Deep Science Publishing
https://doi.org/10.70593/978-81-981271-4-3_9

Chapter 9

Explainable and trustworthy artificial intelligence, machine learning, and deep learning

Nitin Liladhar Rane [1], Suraj Kumar Mallick [2], Ömer Kaya [3], Jayesh Rane [4]

[1] *Vivekanand Education Society's College of Architecture (VESCOA), Mumbai 400074, India*
[2] *Shaheed Bhagat Singh College, University of Delhi, New Delhi 110017, India*
[3] *Engineering and Architecture Faculty, Erzurum Technical University, Erzurum 25050, Turkey*
[4] *Pillai HOC College of Engineering and Technology, Rasayani, India*
[1] nitinrane33@gmail.com

Abstract: The swift progression of artificial intelligence (AI), machine learning (ML), and deep learning (DL) has transformed different industries, offering unprecedented efficiency and innovation. Nevertheless, the growing intricacy and lack of transparency of these technologies have led to important worries about their reliability and ethical consequences. This research explores the growing area of Explainable Artificial Intelligence (XAI) that seeks to improve the clarity, comprehensibility, and responsibility of AI, ML, and deep learning models. XAI helps to increase trust and acceptance by making these technologies easier to understand for users and stakeholders, therefore tackling the "black box" issue. This research provides a thorough examination of the most recent approaches and structures in XAI, with a focus on important strategies like model-agnostic explanations, interpretable models, and post-hoc interpretability techniques. It also examines the important function of XAI in guaranteeing adherence to regulatory standards and ethical guidelines, which are becoming stricter globally. Moreover, the review assesses how XAI is incorporated into different fields such as healthcare, finance, and autonomous systems, illustrating its ability to reduce biases, enhance decision-making, and increase user confidence. This research highlights the significance of XAI in creating AI systems that are both strong and ethical by discussing current trends and developments.

Keywords: Artificial intelligence, Deep learning, Machine learning, Explainable artificial intelligence, Trustworthy artificial intelligence, Explainability.

Citation: Rane, N. L., Mallick, S. K., Kaya, O., & Rane, J. (2024). Explainable and trustworthy artificial intelligence, machine learning, and deep learning. In *Applied Machine Learning and Deep Learning: Architectures and Techniques* (pp. 167-191). Deep Science Publishing. https://doi.org/10.70593/978-81-981271-4-3_9

9.1 Introduction

The healthcare, banking, transportation, and entertainment sectors have all seen significant changes in the last several years due to the widespread use of artificial intelligence (AI) technology (Minh et al., 2022; Kaur et al., 2022; Angelov et al., 2021; Rawal et al., 2021). Concerns over the transparency and dependability of these systems are growing even while AI advances remarkably (Angelov et al., 2021; Islam et al., 2022; Markus et al., 2021; Vilone & Longo, 2020). To address these challenges, explainable AI (XAI) has emerged as a significant area of research that focusses on enhancing the interpretability and comprehensibility of AI models (Kaur et al., 2022; Angelov et al., 2021; Albahri et al., 2023; Speith, 2022; Nazar et al., 2021). By making complex machine learning (ML) and deep learning (DL) algorithms' decision-making processes more transparent, XAI hopes to win over users' and stakeholders' support and confidence. There are multiple reasons why explainability is important in AI. First off, artificial intelligence (AI) systems are being employed in high-stakes scenarios where making the wrong decision could have serious consequences. The ability to rationalise and explain AI-generated conclusions is critical in domains such as financial forecasts, medical diagnostics, and autonomous driving (Markus et al., 2021; Ali et al., 2023; Nassar et al., 2020). Additionally, regulations and ethical standards require transparency in the functioning of AI systems to uphold fairness, accountability, and adherence to guidelines (Markus et al., 2021; Vilone & Longo, 2020; Arrieta et al., 2020; Loh et al., 2022; Vilone & Longo, 2021). The significance of giving explanations for automated decisions is highlighted by the General Data Protection Regulation (GDPR) and other related laws, which stress the importance of XAI.

Each feature in Fig 9.1 reveals how an explainable model can be evaluated against these criteria and how the internal workings, performance and reliability of the model can be understood. For example, fairness and verifiability refer to the ability of a model's output to be judged and subject to review; Accessibility and interactivity refer to the ability of users to directly interact with the model's decision-making process. Confidentiality ensures that a third party can disclose the internal operations of the model, while reliability refers to the certainty of whether the model performs as expected when performing a particular task. Figure 9.1 visually summarizes the criteria which explainable will evaluate reliable AI, ML and DL models evaluated, and what these criteria mean.

However, there are significant barriers to the adoption and confidence of the existing ML and DL models—often referred to as "black boxes"—due to their complexity and lack of transparency (Das & Rad, 2020; Kaur et al., 2021; Machlev et al., 2022). For users and decision-makers to have confidence in these models' dependability and power, they need to comprehend how they arrive at particular conclusions (Markus et al., 2021; Vilone &

Longo, 2020; Confalonieri et al., 2021; Ghassemi et al., 2021; Hassija et al., 2024). Understanding the internal workings of these models requires the application of XAI techniques, such as emphasising significant features, general methods, and tools for visual representation (Markus et al., 2021; Bharati et al., 2023; Saranya & Subhashini, 2023; Zhang et al., 2022).

