Knowledge-Based Systems 263 (2023) 110273

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

Explainable AI (XAI): A systematic meta-survey of current challenges and future opportunities

Waddah Saeed [a,b,*], Christian Omlin [a]

[a] Center for Artificial Intelligence (CAIR), University of Agder, Jon Lilletuns vei 9, Grimstad, 4879, Agder, Norway
[b] School of Computer Science and Informatics, De Montfort University, Leicester, LE1 9BH, UK

ARTICLE INFO

Article history:
Received 16 May 2022
Received in revised form 2 January 2023
Accepted 3 January 2023
Available online 11 January 2023

Keywords:
Explainable AI (XAI)
Interpretable AI
Black-box
Machine learning
Deep learning
Meta-survey
Responsible AI

ABSTRACT

The past decade has seen significant progress in artificial intelligence (AI), which has resulted in algorithms being adopted for resolving a variety of problems. However, this success has been met by increasing model complexity and employing black-box AI models that lack transparency. In response to this need, Explainable AI (XAI) has been proposed to make AI more transparent and thus advance the adoption of AI in critical domains. Although there are several reviews of XAI topics in the literature that have identified challenges and potential research directions of XAI, these challenges and research directions are scattered. This study, hence, presents a systematic meta-survey of challenges and future research directions in XAI organized in two themes: (1) general challenges and research directions of XAI and (2) challenges and research directions of XAI based on machine learning life cycle's phases: design, development, and deployment. We believe that our meta-survey contributes to XAI literature by providing a guide for future exploration in the XAI area.

1. Introduction

Artificial intelligence (AI) has undergone significant and continuous progress in the past decade, resulting in the increased adoption of its algorithms (e.g., machine learning (ML) algorithms) for solving many problems, even those that were difficult to resolve in the past. However, these outstanding achievements are accompanied by increasing model complexity and utilizing black-box AI models that lack transparency. Therefore, it becomes necessary to come up with solutions that can contribute to addressing such a challenge, which could help expand utilizing AI systems in critical and sensitive domains (e.g., healthcare and security domains) where other criteria must be met besides the high accuracy.

Explainable artificial intelligence (XAI) has been proposed as a solution that can help to move towards more transparent AI and thus avoid limiting the adoption of AI in critical domains [1,2]. Generally speaking, according to [3], XAI focuses on developing explainable techniques that empower end-users in comprehending, trusting, and efficiently managing the new age of AI systems. Historically, the need for explanations dates back to the early works in explaining expert systems and Bayesian networks [4]. Deep learning (DL), however, has made XAI a thriving research area.

Every year, a large number of studies dealing with XAI are published. At the same time, various review studies are published covering a range of general or specific aspects of XAI. With many of these review studies, several challenges and research directions are discussed. While this has led to identifying challenges and potential research directions, however, they are scattered.

To the best of our knowledge, this is the first meta-survey that explicitly organizes and reports on the challenges and potential research directions of XAI. This meta-survey aims to provide a reference point for researchers interested in working on challenges and potential research directions in XAI.

The organization of the paper is as follows (also shown in Fig. 1). In Section 2, we discuss the need for XAI from various perspectives. Following that, Section 3 tries to contribute to a better distinction between explainability and interpretability. The protocol used in planning and executing this systematic meta-survey is presented in Section 4. Afterward, Section 5 discusses the challenges and research directions of XAI. Section 6 shows how some of the discussed challenges and research directions can be considered in medicine (which can also be tailored to any other domains). Lastly, final remarks are highlighted in Section 7

2. Why explainable AI is needed?

Nowadays, we are surrounded by black-box AI systems utilized to make decisions for us, as in autonomous vehicles, social

* Corresponding author at: School of Computer Science and Informatics, De Montfort University, Leicester, LE1 9BH, UK.
E-mail addresses: waddah.saeed@dmu.ac.uk (W. Saeed), christian.omlin@uia.no (C. Omlin).

https://doi.org/10.1016/j.knosys.2023.110273

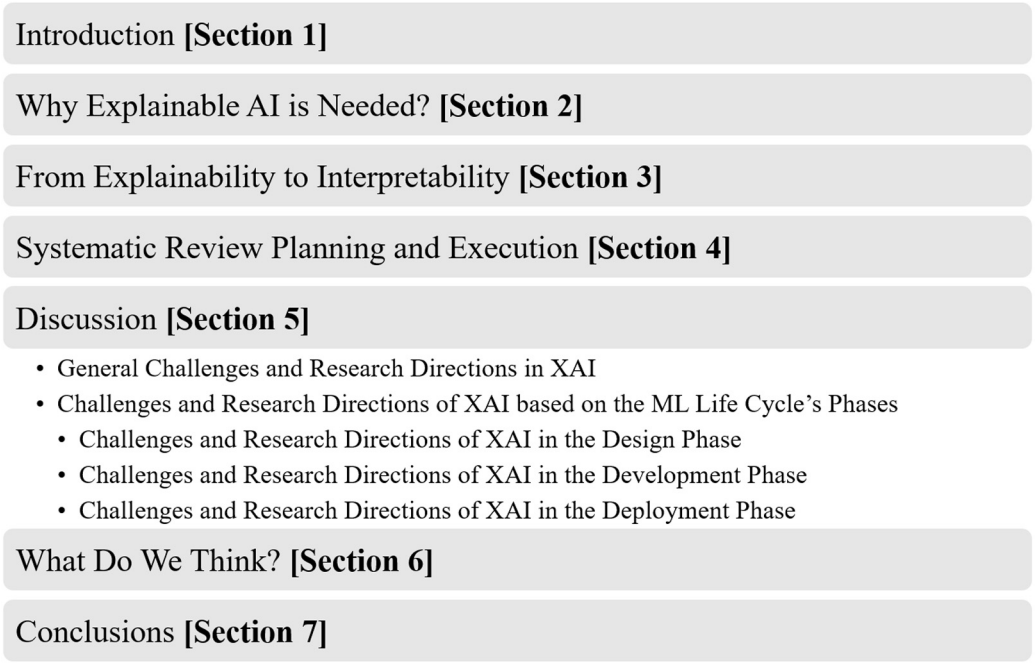


Fig. 1. The organization of this meta-survey paper.

networks, and medical systems. Most of these decisions are taken without knowing the reasons behind these decisions.

