis to assess the validity and reliability of the items in the questionnaire.
2. **The relation to the predictive accuracy**: If an AI model has a low predictive accuracy, meaning that it has learned a poor approximation of the underlying relationships in the data, then any explanation extracted from the model is unlikely to have high quality [13]. Thus, for a low-performance model, it is reasonable to evaluate the explanations as poor [20]. But, for a high-performance model, the quality of the explanations is not implicit. This should be proved by additional techniques (e.g., in [18,19,22,24] the explanations are in line with the experts' knowledge in that field).
   In [25], the problem was to identify users by their hand-tracking data. The feature set for which the random forest model achieved the highest predictive accuracy also contained, in addition to hand-related features, head-related features that were rated as the first two most important features (in particular, "Head.pos.y" and "Head.rot.z"). "Head.pos.y" is the height of the HMD above the ground, which is approximately

the user's height. "Head.rot.z" is the pitch of the head, i.e., whether they tilted their head down or up. These two head-related features ended up being important for user identification based on hand-tracking data because of the design of the experiment: the user interface elements were always at the same height above the ground, so a taller person had to tilt their head further down. In such a situation, the relevancy of explanations to the real-world problem should be demonstrated.

3. **The importance of feature selection**: Having a smaller number of features that are more informative makes the relationship that needs to be learned by the model simpler, becoming easier to be explained [13]. In many domains, expert knowledge can help in constructing useful feature sets for predictive models, to ensure the relevancy of the features (i.e., domain-based feature engineering) (e.g., in [19], the use of EEG data to classify attention levels). The domain-based feature engineering could be further continued with dimensionality reduction techniques [26] and feature selection techniques [21,26,27,30,36].
4. **Visualizations**: High-dimensional datasets often present challenges in quickly grasping the intricate relationships a model has learned. Therefore, we suggest visualizing heatmaps/topographies of the most informative features used by the predictive models (for example, see [19,26] for classification based on EEG signals; [18] for classification based on motion trajectories).
5. **Display styles in XR**: When the target audience for the explanations is made up of lay persons (i.e., non-experts in the field of application), it would be very important how the explanations are presented in XR [31,32,38].

7. Discussions and Conclusions

This paper provides an overview of the current research landscape at the intersection of XR and XAI. The need for interdisciplinary collaboration to address various challenges is revealed when developing transparent and interpretable AI systems and creating immersive experiences that leverage the strengths of both XR and XAI. To identify the key areas of focus, four research questions were formulated and further addressed, thus contributing to the development of a comprehensive research agenda for integrating XAI in XR environments, paving the way for more user-centered, trustworthy, and engaging immersive experiences.

Regarding **domains of applications, benefits, and opportunities**, our systematic review reveals a clear dominance of healthcare applications (30%) in XAI-enhanced XR systems, followed by recommendation systems (16%), and equal representation (12%) of cybersickness, training, and experience personalization. This distribution demonstrates the particular value of explainable AI in critical domains where transparency and trust are paramount. The benefits span from enhanced user trust and improved system transparency to better decision comprehension, particularly in high-stakes applications like medical training and diagnosis.

The advancements in the field have also introduced promising opportunities, including the development of tailored frameworks to address specific industry needs, such as those in healthcare, and the exploration of innovative methodologies, like cybersickness detection using xML or forecasting cybersickness severity by using deep temporal convolutional forecasting models. The emerging applications and opportunities open up numerous perspectives for further research, such as expanding findings to larger populations, leveraging AR to enhance user experiences, and using AR for the design of explanation artifacts and visualizations within recommendation systems. Other potential perspectives include exploring diverse use cases, such as user authentication and access control via

behavioral biometrics, and investigating human–robot team dynamics in scenarios like rescue missions.