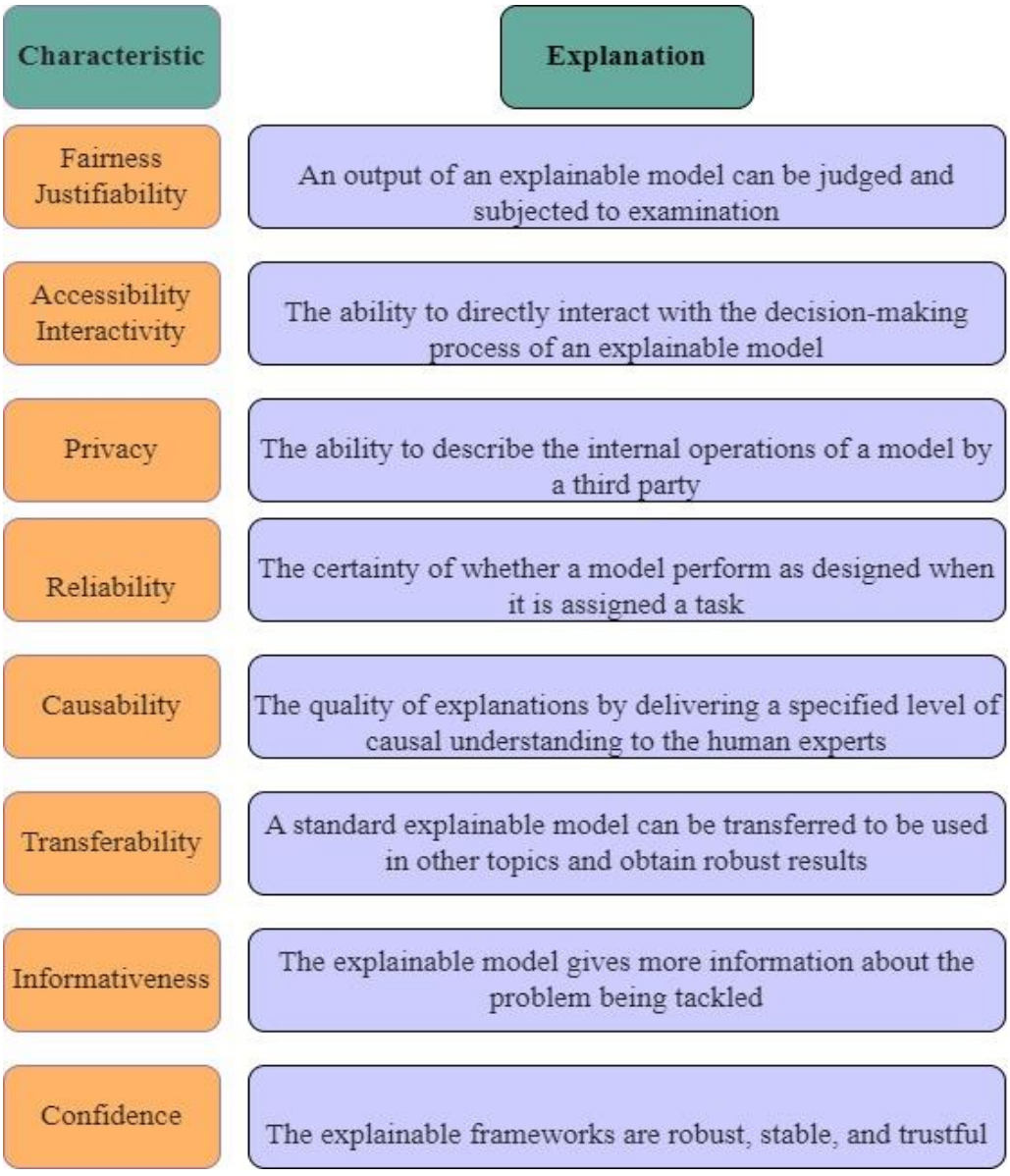

Fig 9.1 Explainable AI characteristics and definition

XAI helps improve model performance and communication between technical experts and lay users by increasing interpretability (Albahri et al., 2023; Alzubaidi et al., 2023; Jiménez-Luna et al., 2020; Novakovsky et al., 2023). Although advancements have been made in XAI studies, there are still numerous challenges and gaps that need to be addressed (Longo et al., 2020; Stepin et al., 2021; Payrovnaziri et al., 2020). This review seeks to offer a thorough examination of the present condition of XAI and its impact on reliable AI, ML, and DL systems. We thoroughly analyze the techniques, uses, and upcoming paths of XAI, focusing on important patterns and advancements in the discipline. Moreover, we delve into the overlap of XAI with ethical AI guidelines, examining how the ability to explain can aid in creating AI solutions that are more conscientious and focused on humans.

Contributions of this research:

- This study thoroughly examines previous works on XAI, highlighting important methods, uses, and areas of limited understanding.
- By conducting thorough keyword analysis and mapping co-occurrences, this study reveals common themes and new trends in the research on XAI.
- Cluster analysis is used in the research to classify and combine the various methods and viewpoints in the XAI field, providing important guidance for future research areas.

9.2 Methodology

The first phase was doing a comprehensive literature study in order to gather significant scholarly articles, conference proceedings, and reliable materials from respectable databases and journals. Google Scholar, IEEE Xplore, ACM Digital Library, Scopus, and Web of Science were the primary databases used in this search. To locate pertinent research, search phrases and keywords such as "explainable AI," "trustworthy AI," "machine learning explainability," "deep learning transparency," and related ideas were used. To guarantee that the most influential and recent research were included, the selection criteria for the literature were centred on relevance, citation count, and recentness. This phase aimed to provide a comprehensive understanding of the current state of research, fundamental theories, techniques, and applications in the field. The most often used terms and phrases in the collected material were identified by a keyword analysis carried out after the literature was reviewed. In this investigation, VOSviewer software and text mining techniques were applied. Titles, abstracts, and keywords sections of the selected publications (Holzinger et al., 2020; Vilone & Longo, 2021; Minh et al., 2022; Ahmed et al., 2022; Islam et al., 2022) were used to extract keywords. To find the major themes and patterns in the field of explainable and reliable AI, ML, and DL, the

analysis concentrated on looking at the frequency and locations of these keywords' appearances. Finding the key concepts and themes that are essential to the research field was made easier at this step.

Table 9.1 Classification of Explainable AI methods

Category	Subcategory	Explanation
Stage	Pre-model	The stage where the model description is made before the model is created
	In-model	The stage where the model description is integrated when creating the model
	Post-model	The stage where the model description is made after the model is created
Scope	Global	Description of the entire model or several models
	Local	Description of a particular sample, feature, or part of a model
Problem Type	Classification	Explaining classification problems
	Regression	Explaining regression problems
Input Data	Numerical/Categorical	Descriptions using numerical or categorical data
	Pictorial	Explanations using visual data
	Textual	Explanations using textual data
	Time series	Explanations using time series data
Output Format	Numerical	Numeric output format
	Rules	Rules-based explanations
	Textual	Textual descriptions
	Visual	Visual descriptions
	Mixed	Descriptions in mixed format (numerical, rules-based, textual and visual)
Model Agnosticism	Model Agnostic	Model-independent methods
	Model Specific	Methods specific to a particular model
Explanation Type	Surrogate	Methods that operate separately from the model or visualize the model
	Visualization	Methods that visualize the internal structure of the model or the decision-making process

Reliability	Reliable	Statements assessing the reliability and performance of the model
Privacy	Privacy-preserving	Ensuring the explainability of the internal operations of the model to a third party
Transferability	Transferable	Obtaining robust results by transferring a standard explainable model to other subjects
Informative	Informative	The explainable model provides more information about the problem at hand
Confidence	Confident	Explainable frameworks are robust, stable and reliable

Table 9.1 comprehensively presents the classification of explainable AI (XAI) methods. It details the various stages, scopes, problem types, input data types and output formats of XAI methods, explaining how different explanatory techniques are positioned and what properties they have. This classification allows XAI methods to be understood from a broad perspective and reveals how and in what situations different types of annotation are used.

In order to delve deeper into the connections between the specified keywords, a co-occurrence analysis was carried out. This study investigated the frequency of pairs of keywords appearing in the same documents, revealing connections and themes in the literature. VOSviewer was utilized to generate co-occurrence maps that visually displayed the intensity of connections between keywords. The maps showed important research areas and demonstrated the inter-connectedness of explainability and trustworthiness in AI, ML, and DL. Cluster analysis was used in the last step of the methodology to group the identified keywords and themes into cohesive categories. Cluster analysis was carried out using the co-occurrence data to pinpoint unique thematic clusters found in the literature. This study used k-means clustering and hierarchical clustering algorithms to categorize related keywords and concepts according to how often they appear together. The clusters were examined and interpreted to pinpoint key research areas and subcategories in the field.