According to [1], not all black-box AI systems need to explain why they take each decision because this could result in many consequences such as reducing systems efficiency and increasing development costs. Generally, explainability/interpretability is not needed in two situations [5]: (1) results that are unacceptable are not accompanied by severe consequences, (2) the problem has been studied in-depth and well-tested in practice, so the decision made by the black-box system is trustworthy, e.g., advertisement system and postal code sorting. Therefore, we should think about why and when explanations/interpretations can be helpful [1].

Based on the retrieved surveys in this work, the need for XAI can be discussed from various perspectives as shown in Fig. 2. The perspective groups below are to some extent based on the work in [6]:

- Regulatory perspective: Black-box AI systems are being utilized in many areas of our daily lives, which could be resulting in unacceptable decisions, especially those that may lead to legal effects. Thus, it poses a new challenge for the legislation. The European Union's General Data Protection Regulation (GDPR)[1] is an example of why XAI is needed from a regulatory perspective. These regulations create what is called the "right to explanation," by which a user is entitled to request an explanation about the decision made by the algorithm that considerably influences them [7]. For example, if an AI system rejects one's application for a loan, the applicant is entitled to request justifications behind that decision to guarantee it is in agreement with other laws and regulations [8]. However, the implementation of such regulations is not straightforward, challenging, and without an enabling technology that can provide explanations, the "right to explanation" is nothing more than a "dead letter" [8–10].
- Scientific perspective: When building black-box AI models, we aim to develop an approximate function to address the given problem. Therefore, after creating the black-box AI model, the created model represents the basis of knowledge, rather than the data [11]. Based on that, XAI can be helpful to reveal the scientific knowledge extracted by the black-box AI models, which could lead to discovering novel concepts in various branches of science.
- Industrial perspective: Regulations and user distrust in black-box AI systems represent challenges to the industry in applying complex and accurate black-box AI systems [12]. Less accurate models that are more interpretable may be preferred in the industry because of regulation reasons [12]. A major advantage of XAI is that it can help in mitigating the common trade-off between model interpretability and performance [2], thus meeting these common challenges. However, it can increase development and deployment costs.
- Model's developmental perspective: Several reasons could contribute to inappropriate results for black-box AI systems, such as limited training data, biased training data, outliers, adversarial data, and model overfitting. Therefore, what black-box AI systems have learned and why they make decisions need to be understood, primarily when they affect humans' lives. For that, the aim will be to use XAI to understand, debug, and improve the black-box AI system to enhance its robustness, increase safety and user trust, and minimize or prevent faulty behavior, bias, unfairness, and discrimination [13]. Furthermore, when comparing models with similar performance, XAI can help in the selection by revealing the features that the models used to produce their decisions [14,15]. In addition, XAI can serve as a proxy function for the ultimate goal because the algorithm may be optimized for an incomplete objective [5]. For instance, optimizing an AI system for cholesterol control with ignoring the likelihood of adherence [5].
- End-user and social perspectives: In the literature of deep learning [10,16,17], it has been shown that altering an image such that humans cannot observe the change can lead the

[1] https://www.privacy-regulation.eu/en/r71.htm

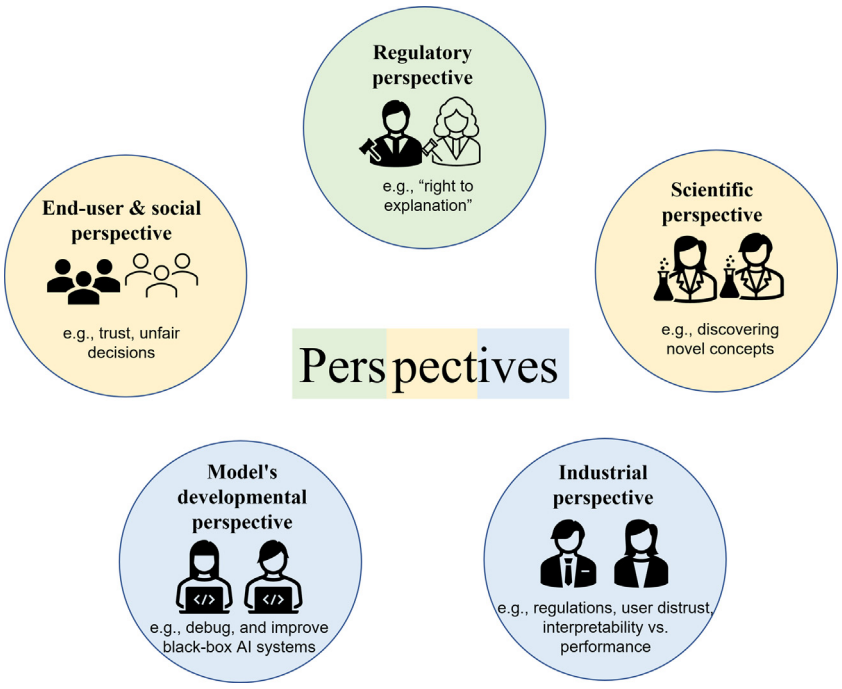


Fig. 2. The five main perspectives for the need for XAI.