The analysis of XR systems integrating XAI demonstrates a significant preference for VR-based approaches, with fourteen publications focusing on pure VR compared to six for pure AR and five for mixed AR/VR systems. The predominance of HMDs, particularly the HTC Vive for VR and the HoloLens for AR, suggests a maturation of these platforms for XAI integration. A critical finding is the extensive use of additional sensors, especially in VR systems, with eye tracking leading in six VR publications. This trend highlights the importance of multimodal data collection for enhanced explainability. However, several limitations were identified, including restricted deployment on actual devices and limited participant diversity. While various XAI methods are used either standalone or in combination with human experts/users, no study considered their use on applications based on recent advancements in the AI domain, in particular, Generative AI techniques such as the ones based on LLMs. We can also mention as limitations the number of application domains which are targeted by XAI-enhanced XR systems and the fact that the questionnaire-driven method has a relatively low incidence for assessing the approaches presented in the selected publications. These limitations suggest critical areas for future research focus, particularly in real-world deployment and diverse user testing.

To address the explainability of AI models, the field of **XAI employs a range of methods and techniques** aimed at enhancing the transparency of AI behaviors and outputs, while preserving high levels of performance. The different approaches which were examined incorporate specific XAI mechanisms within an established taxonomy, which categorizes them based on the model dependency and the type of explanation offered. These distinctions include model-based versus model-agnostic methods, as well as global versus local explanations. Each of these categories is discussed in detail in the review, along with their relevance to the analyzed papers.

The evaluation of XAI methods in XR environments, focusing on assessing explanations, revealed a complex landscape with four distinct approaches: with users (including subjective and objective assessments) and without users (in relation to domain knowledge and model performance). This diversity in evaluation methods reflects the multifaceted nature of assessing the quality of the explanations for AI systems in XR environments, our paper providing researchers with a structured perspective to validate their explanatory methods.

Our review also led to five key recommendations for future research: ensuring high-quality data collection, establishing clear relationships between predictive accuracy and explanation quality, emphasizing feature selection, utilizing effective visualizations, and carefully considering XR-specific display modalities.

These recommendations are particularly crucial given that poor data quality can lead to nonsensical model relationships, while effective feature selection and visualization can significantly enhance the interpretability and relevance of explanations for both expert and non-expert users.

Still, a significant limitation is the persistent challenge of quantifying how well explanations align with human intuition/understanding, particularly due to the subjective nature of interpretability metrics and the lack of standardized evaluation frameworks.

These findings collectively suggest several important directions for future research in the field of XAI-enhanced XR systems. First, there is a need for more standardized evaluation frameworks (including benchmarks and metrics) that can effectively assess both technical accuracy and human interpretability (the quality of explanations). Also, the integration of XAI in XR environments necessitates special attention to the presentation modalities of explanations, particularly when targeting non-expert users.

Second, the development of more sophisticated sensor-integration techniques, particularly in AR systems, would enrich the data available, which could enhance the quality and reliability of explanations.

Finally, there is a clear opportunity for expanding applications into other critical domains (e.g., education, security) where transparency and trust are essential either individually or at a collective level. At the same time, standardized evaluation metrics and benchmarks for evaluating interpretability in an interdisciplinary setting should be considered. Another future research area is represented by the integration of Generative AI modeling techniques and their corresponding explanation mechanisms that would enhance the overall performance and trust in XR systems by using adaptive natural language and visual explanations tailored to different user contexts through real-time text-based explanations and scenario adaptation. For example, one concrete domain application could be in security, with the integration of behavioral biometrics for security incident prevention or response, and rescue missions could use human–robot collaboration with real-time visualization of user-adaptive explanations.

Author Contributions: Conceptualization, C.M.; methodology, C.M.; formal analysis, C.M., M.A.C., D.D. and L.M.; data curation, C.M., M.A.C., D.D. and L.M.; writing—original draft preparation, C.M., M.A.C., D.D. and L.M.; writing—review and editing, C.M. and M.A.C.; visualization, C.M., D.D. and L.M.; supervision, C.M. and M.A.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: This is a survey paper, where no new data were created.