9.3 Results and discussions

Co-occurrence and cluster analysis of the keywords used in XAI and trustworthy AI studies

The network diagram (Fig. 9.2) shows three main groups, each indicated by distinct colors: red, green, and blue. These clusters show regions where specific keywords often appear together, highlighting a thematic consistency in those regions.

Explanation about AI and security in a cluster that is colored red

Terms related to explainable AI (XAI) as well as various aspects of security and decision-making are prominently displayed in the red group. Key phrases like "machine learning," "explainable AI," "security," "network security," "cybersecurity," and "trustworthy AI" highlight the need of openness and understandability in AI systems, particularly when security is at stake. The inclusion of words like "blockchain," "decision making," "trust," and "ethics" implies that the goal of this group is to ensure that AI systems are trustworthy, ethical, and secure. Explainable AI (XAI) is crucial in this situation since it aims to make AI systems more transparent, which will eventually lead to the development of user and stakeholder trust. In security-related circumstances, it is crucial to comprehend the reasoning behind how an AI system reaches a specific decision because of the potential impact of those decisions. This grouping emphasizes the significance of incorporating XAI into AI systems to improve their dependability and approval in important fields such as cybersecurity and network security.

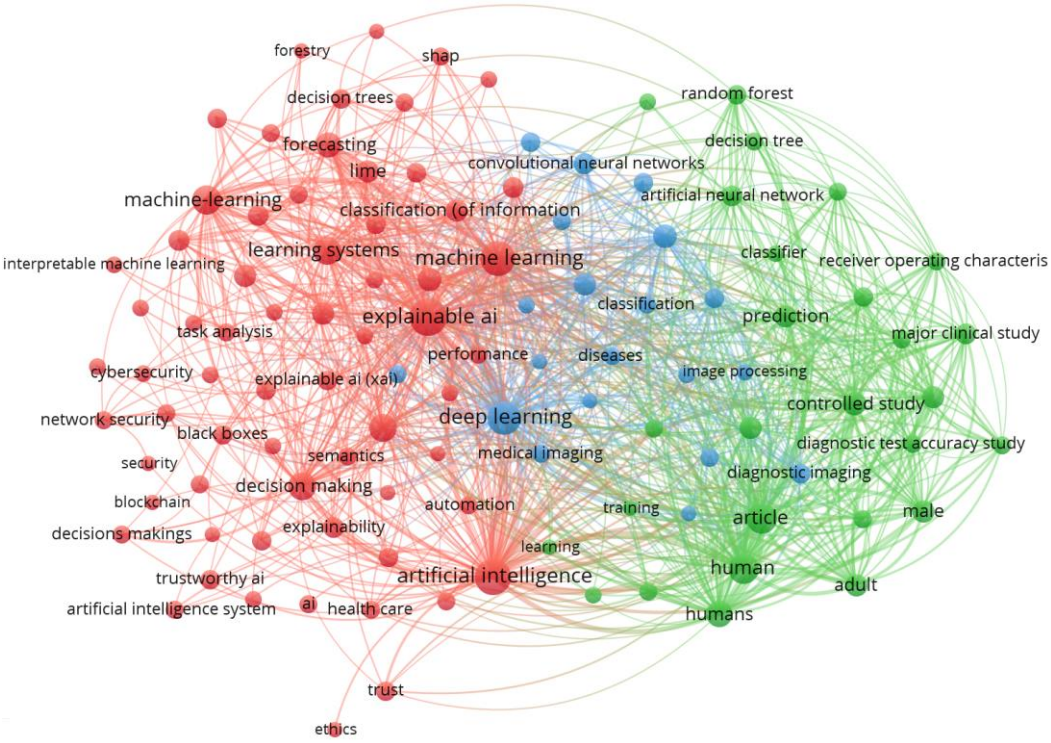

Fig. 9.2 Co-occurrence analysis of the keywords used in XAI and trustworthy AI studies

Cluster of Green: Applications in the Medical and Clinical Field

Keywords related to medical and clinical applications of AI, like "diagnostic imaging," "prediction," "controlled study," "major clinical study," "diagnostic test accuracy study," and "random forest," dominate the green cluster. This indicates a concentration on utilizing AI and machine learning methods to enhance healthcare results, especially through improved diagnostic and predictive abilities. Medical imaging greatly benefits from machine learning algorithms that can rapidly and accurately analyze large amounts of data. Phrases such as "image processing," "artificial neural network," and "classifier" emphasize the application of sophisticated AI methods in examining medical images to help with diagnosis and planning treatment. The regular appearance of these terms shows a significant interest in using AI to improve the precision and effectiveness of medical diagnostics, ultimately resulting in improved patient care.

Blue Cluster: Fundamental AI and Machine Learning Principles

The blue cluster, which contains terms like "deep learning," "convolutional neural networks," "classification," "performance," "learning," and "training," is centred on the fundamentals of artificial intelligence and machine learning. This group represents the fundamental tools and methods that enable advancements in artificial intelligence in various domains. CNNs and deep learning are essential in many AI applications because they provide the computational power needed to examine and learn from big datasets. Phrases like "medical imaging," "illnesses," and "healthcare" are included in this group, suggesting a close relationship with the green group and emphasising the important role these foundational techniques play in the advancement of medical AI.

Connections and mutual presence

The fact that these clusters are interconnected highlights how multidisciplinary AI research is. The red and blue clusters that share phrases such as "explainable AI" and "machine learning" demonstrate how advancements in fundamental AI techniques are contributing to the development of more dependable and transparent AI systems. Similarly, by utilising terminology like "medical imaging" and "classification," the connection between the blue and green groupings illustrates how fundamental AI technologies are applied to solve complex problems in healthcare. The diagram's primary placement of "machine learning" and "deep learning" emphasises their critical importance and highlights their significance in AI research. These terms act as links between various groups, demonstrating their versatility in a wide range of contexts, from healthcare to

security. This centrality also emphasizes the need for ongoing progress in these fields to facilitate a variety of uses.

Research and Development implications

Analyzing patterns of occurrences and grouping of data offers useful perspectives on ongoing research directions and possible avenues for further exploration. The increasing focus on creating AI systems in the red cluster that are both powerful and transparent is highlighted by the importance of "explainable AI" and "trustworthy AI". Researchers and professionals need to prioritize developing algorithms and structures that can clarify their reasoning in a manner that is comprehensible to people, especially in critical sectors such as security and healthcare. The emphasis on medical uses by the green cluster indicates that AI has a great chance to transform the healthcare field with better diagnostics and predictive analytics. Nevertheless, this also brings up crucial ethical concerns, like guaranteeing patient confidentiality and dealing with biases in AI algorithms. Collaboration among AI researchers, medical professionals, and ethicists will be essential for tackling these challenges. The blue group's focus on fundamental AI methods underscores the continuous requirement for studying improved and successful algorithms. Advancements in AI applications will be propelled by ongoing innovations in deep learning, neural networks, and other ML techniques. It will be crucial to guarantee accessibility and usability across various sectors to promote widespread adoption and impact as these technologies continue to advance.