model in producing a wrong class label. On the contrary, completely unrecognizable images to humans can be recognizable with high confidence using DL models [17]. Such findings could raise doubts about trusting such black-box AI models [10]. The possibility to produce unfair decisions is another concern about black-box AI systems. This could happen in case black-box AI systems are developed using data that may exhibit human biases and prejudices [10]. Therefore, producing explanations and enhancing the interpretability of the black-box AI systems will help in increasing trust because it will be possible to understand the rationale behind the model's decisions, and we can know if the system serves what it is designed for instead of what it was trained for [10,18]. Furthermore, the demand for the fairness of black-box AI systems' decisions, which cannot be ensured by error measures, often leads to the need for interpretable models [19].

The above list is far from complete, and there may be an overlap between these perspectives. However, we believe that these perspectives highlight the most critical reasons why XAI is needed.

3. From explainability to interpretability

In the literature, there seems to be no agreement on what "explainability" or "interpretability" mean. While both terms are often used interchangeably in the literature, some examples from the selected papers distinguish them [2,20–24]. To show that there is no agreement upon definitions, three different definitions from [2,20,21] are provided. In [20], the authors stated that "*... we consider interpretability a property related to an explanation and explainability a broader concept referring to all actions to explain.*". In another work [2], interpretability is defined as "*the ability to explain or to provide the meaning in understandable terms to a human.*", while "*explainability is associated with the notion of explanation as an interface between humans and a decision-maker that is, at the same time, both an accurate proxy of the decision-maker and comprehensible to humans...*". Another distinction is drawn in [21], in which authors stated that "*... In the case of interpretation, abstract concepts are translated into insights useful for domain knowledge (for example, identifying correlations between layers in a neural network for language analysis and linguistic knowledge). An explanation provides information that gives insights to users as to how a model came to a decision or interpretation.*". It can be noticed from these distinctions that the authors have different definitions for these two terms. In addition, there is still considerable ambiguity in some of the given distinctions.

To contribute to a distinction between explainability and interpretability, this paper attempts to present a distinction between these terms as follows:

> Explainability provides **insights** to a **targeted audience** to fulfill a **need**, whereas interpretability is the degree to which the provided **insights** can make sense for the **targeted audience**'s domain knowledge.

There are three components in the definition of explainability, as shown in the above distinction: **insights**, **targeted audience**, and **need**. **Insights** are the output from explainability techniques used (e.g., text explanation, feature relevance, local explanation). These insights are provided to a **targeted audience** such as domain experts (e.g., medical doctors), end-users (e.g., users affected by the model decision), and modeling experts (e.g., data scientists). The **need** for the provided insights may be to handle any issues discussed in Section 2 such as justifying decisions, discovering new knowledge, improving the black-box AI model, and ensuring fair decisions. That means explainability aims to help the targeted audience to fulfill a need based on the provided insights from the explainability techniques used.

As for interpretability, are the provided explanations consistent with the targeted audience's knowledge? Do the explanations make sense to the targeted audience? Is the targeted audience able to reason/infer to support decision-making? Are the provided explanations reasonable for the model's decision?

Although the distinction is not ideal, we believe that it represents an initial step towards understanding the difference between explainability and interpretability. Because the interpretability definition cannot be generalized [25,26], it is crucial to take

Table 1
The selected databases and a search engine.

Database/search engine	Link
Scopus	https://www.scopus.com/
Web of Science	https://www.webofscience.com
Science Direct	https://www.sciencedirect.com/
IEEEXplore	https://ieeexplore.ieee.org/Xplore/home.jsp
Springer Link	https://link.springer.com/
Association for Computing Machinery Digital Library (ACM)	https://dl.acm.org/
Google Scholar (Search engine)	https://scholar.google.com/
arXiv	https://arxiv.org/

into account the problem domain and user type [26,27] when measuring interpretability properties. Establishing formalized rigorous evaluation metrics is one of the challenges in XAI, which has been discussed in Towards more formalism section. This paper will use this proposed distinction when discussing the challenges and research directions in XAI.

4. Systematic review planning and execution

This work is mainly based on a systematic literature review (SLR) introduced by Kitchenham and Charters [28]. SLR is used to identify all relevant papers that can help in answering a specified research question in an unbiased manner. SLR has been used in many published works, for example, the works in [29–31]. We started our SLR by specifying the following research question: *What are the challenges and research directions in XAI reported in the existing survey studies?* The answer to this question will help researchers and practitioners to know the various dimensions that one can consider when working in the XAI research area.

Having the research question established, the search terms based on the research question are:

- XAI keywords: explainable, XAI, interpretable.
- Review keywords: survey, review, overview, literature, bibliometric, challenge, prospect, agenda, trend, insight, opportunity, lesson, research direction

With these selected search terms, Boolean ANDs were used between groups and ORs within groups to construct search strings as follows: (explainable OR XAI OR interpretable) AND (survey OR review OR overview OR literature OR bibliometric OR challenge OR prospect OR agenda OR trend OR insight OR opportunity OR lesson OR "research direction").

Relevant and important electronic databases and a search engine were used for searching the primary studies based on the search terms. These databases and the search engine are shown in Table 1. The search using these databases and the search engine was based on different search schemes as shown in Table 2. That was adapted depending on the needs of these databases and the search engine. We retrieved papers published before 1 September 2021.

Inclusion and exclusion criteria were used to select or discard the retrieved studies. The inclusion criteria are the following:

- The study presents a survey of explainable AI.
- The study presents challenges and/or research directions for XAI.