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

References

1. Milgram, P.; Kishino, F. A Taxonomy of Mixed Reality Visual Displays. *IEICE Trans. Inf. Syst.* **1994**, *77*, 1321–1329.
2. Hirzle, T.; Müller, F.; Draxler, F.; Schmitz, M.; Knierim, P.; Hornbæk, K. When XR and AI Meet—A Scoping Review on Extended Reality and Artificial Intelligence. In Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, Hamburg, Germany, 23–28 April 2023; pp. 1–45.
3. Reiners, D.; Davahli, M.R.; Karwowski, W.; Cruz-Neira, C. The Combination of Artificial Intelligence and Extended Reality: A Systematic Review. *Front. Virtual Real.* **2021**, *2*, 721933. [CrossRef]
4. Devagiri, J.S.; Paheding, S.; Niyaz, Q.; Yang, X.; Smith, S. Augmented Reality and Artificial Intelligence in Industry: Trends, Tools, and Future Challenges. *Expert Syst. Appl.* **2022**, *207*, 118002. [CrossRef]
5. Sahu, C.K.; Young, C.; Rai, R. Artificial Intelligence (AI) in Augmented Reality (AR)-Assisted Manufacturing Applications: A Review. *Int. J. Prod. Res.* **2020**, *59*, 4903–4959. [CrossRef]
6. Longo, U.G.; De Salvatore, S.; Candela, V.; Zollo, G.; Calabrese, G.; Fioravanti, S.; Giannone, L.; Marchetti, A.; De Marinis, M.G.; Denaro, V. Augmented Reality, Virtual Reality and Artificial Intelligence in Orthopedic Surgery: A Systematic Review. *Appl. Sci.* **2021**, *11*, 3253. [CrossRef]
7. Papakostas, C.; Troussas, C.; Sgouropoulou, C. Review of the Literature on AI-Enhanced Augmented Reality in Education. In *Special Topics in Artificial Intelligence and Augmented Reality*; Cognitive Technologies; Springer Nature: Cham, Switzerland, 2024; pp. 13–50. ISBN 978-3-031-52004-4.
8. Bassyouni, Z.; Elhajj, I.H. Augmented Reality Meets Artificial Intelligence in Robotics: A Systematic Review. *Front. Robot. AI* **2021**, *8*, 724798. [CrossRef] [PubMed]
9. Luck, M.; Aylett, R. Applying Artificial Intelligence to Virtual Reality: Intelligent Virtual Environments. *Appl. Artif. Intell.* **2000**, *14*, 3–32. [CrossRef]
10. Levine, E. The Next Frontier in Education: How Generative AI and XR Will Evolve the World of Learning in the Next Decade. Available online: https://www.qualcomm.com/news/onq/2024/05/the-next-frontier-in-education-how-generative-ai-and-xr-evolve-learning (accessed on 1 June 2024).
11. Kreuzer, T.; Papapetrou, P.; Zdravkovic, J. Artificial Intelligence in Digital Twins—A Systematic Literature Review. *Data Knowl. Eng.* **2024**, *151*, 102304. [CrossRef]