Techniques for Explainability in Machine Learning and Deep Learning

The rising complexity of models in areas like healthcare, finance, and autonomous driving has made explainability in ML and deep learning a crucial research focus (Vilone & Longo, 2020; Albahri et al., 2023; Speith, 2022; Machlev et al., 2022; Confalonieri et al., 2021; Weber et al., 2024; Van der Velden et al., 2022). Clarifying means being able to explain the inner workings of a model in terms that humans can understand (Islam et al., 2022; Markus et al., 2021; Vilone & Longo, 2020; Lötsch et al., 2021; Alicioglu & Sun, 2022; Giuste et al., 2022). Maintaining transparency is crucial for establishing trust, meeting regulatory requirements, and easing the process of debugging and enhancing models (Chakrobartty & El-Gayar, 2021; Hauser et al., 2022; Schwalbe & Finzel, 2023).

Model-Agnostic Methods

Techniques that are not dependent on models are designed to provide explanations for any opaque model without requiring changes to its internal workings. Shapley Additive exPlanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME) are the most well-known examples of these. LIME works by making minor adjustments to

the input data and seeing how those changes affect the output. In doing so, it approximates the model's behaviour in the vicinity of the explained case. This method works with any kind of machine learning model and is particularly useful for analysing individual predictions. Based on cooperative game theory, SHAP values provide a thorough evaluation of the importance of features. They provide reliable and theoretically sound explanations by assigning the forecast to specific characteristics. SHAP values are summed up and offer a detailed explanation of the impact of each feature on the ultimate prediction, making it a flexible tool for explaining both individual instances and overall trends.

Post-hoc Analysis

Post-hoc analysis involves methods used to interpret and comprehend the model's decisions after it has been trained. These techniques are especially valuable for intricate DL models like convolutional neural networks (CNNs) and recurrent neural networks (RNNs). Feature Visualization entails the visualization of the features that have been acquired by a model. In convolutional neural networks, this can be achieved by creating activation maximization visuals that display the input patterns that activate specific neurons the most. This aids in comprehending the specific characteristics that each neuron is seeking within the input data. Saliency Maps indicate which areas of the input data have the greatest impact on the model's prediction. Methods such as Gradient-weighted Class Activation Mapping (Grad-CAM) enhance this capability by generating saliency maps that are specific to a particular class, revealing the main contributing regions of an image for that classification. Layer-wise Relevance Propagation (LRP) is an additional method mainly utilized in neural networks. LRP functions by sending the prediction back through the network layers, reallocating the prediction score to the input characteristics based on their significance. This assists in determining which characteristics are the most important for a specific forecast. Table 9.2 shows the techniques for explainability in ML and deep learning.

Intrinsic Interpretability

The goal of inherent interpretability is to develop models that are inherently simple to comprehend. These models are designed with an easy-to-understand decision-making process in mind. A classic example of a model that is naturally interpretable is a decision tree. Understanding a decision tree's decision-making process is as simple as tracing the path from the root to the leaf nodes. Similarly, linear models that provide clear insights into how each feature affects the prediction include logistic regression and linear regression. With the ability to allow for non-linear relationships between predictors and the response variable while maintaining interpretability, generalised additive models

(GAMs) extend linear models. To do this, GAMs first simulate each feature's relationship to the target separately, and then they combine these influences.

Table 9.2 Techniques for explainability in ML and deep learning

References	Technique	Description	Examples	Pros
(Kaur et al., 2022; Angelov et al., 2021; Rawal et al., 2021; Albahri et al., 2023; Speith, 2022; Nazar et al., 2021; Ali et al., 2023)	Model-Agnostic Methods	Techniques that are applicable across any ML model irrespective of its architecture or type.	LIME (Local Interpretable Model-agnostic Explanations), SHAP (SHapley Additive exPlanations), Partial Dependence Plots (PDP)	Versatile in nature, applicable to a myriad of models
(Angelov et al., 2021; Albahri et al., 2023; Hassija et al., 2024; Bharati et al., 2023; Saranya & Subhashini, 2023)	Intrinsic Methods	Approaches that inherently incorporate interpretability within the model's design.	Decision Trees, Linear Regression, Rule-Based Models	Intuitive and straightforward, inherently interpretable
(Minh et al., 2022; Vilone & Longo, 2020; Albahri et al., 2023; Speith, 2022)	Post-Hoc Interpretability Methods	Techniques employed post model training to elucidate predictions.	Feature Importance, Counterfactual Explanations, Saliency Maps	Capable of providing insights into individual predictions, applicable to pre-trained models
(Nassar et al., 2020; Machlev et al., 2022; Confalonieri et al., 2021; Ghassemi et al., 2021)	Visualization Techniques	Methods utilizing visual tools to aid in model interpretation.	t-SNE (t-distributed Stochastic Neighbor Embedding), PCA (Principal Component Analysis), Heatmaps	Visually intuitive, effective for identifying patterns and anomalies

(Ali et al., 2023; Nassar et al., 2020; Arrieta et al., 2020; 40)	Surrogate Models	Simplified, interpretable models that approximate the behavior of more complex models.	Training a Decision Tree to mimic a Neural Network, Training a Linear Model to approximate a Random Forest	Provides insight into the behavior of complex models, useful for diagnostic purposes
(Albahri et al., 2023; Speith, 2022; Nazar et al., 2021; Vilone & Longo, 2021; Das & Rad, 2020; Kaur et al., 2021)	Example-Based Explanations	Techniques leveraging examples to elucidate model predictions.	Prototypes and Criticisms, Case-Based Reasoning, K-Nearest Neighbors (KNN)	Offers concrete instances for interpretation, easily comprehensible for users
(Rawal et al., 2021; Islam et al., 2022; Markus et al., 2021; Albahri et al., 2023; Speith, 2022)	Self-Explaining Models	Models designed to offer their own explanations for their predictions.	Attention Mechanisms in Neural Networks, Self-Explaining Neural Networks	Intrinsically interpretable, capable of providing both predictions and explanations concurrently
(Nazar et al., 2021; Ali et al., 2023; Nassar et al., 2020; Zhang et al., 2022; Alzubaidi et al., 2023)	Feature Attribution Methods	Techniques attributing the significance of individual features to the model's predictions.	Integrated Gradients, Gradient Shapley, Layer-Wise Relevance Propagation (LRP)	Offers detailed insights into feature importance, essential for understanding the driving factors behind predictions
(Payrovnaziri et al., 2020; Weber et al., 2024; Van der Velden et al., 2022; Schwalbe & Finzel, 2023)	Hybrid Methods	Integrating multiple interpretability techniques to leverage their strengths while mitigating their weaknesses.	Combining SHAP with LIME, Utilizing both PDP and Feature Importance for comprehensive analysis	Provides a holistic understanding, balances trade-offs inherent in different techniques

Explainable Neural Networks

Despite the common perception of classic neural networks as opaque, recent efforts have yielded interpretable neural network architectures. Attention mechanisms provide a way to interpret which particular regions of the input data the model is focussing on. Examples of these models include transformers and attention-based RNNs. The attention weights visualisation helps us to comprehend how the model evaluates and prioritises different parts of the input. The explainability of linear models and the adaptability of neural networks are combined in Self-Explaining Neural Networks (SENN). With the help of SENNs, forecasts may be broken down into easily understood components, making it easier to evaluate how each element affects the final prediction.