On the other hand, the exclusion criteria are the following:

- The study is not written in English.
- The study presents a survey of XAI without discussing any challenges or research directions.

After obtaining search results, all studies were analyzed individually by the first author to assess their relevance in the context of this SLR considering the inclusion and exclusion criteria. These studies were first analyzed by their titles and abstracts to decide if the paper matched the first inclusion criterion. If matched, the paper was analyzed in detail in the second step. In the second step, the exclusion criteria and the second inclusion criterion were checked. Throughout the work on this SLR, the authors met regularly to discuss what has been done and what needs to be done next.

Table 2
Search schemes for the databases and search engine used along with the number of retrieved papers.

Database/search engine	Search scheme	Retrieved
Scopus	Title	148
Web of Science	Title	100
Science Direct	Title	N/A[a]
IEEEXplore	Title	34
Springer Link	Title	N/A[a]
ACM	Title	93
Google Scholar	Full text	200[b]
arXiv	Title	N/A[a]

[a]It is not possible to use all keywords in one search string. Therefore, the count of retrieved papers is not correct as some papers appeared more than one time.
[b]The first 20 pages.

Table 3
Distribution of selected papers per year.

Year	Number of papers
2017	4
2018	12
2019	10
2020	22
2021	25[a]
Total	73

[a]Published before 1 September 2021.

We reviewed the list of references of the selected studies to include other papers that may not be retrieved from the selected electronic databases, which resulted in retrieving eight non-survey papers that reported challenges and/or research directions in XAI [5,11,19,32–36]. It is good to note that each arXiv paper was only included if it did not have a peer-reviewed version, otherwise, the peer-reviewed version was included.

Overall, the total number of selected papers is 73, as shown in Table 3. As shown in Fig. 3, the primary outlet for the selected papers is journal articles followed by conference papers and arXiv papers. The distribution of the selected papers per publisher is shown in Fig. 4.

5. Discussion

To our best knowledge, there are two meta-survey papers on XAI that primarily used survey papers as a basis for their discussions. The first meta-survey focused its discussion on the visual interpretation of ML models using 15 survey papers and 3 non-survey papers published between 2014–2018 (17 between 2016–2018 and 1 in 2014) [24]. The second meta-survey paper [37] included over 70 survey papers published up to the

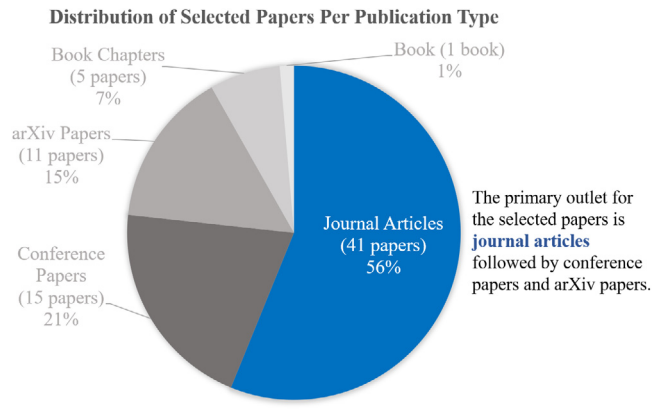


Fig. 3. Distribution of selected papers per publication type.

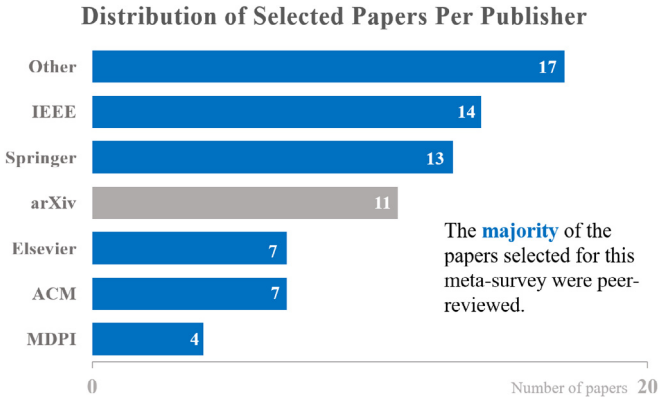


Fig. 4. Distribution of selected papers per publisher.

beginning of 2021 with a focus on XAI and XAI methods. In our survey, we included 73 papers that covered a range of general and specific aspects of XAI with a focus on challenges and research directions in XAI reported in these included papers.

There are various taxonomies proposed for XAI which mainly depend on the aspects discussed in the papers. For example, the authors of [37] proposed a taxonomy that has three main components: explainability problem definition, explanator properties, and metrics. In [2,38], the taxonomy is based on the explainability/interpretability methods (i.e., ad-hoc and post-hoc methods). A three dimensions taxonomy based on interpretability intervention (passive vs. active) methods, type of explanations, and input space (global vs. local) was used in [23], and a taxonomy based on explanation forms (visual, text, graph, and symbol explanations) was used in [39].

In order to place the challenges and research directions in XAI in a meaningful context, the discussion in our meta-survey is based on a taxonomy based on two themes, as shown in Fig. 5. The first theme focuses on the general challenges and research directions in XAI and the second theme is about the challenges and research directions of XAI based on the ML life cycle's phases. Before stating the reason for using ML life cycle's phases in our taxonomy, it is important to highlight that ML models are data-driven models so they learn through experience (i.e., data). If training data contain biases and/or specific design choices can result in biased behavior in ML algorithms, ML models can produce biased outcomes and these biased outcomes can affect users' decisions and these decisions can result in more biased data that could be used to train other ML algorithms [40]. That means bias can occur along all phases of the ML pipeline. Since XAI is meant to detect and prevent or at least mitigate bias, we categorized the challenges and research directions in XAI based on ML life cycle's phases.