12. Available online: https://Commission.Europa.Eu/Law/Law-Topic/Data-Protection/Data-Protection-Eu_en (accessed on 10 September 2024).
13. Murdoch, W.J.; Singh, C.; Kumbier, K.; Abbasi-Asl, R.; Yu, B. Definitions, Methods, and Applications in Interpretable Machine Learning. *Proc. Natl. Acad. Sci. USA* **2019**, *116*, 22071–22080. [CrossRef]
14. Nauta, M.; Trienes, J.; Pathak, S.; Nguyen, E.; Peters, M.; Schmitt, Y.; Schlötterer, J.; Van Keulen, M.; Seifert, C. From Anecdotal Evidence to Quantitative Evaluation Methods: A Systematic Review on Evaluating Explainable AI. *ACM Comput. Surv.* **2023**, *55*, 1–42. [CrossRef]
15. Denyer, D.; Tranfield, D. Producing a Systematic Review. In *The Sage Handbook of Organizational Research Methods*; Sage Publications Ltd.: Thousand Oaks, CA, USA, 2009; pp. 671–689. ISBN 978-1-4129-3118-2.
16. 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. Declaración PRISMA 2020: Una guía actualizada para la publicación de revisiones sistemáticas. *Rev. Esp. Cardiol.* **2021**, *74*, 790–799. [CrossRef]
17. Liberati, A.; Altman, D.G.; Tetzlaff, J.; Mulrow, C.; Gøtzsche, 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**, *6*, e1000100. [CrossRef]
18. Zhao, S.; Xiao, X.; Wang, Q.; Zhang, X.; Li, W.; Soghier, L.; Hahn, J. An Intelligent Augmented Reality Training Framework for Neonatal Endotracheal Intubation. In Proceedings of the 2020 IEEE International Symposium on Mixed and Augmented Reality (ISMAR), Porto de Galinhas, Brazil, 9–13 November 2020; pp. 672–681.
19. Li, G.; Adeel Khan, M. Deep Learning on VR-Induced Attention. In Proceedings of the 2019 IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR), San Diego, CA, USA, 9–11 December 2019; pp. 163–1633.
20. Kundu, R.K.; Islam, R.; Calyam, P.; Hoque, K.A. TruVR: Trustworthy Cybersickness Detection Using Explainable Machine Learning. In Proceedings of the 2022 IEEE International Symposium on Mixed and Augmented Reality (ISMAR), Singapore, 17–21 October 2022; pp. 777–786.
21. Kundu, R.K.; Elsaid, O.Y.; Calyam, P.; Hoque, K.A. VR-LENS: Super Learning-Based Cybersickness Detection and Explainable AI-Guided Deployment in Virtual Reality. In Proceedings of the 28th International Conference on Intelligent User Interfaces, Sydney, NSW, Australia, 27–31 March 2023; pp. 819–834.
22. Liebers, J.; Horn, P.; Burschik, C.; Gruenefeld, U.; Schneegass, S. Using Gaze Behavior and Head Orientation for Implicit Identification in Virtual Reality. In Proceedings of the 27th ACM Symposium on Virtual Reality Software and Technology, Osaka, Japan, 8–10 December 2021; pp. 1–9.
23. Losada, Á.; Páez, F.J.; Luque, F.; Piovano, L. Application of Machine Learning Techniques for Predicting Potential Vehicle-to-Pedestrian Collisions in Virtual Reality Scenarios. *Appl. Sci.* **2022**, *12*, 11364. [CrossRef]
24. Mirchi, N.; Bissonnette, V.; Yilmaz, R.; Ledwos, N.; Winkler-Schwartz, A.; Del Maestro, R.F. The Virtual Operative Assistant: An Explainable Artificial Intelligence Tool for Simulation-Based Training in Surgery and Medicine. *PLoS ONE* **2020**, *15*, e0229596. [CrossRef] [PubMed]
25. Liebers, J.; Brockel, S.; Gruenefeld, U.; Schneegass, S. Identifying Users by Their Hand Tracking Data in Augmented and Virtual Reality. *Int. J. Hum.–Comput. Interact.* **2024**, *40*, 409–424. [CrossRef]
26. McDermott, E.J.; Raggam, P.; Kirsch, S.; Belardinelli, P.; Ziemann, U.; Zrenner, C. Artifacts in EEG-Based BCI Therapies: Friend or Foe? *Sensors* **2021**, *22*, 96. [CrossRef] [PubMed]
27. Kundu, R.K.; Islam, R.; Quarles, J.; Hoque, K.A. LiteVR: Interpretable and Lightweight Cybersickness Detection Using Explainable AI. In Proceedings of the 2023 IEEE Conference Virtual Reality and 3D User Interfaces (VR), Shanghai, China, 25–29 March 2023; pp. 609–619.
28. Gao, H.; Bozkir, E.; Stark, P.; Goldberg, P.; Meixner, G.; Kasneci, E.; Göllner, R. Detecting Teacher Expertise in an Immersive VR Classroom: Leveraging Fused Sensor Data with Explainable Machine Learning Models. In Proceedings of the 2023 IEEE International Symposium on Mixed and Augmented Reality (ISMAR), Sydney, Australia, 16–20 October 2023; pp. 683–692.
29. Islam, R.; Desai, K.; Quarles, J. Towards Forecasting the Onset of Cybersickness by Fusing Physiological, Head-Tracking and Eye-Tracking with Multimodal Deep Fusion Network. In Proceedings of the 2022 IEEE International Symposium on Mixed and Augmented Reality (ISMAR), Singapore, 17–21 October 2022; pp. 121–130.
30. Choi, D.; Seo, S.; Park, H.; Hong, T.; Koo, C. Forecasting Personal Learning Performance in Virtual Reality-Based Construction Safety Training Using Biometric Responses. *Autom. Constr.* **2023**, *156*, 105115. [CrossRef]
31. Yang, P.-K.; Alvarado Rodriguez, O.L.; Gutierrez, F.; Verbert, K. Touching the Explanations: Explaining Movie Recommendation Scores in Mobile Augmented Reality. In Proceedings of the 2022 IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR), San Diego, CA, USA, 12–14 December 2022; pp. 157–162.
32. Zheng, M.; Thomas, R.J.; Pan, X.; Xu, Z.; Liang, Y.; Campbell, A.G. Augmenting Feature Importance Analysis: How Color and Size Can Affect Context-Aware AR Explanation Visualizations? In Proceedings of the 2022 IEEE International Symposium on Mixed and Augmented Reality (ISMAR), Singapore, 17–21 October 2022; pp. 508–517.