Counterfactual Explanations

Understanding how a model's forecast would change if certain input data points were altered is possible using counterfactual explanations. Understanding decision boundaries and identifying potential biases in the model are two areas in which this technique really shines. A counterfactual explanation in a loan approval model could show that the loan would be approved if the applicant made $10,000 more. In order to ensure fairness and openness, this type of explanation helps users understand the elements that go into the model's judgements.

Causal Inference

Rather of focussing only on links, causal inference approaches aim to establish cause-and-effect correlations. These techniques come in handy when understanding the cause-and-effect link between traits and outcomes is crucial. For the goal of expressing and researching causal relationships, causal graphs and structural equation models (SEMs) are used. We are able to more thoroughly examine the effects of different features on the outcome by incorporating knowledge from a particular field into these models. With the help of the popular Python causal inference module DoWhy, users can identify, compute, and refute causal relationships by providing a comprehensive framework for causal analysis.

Rule-based Methods

Human-understandable rules provide arguments in rule-based approaches. These methods are particularly helpful in circumstances where people need direction that is both clear and useful. Models such as decision trees can be utilised to extract decision rules, or association rule mining techniques can be applied to create them. These rules specify the conditions under which particular projections are produced, offering a clear and simple way to comprehend the model's output. RuleFit is an ensemble method that combines

rule-based and linear models. By generating rules from tree-based models and utilising these rules to build a sparse linear model, it provides a balance between interpretability and accuracy.

Prototype and Criticism-based Methods

These methods provide insights by identifying representative examples (prototypes) and examples that highlight the shortcomings of the model (criticisms). Few-shot learning makes use of prototype networks to identify class prototypes. By serving as typical examples, these models help others understand what each society considers to be normal. Criticism based strategies identify situations in which one would expect the model's forecasts to be less reliable. We can discover the model's shortcomings and pinpoint areas that can benefit from further improvements by looking at these criticisms.

Interactive and Visual Explanations

To make sure that explanations are understandable and accessible to consumers, interactive tools and visualisations are crucial. Users are able to examine feature significance and model behaviour directly with the use of tools such as TensorBoard and LIME's interactive visualisations. These resources provide an easy-to-use means of investigating and comprehending model predictions. Explainability Dashboards provide a comprehensive overview of the operations of the model by combining various explainability techniques into a single interface. These dashboards offer saliency maps, feature importance ratings, and counterfactual explanations to aid in the thorough understanding and interpretation of model decisions.

Ethical and Fair AI

Explainability requires, among other things, that AI models be impartial and free from bias. Techniques for detecting and mitigating bias are essential for creating trustworthy AI systems. Algorithms that consider fairness adjust the training process to ensure that the model's predictions are impartial towards particular groups. To attain more equitable outcomes, techniques such as modifying the training data's weights, putting adversarial debiasing into practice, and utilising fairness restrictions are helpful. Measurements and charts are provided by bias detection tools, such as AI Fairness 360 and Fairness Indicators, to help identify and quantify biases in machine learning models. These tools make it easier to understand how different traits and subgroups are treated by the model, which promotes the development of more egalitarian AI systems.

Trustworthy AI

One essential component of reliable AI is transparency (Kaur et al., 2022; Angelov et al., 2021). Ensuring users and stakeholders understand how AI systems work is a prerequisite for transparency. Explainability, as defined by Zhang et al. (2022; Zhang et al., 2022; Alzubaidi et al., 2023; 40), entails providing justifications for the judgements made by AI models. The need to explain complex algorithms and foster user trust has prompted the AI community to place a high priority on Explainable AI (XAI) (Angelov et al., 2021; Rawal et al., 2021; Islam et al., 2022; Vilone & Longo, 2021). Progress in XAI techniques, such as building interpretable models and post-hoc justifications, has enhanced the ability to analyse AI decisions. Examples include tools such as Local Interpretable Model-agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP) that enable individuals to comprehend the factors that impacted a specific decision. These strategies are especially important in critical areas like healthcare, finance, and criminal justice, where unclear AI choices can have significant consequences (Kaur et al., 2022; Albahri et al., 2023; Speith, 2022; Vilone & Longo, 2021).

Accountability in AI involves having systems in position to guarantee AI operates within legal and ethical limits and provides a way to address issues when they occur. Creating accountability requires determining specific duties and obligations for AI creators, users, and overseers. It also requires strong governance frameworks that specify the procedures for auditing and supervision. Recent trends indicate the increased focus on regulatory frameworks and standards to oversee the ethical advancement and implementation of AI. For instance, the AI Act proposed by the European Union focuses on creating a legal structure for AI, sorting applications by their level of risk, and implementing strict regulations for high-risk AI systems. Regulatory initiatives play a vital role in ensuring organizations are held accountable and that AI technologies comply with ethical principles and legal norms.

Ensuring fairness in AI is essential for building trust, as it involves minimizing biases that may result in discrimination. AI systems that are trained on biased data have the potential to continue or worsen current inequalities. Hence, guaranteeing fairness requires implementing both technical and procedural steps to detect and address biases in AI models. Recent studies in AI ethics have concentrated on creating methods to identify and address bias. This consists of techniques to preprocess data and eliminate biases, techniques to adjust algorithms during training, and methods to modify outputs for fair outcomes. Furthermore, there is an increasing acknowledgment of the significance of varied datasets and inclusive methodologies in the development of AI to reduce biases from the beginning. The rise of AI systems has led to increased worries regarding privacy and data security. Trustworthy AI must guarantee responsible handling of personal data and protection of users' privacy. This includes following data protection laws like the

General Data Protection Regulation (GDPR) in Europe, which sets strict rules for how data is collected, processed, and stored. Privacy-focused AI methods, like federated learning and differential privacy, are increasingly being used as ways to deal with these issues. Federated learning allows AI models to be trained on multiple separate devices without centralizing data, which helps to improve privacy. Differential privacy involves adding random noise to data to safeguard individual privacy without compromising the ability to perform meaningful analysis. These methods are crucial for upholding user confidence and guaranteeing that AI systems uphold privacy rights.

AI safety involves making sure AI systems are dependable and strong, working as planned and able to withstand attacks and unexpected situations. Trustworthy AI needs to be created to navigate various situations, even ones that are uncommon or unforeseen. Current developments in AI resilience concentrate on enhancing the ability of models to withstand attacks from adversaries, specifically through small alterations to input data that may disrupt the functionality of AI systems. Methods like adversarial training, where models are trained with adversarial examples, and robust optimization techniques have been created to improve the security of AI. Moreover, it is essential to conduct ongoing monitoring and thorough testing in order to guarantee the continued safety and reliability of AI systems. AI systems need to align with ethical principles and human values, in addition to being technically robust, in order to be considered trustworthy. This involves making sure that AI technologies are created and used in ways that honor human dignity, independence, and rights. Centering on humans, AI design highlights the significance of engaging users and stakeholders during development to ensure AI systems meet their needs and address their concerns. Collaborative design methods involving end-users are being more widely acknowledged as successful strategies for achieving this objective. AI developers can build systems that are socially responsible by giving importance to human values and ethics alongside technological advancement.