For simplicity, we divided the life cycle into three main phases: design, development, and deployment phases. It is good to note that the reported challenges and research directions were sorted from the most commonly reported in the selected papers to the least. The following subsections shed light on these challenges and research directions.

5.1. General challenges and research directions in XAI

In this section, we reported XAI's general challenges and research directions. These general challenges and research directions in XAI are shown in Fig. 6. The selected papers that discussed these general challenges and research directions are mentioned in Table 4. As shown in Fig. 7, the majority of these challenges and research directions have received much attention recently (dark blue colors in 2020 and 2021).

5.1.1. Towards more formalism

It is one of the most raised challenges in the literature of XAI [1,2,8–10,18,26,32,41–57,59–61]. It was suggested that more formalism should be considered in terms of systematic definitions, abstraction, and formalizing and quantifying [1].

Starting from the need for systematic definitions, until now, there is no agreement on what an explanation is [10]. Furthermore, it has been found that similar or identical concepts are called by different names and different concepts are called by the same names [1,41]. In addition, without a satisfying definition of interpretability, how it is possible to determine if a new approach better explains ML models [42]? Therefore, to facilitate easier sharing of results and information, definitions must be agreed upon [1,41].

With regards to abstraction, many works have been proposed in an isolated way; thus, there is a need to be consolidated to build generic explainable frameworks that would guide the development of end-to-end explainable approaches [1]. Additionally, taking advantage of the abstraction explanations in identifying properties and generating hypotheses about data-generating processes (e.g., causal relationships) could be essential for future artificial general intelligence (AGI) systems [41].

Regarding the formalization and quantification of explanations, it was highlighted in [1] that some current works focus on a detailed problem formulation which becomes irrelevant as the method of interpretation or the explanation differs. Therefore, regardless of components that may differ, the expansibility problem must be generalized and formulated rigorously, and this will improve the state-of-the-art for identifying, classifying, and evaluating sub-issues of explainability [1]. The work in [53] also stressed the need for a thorough formalization and theoretical understanding of XAI to answer important and unanswered theoretical questions such as the weighing of model and data distribution into the generated explanation.

Establishing formalized rigorous evaluation metrics needs to be considered as well [1]. However, due to the absence of an agreement on the definitions of interpretability/explainability, no established approach exists to evaluate XAI results [9]. The lack of ground truth explanations in most cases is the biggest challenge for rigorous evaluations [42,50], and because of that and other factors such as the cost of predictability or run-time efficiency, the question of what is the optimal explanation remains an open question [53]. So far, different evaluation metrics have been proposed, such as reliability, trustworthiness, usefulness, soundness, completeness, compactness, comprehensibility, human-friendly or human-centered, correctness or fidelity, complexity, and generalizability [50]. However, it seems

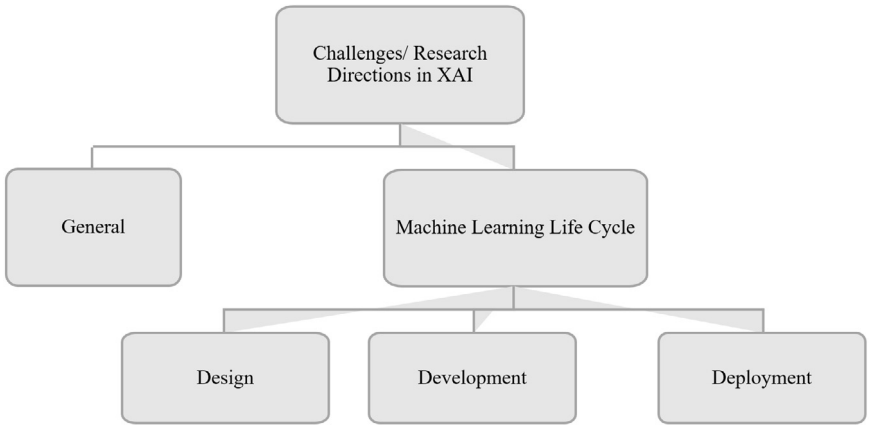


Fig. 5. The proposed organization to discuss the challenges and research directions in XAI. For simplicity, the arrows that show the flow in the life cycle are removed.

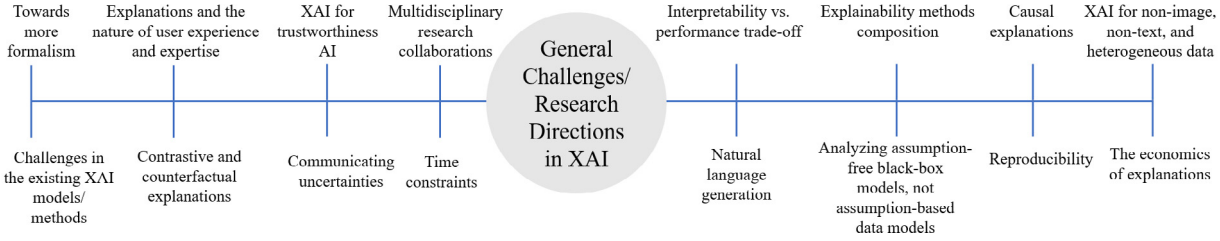


Fig. 6. General Challenges and Research Directions in XAI.

Table 4
A summary of the selected papers for the general challenges and research directions in XAI.