33. Faulhaber, A.K.; Ni, I.; Schmidt, L. The Effect of Explanations on Trust in an Assistance System for Public Transport Users and the Role of the Propensity to Trust. In Proceedings of the Mensch und Computer 2021, Ingolstadt, Germany, 5–8 September 2021; pp. 303–310.
34. Zimmermann, R.; Mora, D.; Cirqueira, D.; Helfert, M.; Bezbradica, M.; Werth, D.; Weitzl, W.J.; Riedl, R.; Auinger, A. Enhancing Brick-and-Mortar Store Shopping Experience with an Augmented Reality Shopping Assistant Application Using Personalized Recommendations and Explainable Artificial Intelligence. *J. Res. Interact. Mark.* **2023**, *17*, 273–298. [CrossRef]
35. Tabrez, A.; Luebbers, M.B.; Hayes, B. Descriptive and Prescriptive Visual Guidance to Improve Shared Situational Awareness in Human-Robot Teaming. In Proceedings of the AAMAS '22: Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems, Virtual Event, 9 May 2022; pp. 1256–1264.
36. Kalatian, A.; Farooq, B. Decoding Pedestrian and Automated Vehicle Interactions Using Immersive Virtual Reality and Interpretable Deep Learning. *Transp. Res. Part C Emerg. Technol.* **2021**, *124*, 102962. [CrossRef]
37. Gillmann, C.; Smit, N.N.; Groller, E.; Preim, B.; Vilanova, A.; Wischgoll, T. Ten Open Challenges in Medical Visualization. *IEEE Comput. Graph. Appl.* **2021**, *41*, 7–15. [CrossRef]
38. Liu, H.; Zhu, Y.; Zhu, S. Patching Interpretable AND-OR-GRAPH Knowledge Representation Using Augmented Reality. *Appl. AI Lett.* **2021**, *2*, e43. [CrossRef]
39. Moreira, C.; Nobre, I.B.; Sousa, S.C.; Pereira, J.M.; Jorge, J. Improving X-Ray Diagnostics through Eye-Tracking and XR. In Proceedings of the 2022 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW), Christchurch, New Zealand, 12–16 March 2022; pp. 450–453.
40. Mozumder, M.A.I.; Ariful, T.P.T.; Sumon, R.I.; Uddin, S.M.I.; Athar, A.; Kim, H.-C. The Metaverse for Intelligent Healthcare Using XAI, Blockchain, and Immersive Technology. In Proceedings of the 2023 IEEE International Conference on Metaverse Computing, Networking and Applications (MetaCom), Kyoto, Japan, 26–28 June 2023; pp. 612–616.
41. Turner, C.J.; Garn, W. Next Generation DES Simulation: A Research Agenda for Human Centric Manufacturing Systems. *J. Ind. Inf. Integr.* **2022**, *28*, 100354. [CrossRef]
42. Zanitti, M.; Ferens, M.; Ferrarin, A.; Trovò, F.; Miskovic, V.; Prelaj, A.; Shen, M.; Kosta, S. MetaLung: Towards a Secure Architecture for Lung Cancer Patient Care on the Metaverse. In Proceedings of the 2023 IEEE International Conference on Metaverse Computing, Networking and Applications (MetaCom), Kyoto, Japan, 26–28 June 2023; pp. 201–208.