Several new developments are influencing the future of trustworthy AI as AI progresses. A significant development is the incorporation of AI ethics into educational curricula and professional training programs. By teaching ethical principles and responsible practices to the future AI practitioners, we are setting the groundwork for a more mindful AI community. One more trend involves utilizing AI for positive societal impact, harnessing AI tools to tackle worldwide issues like climate change, healthcare, and education. These projects showcase how AI can have a positive impact on society if it is developed and used responsibly. Furthermore, the need for interdisciplinary collaboration is growing more significant in the realm of trustworthy AI. Collaborating with specialists in various fields like computer science, law, philosophy, and social sciences promotes a

comprehensive strategy for dealing with the intricate ethical, legal, and societal impacts of AI.

Evaluation Metrics and Benchmarks for Trustworthy AI

AI has made notable advancements in various sectors like healthcare and finance in recent years (Kaur et al., 2022; Angelov et al., 2021; Rawal et al., 2021; Markus et al., 2021; Vilone & Longo, 2020; Albahri et al., 2023). Nevertheless, there are worries about the reliability of AI systems due to issues with fairness, transparency, accountability, and robustness (Rawal et al., 2021; Speith, 2022; Machlev et al., 2022). Developing in-depth evaluation metrics and benchmarks has become essential in order to tackle these issues (Saranya & Subhashini, 2023; Zhang et al., 2022; Zhang et al., 2022). Fig. 9.2 shows the evaluation metrics and benchmarks for trustworthy AI.

Fairness Metrics

The term "fairness in AI" describes the objectivity with which AI systems reach judgements. A crucial component of fairness is demographic parity, which ensures that the results are distributed equally among the different demographic groups. Metrics like statistical parity difference, differential impact, and equalised odds can be used to measure this more precisely. The emphasis on fairness-aware machine learning has led to the development of new tools, such as Google's What-If Tool and IBM's AI Fairness 360, which help evaluate and minimise bias in AI models. Equity is a crucial component of representation learning, as it aims to produce embeddings that are fair to different groups. To this end, techniques including fairness-aware data preparation and adversarial debiasing are applied. Recent studies have also investigated the idea of intersectional equity, acknowledging that people may be part of various disadvantaged groups, meaning that equity measures need to consider these intersections to prevent amplifying biases.

Transparency and Explainability Metrics

Building confidence in AI systems requires openness and the capacity to explain decisions. Explainability measures assess how simple it is for people to understand how an AI system makes decisions. Various metrics and methods have been introduced in explainable AI (XAI) recently to evaluate this aspect. Commonly utilised techniques that provide insight into individual predictions made by complex models are LIME and SHAP. Metrics like completeness, which gauges how much of the model's behaviour is captured in the explanations, and fidelity, which assesses how well explanations match model predictions, can be used to assess how well models are interpreted. Furthermore, there's a growing tendency of using human-centered user research and assessments to gauge how well explanations help consumers understand and feel more confident.

Robustness Metrics

The ability of AI systems to resist hostile attacks and disruptions is referred to as robustness. When assessing an AI model's resilience, it is crucial to evaluate its performance in various hostile scenarios. The ability of an AI system to withstand hostile inputs is measured using metrics like verified robustness, adversarial robustness score, and accuracy against attacks. In order to increase the resilience of AI models, recent research has focused on developing effective training techniques including adversarial training and defensive distillation. Benchmark datasets like the Adversarial Robustness Toolbox (ART) and the CleverHans library provide standardised techniques to evaluate and compare the resilience of different AI systems. To ensure secure and reliable deployment in real-world scenarios, robustness evaluation must be integrated into the AI model building process.

Accountability Metrics

Ensuring that the AI system can be held accountable for its activities and tracking decisions back to their source are essential components of accountability in AI. Important evaluation indicators for accountability include the capacity to easily evaluate and verify judgements made by AI systems, as well as the tracking of origin and decision-making phases. Technological developments in blockchain have created new avenues for enhancing AI accountability. Blockchain can be used to create permanent records of the decisions and interactions performed by AI systems, ensuring transparency and monitoring capabilities. Furthermore, stricter accountability standards are being promoted by regulatory guidelines like the AI Act of the European Union, which calls for the development of robust assessment standards in order to comply with the regulations.

Ethical and Social Impact Metrics

Assessing the ethical and social consequences of AI systems is becoming more important as stakeholders acknowledge the wider effects of utilizing AI. Measurements within this field evaluate how well AI systems adhere to ethical principles and societal values. One example is the AI Ethics Impact Assessment (AIEIA) framework, which offers a systematic method to assess the ethical consequences of AI systems by taking into account aspects such as human rights, societal welfare, and environmental sustainability. Furthermore, social impact measures, like the ones suggested by the IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, provide criteria for evaluating the wider societal impacts of AI integration. These measurements assist in guaranteeing that AI systems benefit society in a positive way and do not worsen current inequalities or introduce fresh ethical dilemmas.

Benchmarking Trustworthy AI

Benchmarking is crucial for establishing a standard way to evaluate AI systems and guaranteeing that they can be compared with each other across various models and uses. The AI community has created multiple benchmarking frameworks for evaluating the reliability of AI systems. For instance, the TRUST-AI framework offers a thorough collection of benchmarks to assess the fairness, transparency, robustness, and accountability of AI models. IBM's AI Explainability 360 (AIX360) toolkit provides various datasets and metrics to assess the explainability of AI models. In the same way, the AIRE framework provides standardized techniques for evaluating the resilience of AI systems to adversarial attacks. These tools for benchmarking play a crucial role in setting initial performance levels and leading the creation of more reliable AI systems.