Challenges and research directions	Papers
Towards more formalism	[1,2,8–10,18,26,27,32,39,41–58], [59–61][a]
Explanations and the nature of user experience and expertise	[2,24,32,34,48,50,55,62–64], [59,60][a]
XAI for trustworthiness AI	[2,20,22,24,36,45,50,58,65], [61,66][a]
Multidisciplinary research collaborations	[18,38,44,55,67–69], [59,61,70][a]
Interpretability vs. performance trade-off	[2,32,34,44,53,58,71]
XAI for non-image, non-text, and heterogeneous data	[43,63,68,72], [59,73][a]
Explainability methods composition	[1,32,36,65], [59][a]
Causal explanations	[32,42,45,65,74]
Challenges in the existing XAI models/methods	[8,32,42,54,75][a]
Contrastive and counterfactual explanations	[39,76], [60][a]
Communicating uncertainties	[24,42]
Time constraints	[50], [5][a]
Natural language generation	[59][a]
Analyzing assumption-free black-box models, not assumption-based data models	[11]
Reproducibility	[9]
The economics of explanations	[1]

[a]A non-peer-reviewed paper from arXiv.

that there are two main evaluation metrics groups: objective and human-centered evaluations [42]. The former is quantifiable mathematical metrics, and the latter relies on user studies [42]. Further progress is needed towards evaluating XAI techniques' performance and establishing objective metrics for evaluating XAI approaches in different contexts, models, and applications [2]. Recently, the work in [27] suggested that further research should be undertaken in objective and human-centered evaluations. For human-centered evaluations, looking into effective designs for human experiments and subjective explanation evaluation measures can help establish agreed criteria on human-centered evaluations thus making comparisons between explanations easier. As for objective evaluations, it seems that previous work has mainly focused on attribution-based explanations (e.g., feature importance). Therefore, considering other types of explanations is needed (e.g., example-based explanations). Additionally, there is a need to come up with more objective evaluations that measure explainability properties of clarity and broadness of interpretability as the current focus is on the evaluation of the soundness of fidelity of explanations. Finally, it has been suggested by the authors the need to integrate both human-centered and objective evaluations for a comprehensive evaluation as well as understand the contribution each metric makes in this comprehensive evaluation. Following the evaluation of the explanations, the recommended explanation is shown based on the task and the type of user, which is ultimately best to build using a model-agnostic framework [26].

5.1.2. Explanations and the nature of user experience and expertise

Based on the nature of the application, users who use ML models can vary (e.g., data scientists, domain experts, decision-makers, and non-experts). The nature of user experience and expertise matters in terms of what kind of cognitive chunks they possess and the complexity they expect in their explanations [5].

Recent research published in 2020 and 2021 has focused on most of these challenges and research directions

Challenges and research directions	2017	2018	2019	2020	2021	Total
Towards more formalism	1	4	4	11	11	31
Explanations and the nature of user experience and expertise	1	1	2	6	2	12
XAI for trustworthiness AI	1	1	0	4	5	11
Multidisciplinary research collaborations	0	1	1	3	5	10
Interpretability vs. performance trade-off	0	1	1	3	2	7
XAI for non-image, non-text, and heterogeneous data	0	0	0	3	3	6
Explainability methods composition	0	3	0	1	1	5
Causal explanations	0	1	0	3	1	5
Challenges in the existing XAI models/methods	0	1	1	1	2	5
Contrastive and counterfactual explanations	0	0	1	0	2	3
Communicating uncertainties	0	0	0	2	0	2
Time constraints	1	0	0	1	0	2
Natural language generation	0	0	0	1	0	1
Analyzing assumption-free black-box models, not assumption-based data models	0	0	1	0	0	1
Reproducibility	0	0	0	1	0	1
The economics of explanations	0	1	0	0	0	1

Fig. 7. Number of research papers per year for each general challenge and research direction. Method: Publication year was used for peer-reviewed papers published before 2022, otherwise, arXiv's publication year was considered. As can be seen from this figure, the majority of these challenges and research directions have received much attention recently (dark blue colors in 2020 and 2021).

In general, users have varying backgrounds, knowledge, and communication styles [5]. However, it seems that the current focus of explanation methods is tailored to users who can interpret the explanations based on their knowledge of the ML process [36,50].

The works in [24,32,34,48,50,60,62] have highlighted what is needed to be considered with regards to explanations and the nature of user experience and expertise. In [50], user-friendly explanations have been suggested so users can interpret the explanations with less technical knowledge. Therefore, figuring out what to explain should follow the identification of the end-user. In [60], it has been highlighted that previous works in explainable AI systems (e.g., expert systems) generally neglected to take into account the knowledge, goals, skills, and abilities of users. Additionally, the goals of users, systems, and explanations were not clearly defined. Therefore, clearly stating goals and purposes are needed to foster explanation testing within the appropriate context. In [62], the authors have discussed that identifying the users' goals and keeping up with their dynamic nature means collecting more data from them. It is also essential to develop change detection approaches to goals and needs for the purpose of adapting these changes to end-users. For a deeper understanding of these dynamics, user studies (e.g., diary studies, interviews, and observation) can help develop guidelines for developing long-term explainable systems and determining which user data to gather to improve personalization [62].

In [34], it has been suggested that abstraction can be used to simplify the explanations. Understanding how abstractions are discovered and shared in learning and explanation is an essential part of the current XAI research. The work in [32] has mentioned that the inclusion of end-users in the design of black-box AI models is essential, especially for specific domains, e.g., the medical domain. That would help to understand better how the end-users will use the outputs and interpret explanations. It is a way to educate them about the predictions and explanations produced by the system. In [24], the authors have discussed that utilizing users' previous knowledge is a significant challenge for visualization tools today. Customizing visualization tools for different user types can be useful at several stages of the ML model pipeline [24]. However, to use prior users' knowledge in predictive models, it is important to establish processes to digitally capture and quantify their prior knowledge [24,77].