43. Ali, S.; Abuhmed, T.; El-Sappagh, S.; Muhammad, K.; Alonso-Moral, J.M.; Confalonieri, R.; Guidotti, R.; Del Ser, J.; Díaz-Rodríguez, N.; Herrera, F. Explainable Artificial Intelligence (XAI): What We Know and What Is Left to Attain Trustworthy Artificial Intelligence. *Inf. Fusion* **2023**, *99*, 101805. [CrossRef]
44. Chamola, V.; Hassija, V.; Sulthana, A.R.; Ghosh, D.; Dhingra, D.; Sikdar, B. A Review of Trustworthy and Explainable Artificial Intelligence (XAI). *IEEE Access* **2023**, *11*, 78994–79015. [CrossRef]
45. Dwivedi, R.; Dave, D.; Naik, H.; Singhal, S.; Omer, R.; Patel, P.; Qian, B.; Wen, Z.; Shah, T.; Morgan, G.; et al. Explainable AI (XAI): Core Ideas, Techniques, and Solutions. *ACM Comput. Surv.* **2023**, *55*, 1–33. [CrossRef]
46. Salih, A.M.; Raisi-Estabragh, Z.; Galazzo, I.B.; Radeva, P.; Petersen, S.E.; Lekadir, K.; Menegaz, G. A Perspective on Explainable Artificial Intelligence Methods: SHAP and LIME. *Adv. Intell. Syst.* **2024**, 2400304. [CrossRef]
47. Nguyen, H.; Cao, H.; Nguyen, K.; Pham, N. Evaluation of Explainable Artificial Intelligence: SHAP, LIME, and CAM. In Proceedings of the FPT AI Conference, Hanoi, Vietnam, 6–7 May 2021.
48. Tintarev, N.; Masthoff, J. Explaining Recommendations: Design and Evaluation. In *Recommender Systems Handbook*; Ricci, F., Rokach, L., Shapira, B., Eds.; Springer: Boston, MA, USA, 2015; pp. 353–382. ISBN 978-1-4899-7636-9.
49. Balog, K.; Radlinski, F. Measuring Recommendation Explanation Quality: The Conflicting Goals of Explanations. In Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, 25–30 July 2020; pp. 329–338.
50. Pu, P.; Chen, L.; Hu, R. A User-Centric Evaluation Framework for Recommender Systems. In Proceedings of the Fifth ACM Conference on Recommender Systems, Chicago, IL, USA, 23 October 2011; pp. 157–164.
51. Körber, M. Theoretical Considerations and Development of a Questionnaire to Measure Trust in Automation. In Proceedings of the 20th Congress of the International Ergonomics Association (IEA 2018), Florence, Italy, 26–30 August 2018; Bagnara, S., Tartaglia, R., Albolino, S., Alexander, T., Fujita, Y., Eds.; Advances in Intelligent Systems and Computing. Springer International Publishing: Cham, Switzerland, 2019; Volume 823, pp. 13–30, ISBN 978-3-319-96073-9.
52. Mcknight, D.H.; Carter, M.; Thatcher, J.B.; Clay, P.F. Trust in a Specific Technology: An Investigation of Its Components and Measures. *ACM Trans. Manag. Inf. Syst.* **2011**, *2*, 1–25. [CrossRef]
53. Hoffman, R.R.; Mueller, S.T.; Klein, G.; Litman, J. Metrics for Explainable AI: Challenges and Prospects. *arXiv* **2018**, arXiv:1812.04608.
54. Hausman, A.V.; Siekpe, J.S. The Effect of Web Interface Features on Consumer Online Purchase Intentions. *J. Bus. Res.* **2009**, *62*, 5–13. [CrossRef]