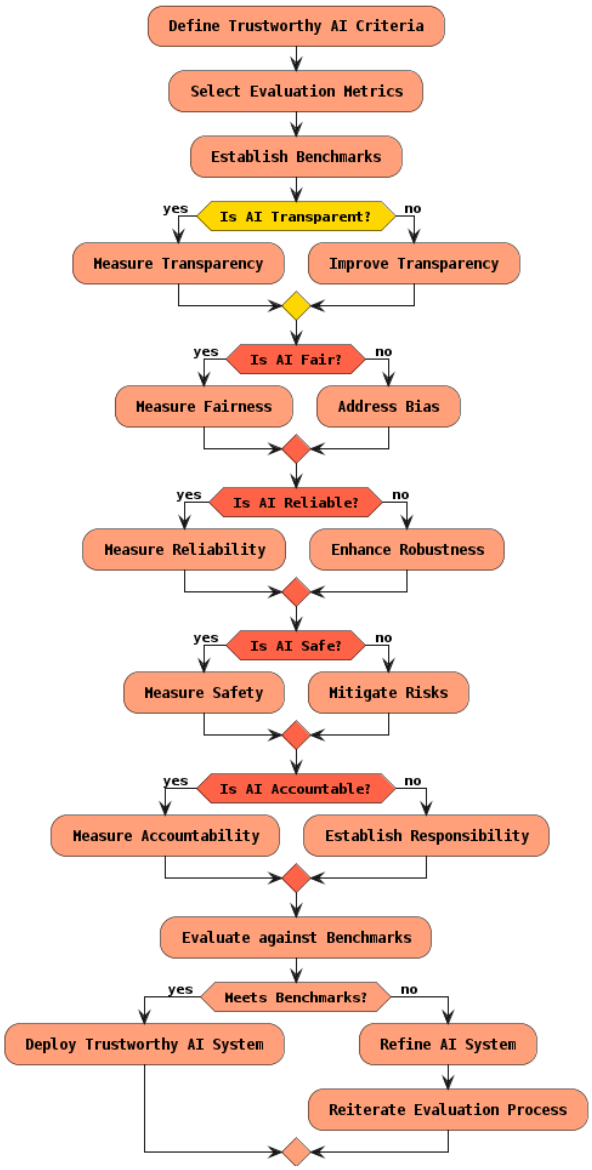

Fig. 9.2 Evaluation metrics and benchmarks for trustworthy AI

Human-Centered AI and User Trust

Human-centered AI (HCAI) is a revolutionary method that focuses on creating AI systems that prioritize human values, ethics, and needs. Incorporating HCAI into the creation and implementation of AI technologies is becoming more important to guarantee that AI systems are efficient, effective, trusted, and accepted by users. The basic principle of HCAI is to design AI systems that improve human abilities and decision-making. This

includes creating AI technologies that are clear, understandable, and in line with ethical standards important to humans. Trust in AI systems is greatly enhanced when users are able to comprehend the decision-making process and have assurance that decisions are fair and ethical. Transparency and explainability play key roles by providing users with a deeper understanding of how AI makes decisions, which in turn builds trust and a feeling of being in control. A recent development in the field of HCAI is the emergence of explainable AI (XAI) methods. XAI's goal is to enhance the interpretability of AI systems, allowing users to understand the process by which specific results are achieved. In the healthcare industry, XAI can assist medical professionals in comprehending the logic behind AI-made diagnostic suggestions, ultimately enhancing their confidence in the technology. Transparent AI decisions increase trust and reliance on these systems, crucial for wider acceptance and integration of AI technologies across different fields.

Another crucial element of HCAI involves ensuring the ethical development of AI systems. Ethical AI guarantees that AI technologies are developed and utilized in accordance with ethical standards and societal norms. This involves tackling concerns like prejudice, equality, and responsibility. AI systems need to be created to reduce biases and guarantee equitable treatment for all users. In recruitment processes, it is important to carefully examine AI algorithms to prevent any discrimination related to race, gender, or other demographic factors. HCAI promotes trust among users by emphasizing ethical considerations, ensuring that AI systems prioritize fairness and integrity. Trust in AI from users is also dependent on the idea of designing with the user in mind. This method includes actively engaging users in the design and development of AI systems. Developers can acquire important information about user needs, preferences, and concerns by involving users in participatory design techniques. This cooperative method guarantees that the AI systems created meet user expectations and needs, leading to increased user satisfaction and trust. For example, when designing smart home devices, including users in the design process can result in the development of interfaces that are easier to use and understand, ultimately building trust and encouraging greater acceptance.

Moreover, the continuous progress in AI technologies is more and more centered on protecting the privacy and security of user data. Users' perceptions of data handling heavily impact trust in AI systems. It is essential to have strong data protection measures and give users control over their personal information in order to establish trust. AI systems that provide transparent privacy policies and tools for users to control their data are more likely to gain users' trust. With increasing worries about data privacy, it is crucial to incorporate robust data protection measures into AI systems. The importance of regulation and governance should not be ignored when considering HCAI and building

trust with users. Governments and regulators have an important role in setting rules and standards for the ethical use of AI. These rules guarantee that AI systems are created and implemented in a way that aligns with societal values and legal obligations. Regulatory structures that support transparency, accountability, and fairness in AI have the potential to greatly increase trust from users. As an example, the European Union's General Data Protection Regulation (GDPR) enforces rules on data protection, affecting AI systems that process personal data. Adhering to these regulations doesn't just guarantee legal compliance, it also fosters user trust in the ethical application of AI.

9.4 Conclusions

The assessment of Explainable AI (XAI) and trustworthy AI, ML, and DL displays notable progress and persistent obstacles in developing transparent, accountable, and dependable AI systems. The need for transparency in AI has grown as AI becomes more prevalent in vital industries like construction, healthcare, finance, and autonomous systems. XAI techniques have the objective of unveiling the decision-making procedures of intricate models, in order to build user confidence and assist in meeting regulatory requirements. The utilization of XAI has demonstrated potential in enhancing the comprehensibility of opaque models like neural networks and ensemble methods by methods such as feature attribution, surrogate models, and visual explanations. Furthermore, the focus on trustworthy AI includes not just transparency but also fairness, resilience, and confidentiality. Guaranteeing fairness requires tackling biases in data and algorithms that may result in discriminatory results. Robustness concerns the ability of AI systems to withstand adversarial attacks and perturbations, while privacy-preserving methods like differential privacy and federated learning work to safeguard sensitive data in AI processes. In spite of these advancements, the field encounters several challenges. Finding a balance between model performance and interpretability is still a key concern, since models that are highly interpretable may not always reach the highest level of accuracy. Moreover, the ability of XAI methods to be applied in real-world, large-scale scenarios presents technical and computational obstacles. Moreover, it is crucial to develop standard criteria and benchmarks for assessing the transparency and reliability of various AI systems in a consistent manner. Sustained cooperation among researchers, practitioners, and policymakers is essential for promoting the progress of transparent, equitable, and strong AI systems that build public confidence and stimulate innovation.

References

Ahmed, I., Jeon, G., & Piccialli, F. (2022). From artificial intelligence to explainable artificial intelligence in industry 4.0: a survey on what, how, and where. IEEE Transactions on Industrial Informatics, 18(8), 5031-5042.

Albahri, A. S., Duhaim, A. M., Fadhel, M. A., Alnoor, A., Baqer, N. S., Alzubaidi, L., ... & Deveci, M. (2023). A systematic review of trustworthy and explainable artificial intelligence in healthcare: Assessment of quality, bias risk, and data fusion. Information Fusion.

Ali, S., Abuhmed, T., El-Sappagh, S., Muhammad, K., Alonso-Moral, J. M., Confalonieri, R., ... & Herrera, F. (2023). Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. Information fusion, 99, 101805.

Alicioglu, G., & Sun, B. (2022). A survey of visual analytics for explainable artificial intelligence methods. Computers & Graphics, 102, 502-520.

Alzubaidi, L., Al-Sabaawi, A., Bai, J., Dukhan, A., Alkenani, A. H., Al-Asadi, A., ... & Gu, Y. (2023). Towards Risk-Free Trustworthy Artificial Intelligence: Significance and Requirements. International Journal of Intelligent Systems, 2023(1), 4459198.

Angelov, P. P., Soares, E. A., Jiang, R., Arnold, N. I., & Atkinson, P. M. (2021). Explainable artificial intelligence: an analytical review. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 11(5), e1424.

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., ... & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information fusion, 58, 82-115.

Bharati, S., Mondal, M. R. H., & Podder, P. (2023). A review on explainable artificial intelligence for healthcare: why, how, and when?. IEEE Transactions on Artificial Intelligence.