In [59], it has been mentioned that DL models often use concepts that are unintelligible to predict outcomes. Therefore, using systems that use such models requires human-centric explanations that can accurately explain a decision and make sense to the users (e.g., medical domain expert) [59]. An approach to come up with human-centric explanations is examining the role of human-understandable concepts acquired by DL models [59]. It is also essential to analyze the features used by the DL models in predicting correct decisions based on incorrect reasoning [59]. Having an understanding of the model's concepts would help reduce reliability concerns and develop trust when deploying the system, especially in critical applications [59]. The authors also highlighted the importance of addressing the domain-specific needs of specific applications and their users when developing XAI methods. Finally, the work in [2] has discussed that XAI can facilitate the process of explaining to non-experts how a model reached a given decision, which can substantially increase information exchange among heterogeneous people regarding the knowledge learned by models, especially when working in projects with a multi-disciplinary team.

To sum up, it is crucial to tailor explanations based on user experience and expertise. Explanations should be provided differently to different users in different contexts [78]. In addition, it is also essential to clearly define the goals of users, systems, and explanations. Stakeholder engagement and system design are both required to understand which explanation type is needed [78].

5.1.3. XAI for trustworthiness AI

Increasing the use of AI in everyday life applications will increase the need for AI trustworthiness, especially in situations where undesirable decisions may have severe consequences [50]. The High-Level Expert Group in European Commission put seven essentials for achieving trustworthy AI[2]: (1)*human agency and oversight*; (2) *robustness and safety*; (3) *privacy and data governance*; (4) *transparency*; (5) *diversity, non-discrimination, and fairness*; (6) *societal and environmental well-being*; and (7) *accountability*. The discussion about privacy, security, and safety is given in XAI and Privacy Section, XAI and Security Section, and XAI and Safety Section, respectively. The discussion in this section is about what is reported in the selected papers regarding fairness and accountability.

[2] https://ec.europa.eu/commission/presscorner/detail/en/IP_19_1893

With regard to fairness, ML algorithms must not be biased or discriminatory in the decisions they provide. However, with the increased usage of ML techniques, new ethical, policy, and legal challenges have also emerged, for example, the risk of unintentionally encoding bias into ML decisions [22]. Meanwhile, the opaque nature of data mining processes and the complexity of ML models make it more challenging to justify consequential decisions of following what the models say [22]. The work in [61] argues that data, algorithmic, and social biases need to be remedied in order to promote fairness. Further, it is imperative to be able to analyze AI systems to have trust in the model and its predictions, especially for some critical applications. Researchers started trying to form a definition of fairness and the meaning of fairness in an algorithm as discussed in [22]. According to [22], it would also be necessary to devise new techniques for discrimination-aware data mining. It is also worth noting that when converting fairness into a computational problem, we need to keep the fairness measures fair [22]. The work in [36] states that it is possible to visualize learned features using XAI methods and assess bias using methods other than explanation methods. On the other hand, regulations and laws are necessary for the suspicion about unfair outcomes [36].

Turning now to accountability, having accountability means having someone responsible for the results of AI decisions if harm occurs. In [78], it has been mentioned that investigating and appealing decisions with major consequences for people is an important aspect of systems of accountability, and some current regulations also aim to achieve this. XAI can be an important factor in systems of accountability by providing users with the means to appeal a decision or modify their behavior in the future to achieve a better result. However, more work should be done to establish an environment that promotes individual autonomy and establish a system of accountability. It has also been discussed in [22] that developing procedures for testing AI algorithms for policy compliance is necessary so that we can establish whether or not a given algorithm adheres to a specific policy without revealing its proprietary information. It is also desirable for a model to specify its purposes and provide external verification of whether these goals are met and, if not, describe the causes of the predicted outcomes.

The use of XAI can enhance understanding, increase trust, and uncover potential risks [61]. Therefore, when designing XAI techniques, it is imperative to maintain fairness, accountability, and transparency [61]. On the other hand, it is necessary to highlight that not only black-box AI models are vulnerable to adversarial attacks, but also XAI approaches [79]. There is also a risk that to promote trust in black-box AI models predictions; explainers may be more persuasive but misleading than informative, so users may become deceived, thinking the system to be trustworthy [45,78]. It is possible to increase trust through explanations, but explanations do not always produce systems that produce trustworthy outputs or ensure that system implementers make trustworthy claims about their abilities [78].

The work in [20] discusses measures to create trustworthy AI. It has been highlighted that before employing AI systems in practice, it is essential to have quantitative proxy metrics to assess explanation quality objectively, compare explanation methods, and complement them with human evaluation methods (e.g., data quality reporting, extensive testing, and regulation).

Finally, it is good to note that a further explore the idea of Responsible AI with a discussion about principles of AI, fairness, privacy, and data fusion can be found in [2].

5.1.4. Multidisciplinary research collaborations

One area of research that can offer new insights for explainable methods is working closely with researchers from other disciplines such as psychology, behavioral and social sciences, human–computer interaction, physics, and neuroscience. Multidisciplinary research is therefore imperative to promote human-centric AI and expand utilizing XAI in critical applications [61]. For example, in healthcare [80], military [81], common law [82], and transportation [66].