55. Hart, S.G.; Staveland, L.E. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In *Advances in Psychology*; Elsevier: Amsterdam, The Netherlands, 1988; Volume 52, pp. 139–183. ISBN 978-0-444-70388-0.
56. Lewis, M.; Sycara, K.; Walker, P. The Role of Trust in Human-Robot Interaction. In *Foundations of Trusted Autonomy*; Abbass, H.A., Scholz, J., Reid, D.J., Eds.; Studies in Systems, Decision and Control; Springer International Publishing: Cham, Switzerland, 2018; Volume 117, pp. 135–159. ISBN 978-3-319-64815-6.
57. Wallkötter, S.; Tulli, S.; Castellano, G.; Paiva, A.; Chetouani, M. Explainable Embodied Agents Through Social Cues: A Review. *ACM Trans. Hum.-Robot Interact.* **2021**, *10*, 1–24. [CrossRef]
58. Castelfranchi, C.; Falcone, R. Principles of Trust for MAS: Cognitive Anatomy, Social Importance, and Quantification. In Proceedings of the Proceedings International Conference on Multi Agent Systems (Cat. No.98EX160), Paris, France, 3–7 July 1998; pp. 72–79.
59. Jian, J.-Y.; Bisantz, A.M.; Drury, C.G. Foundations for an Empirically Determined Scale of Trust in Automated Systems. *Int. J. Cogn. Ergon.* **2000**, *4*, 53–71. [CrossRef]
60. Bilgic, M.; Mooney, R.J. Explaining Recommendations: Satisfaction vs. Promotion. In Proceedings of the Workshop on Beyond Personalization: The Next Stage of Recommender System Research at the International Conference on Intelligent User Interfaces, San Diego, CA, USA, 10–13 January 2005.
61. Gedikli, F.; Jannach, D.; Ge, M. How Should I Explain? A Comparison of Different Explanation Types for Recommender Systems. *Int. J. Hum.-Comput. Stud.* **2014**, *72*, 367–382. [CrossRef]
62. Szasz, P.; Louridas, M.; Harris, K.A.; Aggarwal, R.; Grantcharov, T.P. Assessing Technical Competence in Surgical Trainees: A Systematic Review. *Ann. Surg.* **2015**, *261*, 1046–1055. [CrossRef] [PubMed]
63. Martin, J.A.; Regehr, G.; Reznick, R.; Macrae, H.; Murnaghan, J.; Hutchison, C.; Brown, M. Objective Structured Assessment of Technical Skill (OSATS) for Surgical Residents. *Br. J. Surg.* **1997**, *84*, 273–278. [CrossRef] [PubMed]
64. Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning*; Springer Series in Statistics; Springer: New York, NY, USA, 2009; ISBN 978-0-387-84857-0.
65. Winkler-Schwartz, A.; Yilmaz, R.; Mirchi, N.; Bissonnette, V.; Ledwos, N.; Siyar, S.; Azarnoush, H.; Karlik, B.; Del Maestro, R. Machine Learning Identification of Surgical and Operative Factors Associated With Surgical Expertise in Virtual Reality Simulation. *JAMA Netw. Open* **2019**, *2*, e198363. [CrossRef] [PubMed]
66. Robnik-Sikonja, M.; Kononenko, I. An Adaptation of Relief for Attribute Estimation in Regression. In Proceedings of the Fourteenth International Conference on Machine Learning, San Francisco, CA, USA, 8–12 July 1997; pp. 296–304.