Chakrobartty, S., & El-Gayar, O. (2021). Explainable artificial intelligence in the medical domain: a systematic review.

Confalonieri, R., Coba, L., Wagner, B., & Besold, T. R. (2021). A historical perspective of explainable Artificial Intelligence. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 11(1), e1391.

Das, A., & Rad, P. (2020). Opportunities and challenges in explainable artificial intelligence (xai): A survey. arXiv preprint arXiv:2006.11371.

Ghassemi, M., Oakden-Rayner, L., & Beam, A. L. (2021). The false hope of current approaches to explainable artificial intelligence in health care. The Lancet Digital Health, 3(11), e745-e750.

Giuste, F., Shi, W., Zhu, Y., Naren, T., Isgut, M., Sha, Y., ... & Wang, M. D. (2022). Explainable artificial intelligence methods in combating pandemics: A systematic review. IEEE Reviews in Biomedical Engineering, 16, 5-21.

Hassija, V., Chamola, V., Mahapatra, A., Singal, A., Goel, D., Huang, K., ... & Hussain, A. (2024). Interpreting black-box models: a review on explainable artificial intelligence. Cognitive Computation, 16(1), 45-74.

Hauser, K., Kurz, A., Haggenmüller, S., Maron, R. C., von Kalle, C., Utikal, J. S., ... & Brinker, T. J. (2022). Explainable artificial intelligence in skin cancer recognition: A systematic review. European Journal of Cancer, 167, 54-69.

Holzinger, A., Goebel, R., Fong, R., Moon, T., Müller, K. R., & Samek, W. (2020). xxAI-beyond explainable artificial intelligence. In International Workshop on Extending Explainable AI Beyond Deep Models and Classifiers (pp. 3-10). Cham: Springer International Publishing.

Islam, M. R., Ahmed, M. U., Barua, S., & Begum, S. (2022). A systematic review of explainable artificial intelligence in terms of different application domains and tasks. Applied Sciences, 12(3), 1353.

Jiménez-Luna, J., Grisoni, F., & Schneider, G. (2020). Drug discovery with explainable artificial intelligence. Nature Machine Intelligence, 2(10), 573-584.

Kaur, D., Uslu, S., & Durresi, A. (2021). Requirements for trustworthy artificial intelligence–a review. In Advances in Networked-Based Information Systems: The 23rd International Conference on Network-Based Information Systems (NBiS-2020) 23 (pp. 105-115). Springer International Publishing.

Kaur, D., Uslu, S., Rittichier, K. J., & Durresi, A. (2022). Trustworthy artificial intelligence: a review. ACM computing surveys (CSUR), 55(2), 1-38.

Loh, H. W., Ooi, C. P., Seoni, S., Barua, P. D., Molinari, F., & Acharya, U. R. (2022). Application of explainable artificial intelligence for healthcare: A systematic review of the last decade (2011–2022). Computer Methods and Programs in Biomedicine, 226, 107161.

Longo, L., Goebel, R., Lecue, F., Kieseberg, P., & Holzinger, A. (2020). Explainable artificial intelligence: Concepts, applications, research challenges and visions. In International cross-domain conference for machine learning and knowledge extraction (pp. 1-16). Cham: Springer International Publishing.

Lötsch, J., Kringel, D., & Ultsch, A. (2021). Explainable artificial intelligence (XAI) in biomedicine: Making AI decisions trustworthy for physicians and patients. BioMedInformatics, 2(1), 1-17.

Machlev, R., Heistrene, L., Perl, M., Levy, K. Y., Belikov, J., Mannor, S., & Levron, Y. (2022). Explainable Artificial Intelligence (XAI) techniques for energy and power systems: Review, challenges and opportunities. Energy and AI, 9, 100169.

Markus, A. F., Kors, J. A., & Rijnbeek, P. R. (2021). The role of explainability in creating trustworthy artificial intelligence for health care: a comprehensive survey of the terminology, design choices, and evaluation strategies. Journal of biomedical informatics, 113, 103655.

Minh, D., Wang, H. X., Li, Y. F., & Nguyen, T. N. (2022). Explainable artificial intelligence: a comprehensive review. Artificial Intelligence Review, 1-66.

Nassar, M., Salah, K., ur Rehman, M. H., & Svetinovic, D. (2020). Blockchain for explainable and trustworthy artificial intelligence. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 10(1), e1340.

Nazar, M., Alam, M. M., Yafi, E., & Su'ud, M. M. (2021). A systematic review of human–computer interaction and explainable artificial intelligence in healthcare with artificial intelligence techniques. IEEE Access, 9, 153316-153348.

Novakovsky, G., Dexter, N., Libbrecht, M. W., Wasserman, W. W., & Mostafavi, S. (2023). Obtaining genetics insights from deep learning via explainable artificial intelligence. Nature Reviews Genetics, 24(2), 125-137.

Payrovnaziri, S. N., Chen, Z., Rengifo-Moreno, P., Miller, T., Bian, J., Chen, J. H., ... & He, Z. (2020). Explainable artificial intelligence models using real-world electronic health record data: a systematic scoping review. Journal of the American Medical Informatics Association, 27(7), 1173-1185.

Rawal, A., McCoy, J., Rawat, D. B., Sadler, B. M., & Amant, R. S. (2021). Recent advances in trustworthy explainable artificial intelligence: Status, challenges, and perspectives. IEEE Transactions on Artificial Intelligence, 3(6), 852-866.

Saranya, A., & Subhashini, R. (2023). A systematic review of Explainable Artificial Intelligence models and applications: Recent developments and future trends. Decision analytics journal, 100230.

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

Speith, T. (2022). A review of taxonomies of explainable artificial intelligence (XAI) methods. In Proceedings of the 2022 ACM conference on fairness, accountability, and transparency (pp. 2239-2250).

Stepin, I., Alonso, J. M., Catala, A., & Pereira-Fariña, M. (2021). A survey of contrastive and counterfactual explanation generation methods for explainable artificial intelligence. IEEE Access, 9, 11974-12001.

Van der Velden, B. H., Kuijf, H. J., Gilhuijs, K. G., & Viergever, M. A. (2022). Explainable artificial intelligence (XAI) in deep learning-based medical image analysis. Medical Image Analysis, 79, 102470.

Vilone, G., & Longo, L. (2020). Explainable artificial intelligence: a systematic review. arXiv preprint arXiv:2006.00093.

Vilone, G., & Longo, L. (2021). Notions of explainability and evaluation approaches for explainable artificial intelligence. Information Fusion, 76, 89-106.

Weber, P., Carl, K. V., & Hinz, O. (2024). Applications of explainable artificial intelligence in finance—a systematic review of finance, information systems, and computer science literature. Management Review Quarterly, 74(2), 867-907.

Zhang, Y., Weng, Y., & Lund, J. (2022). Applications of explainable artificial intelligence in diagnosis and surgery. Diagnostics, 12(2), 237.

Zhang, Z., Al Hamadi, H., Damiani, E., Yeun, C. Y., & Taher, F. (2022). Explainable artificial intelligence applications in cyber security: State-of-the-art in research. IEEE Access, 10, 93104-93139.