Several studies, for instance [18,38,44,59,67–69], have suggested some potential multidisciplinary research works. In [44], it has been highlighted that approaching the psychology discipline can help to get insights into both the structure and the attributes of explanations and the way they can influence humans. They also have suggested that defining the context of explanations is an important research direction. Here, it is essential to consider the domain of the application, the users, the type of explanations (e.g., textual, visual, combinations of solutions), and how to provide the explanations to the users. This research direction can form a connection with behavioral and social sciences. The paper in [67] also has shown that XAI can benefit from the work in philosophy, cognitive psychology/science, and social psychology. The paper summarizes some findings and suggests ways to incorporate these findings into work on XAI.

Approaching Human–Computer Interaction (HCI) studies are essential to XAI. However, few user experiments have been conducted in the area of explainability [18]. Therefore, more should be conducted to study the topic adequately [18]. Humans must be included in the process of creating and utilizing XAI models, as well as enhancing their interpretability/explainability [44]. In [59], it has been highlighted that interactive tools may help users understand, test, and engage with AI algorithms, thereby developing new approaches that can improve algorithms' explainability. Furthermore, interactive techniques can help users to interpret predictions and hypothesis-test users' intuitions rather than relying solely upon algorithms to explain things to them. In [69], it has been suggested to draw from the HCI research on interaction design and software learnability to improve the usability of intelligible or explainable interfaces. Additionally, HCI researchers can take advantage of the theoretical work on the cognitive psychology of explanations to make understandable explanations. They can also empirically evaluate the effectiveness of new explanation interfaces.

The advances in neuroscience should be of great benefit to the development and interpretation of DL techniques (e.g., cost function, optimization algorithm, and bio-plausible architectural design) owing to the close relationship between biological and neural networks [38]. It is imperative to learn from biological neural networks so that better and explainable neural network architectures can be designed [38]. Finally, connecting with physics and other disciplines that have a history of explainable visual methods might provide new insights for explainable methods [68].

5.1.5. Interpretability vs. performance trade-off

The belief that complicated models provide more accurate outcomes is not necessarily correct [25]. However, this can be incorrect in cases when the given data is structured and with meaningful features [25]. In a situation where the function being approximated is complex, that the given data is widely distributed among suitable values for each variable and the given data is sufficient to generate a complex model, the statement *"models that are more complex are more accurate"* can be true [2]. In such a situation, the trade-off between interpretability and performance becomes apparent [2].

When the performance is coupled with model complexity, model interpretability is in question [2]. Explainability techniques, however, could help in minimizing the trade-off [2]. However, according to [32], what determines this trade-off? and who determines it? The authors have highlighted the importance of discussing with end-users this trade-off so that they can be aware of the potential risks of misclassification or opacity. Another point that should be considered is the approximation dilemma: models need to be explained in enough detail and in a way that matches the audience for whom they are intended while keeping in mind that explanations reflect the model and do not oversimplify its essential features [2]. Even though studying the trade-off is essential, it is impossible to proceed without standardized metrics for assessing the quality of explanations [71].

Another possible solution for the trade-off is suggested in [44] which is developing fully transparent models throughout the entire process of creation, exploitation, and exploration and can provide local and global explanations. In turn, this leads to using methods that embed learning capabilities to develop accurate models and representations [44]. The methods should also be able to describe these representations in effective natural language consistent with human understanding and reasoning [44].

5.1.6. Explainability methods composition

For specific applications in healthcare (e.g., predicting disease progression), several types of explanations at different levels are needed (e.g., local and global explanations) [32] in order to provide the most complete and diverse explanations we can [59]. This is derived from the way clinicians communicate decisions utilizing visualizations and temporal coherence as well as textual descriptions [59].

Some overlap exists between explainability methods, but for the most part, each seems to address a different question [65]. According to [1], combining various methods to obtain more powerful explanations is rarely considered. In addition, rather than using disparate methods separately, we should investigate how we can use them as basic components that can be linked and synergized to develop innovative technologies [1]. Furthermore, it could help to provide answers in a simple human interpretable language [36]. First efforts, as cited in [59], have been made as in [83] where the authors proposed a model that can provide visual relevance and textual explanations. Recently, the work in [84] proposed a new framework with multi-modal explanations component derived based on a wide range of images and vocabulary. In this model, the generated textual explanations are paired with their corresponding visual regions in the image. As a result of this explanation generation method and other components in the framework, better reasoning and interpretability were achieved than in some state-of-the-art models. These findings suggest opportunities for future research with the aim to enhance both interpretability and accuracy [1].

5.1.7. Causal explanations

Developing causal explanations for AI algorithms (i.e., why they made those predictions instead of how they arrived at those predictions) can help increase human understanding [45]. In addition, causal explanations strengthen models' resistance to adversarial attacks, and they gain more value when they become part of decision-making [42]. However, there can be conflicts between predicting performance and causality [42]. For example, when the confounder, which is a variable that influences both the dependent variable and independent variable, is missing from the model [42].

Causal explanations are anticipated to be the next frontier of ML research and to become an essential part of the XAI literature [32,65]. There is a need for further research to determine when causal explanations can be made from an ML model [42]. In addition, according to a recent survey on causal interpretability for ML [74], it has been highlighted the absence of ground truth data for causal explanations and verification of causal relationships makes evaluating causal interpretability more challenging. Therefore, more research is needed to guide on how to evaluate causal interpretability models [74].

5.1.8. XAI for non-image, non-text, and heterogeneous data

The focus of XAI works is mainly on image and text data (e.g., Natural language generation and Interpretability for natural language processing sections for text explanations and Developing visual analytics approaches for advanced DL architectures for visual explanations). Other data types exist, however, but they have received less attention [63,73], such as time-series [85], graphs [86], and spatio-temporal data [87].

Using visualization to transform non-image data into images creates opportunities to discover explanations through salient pixels and features [68]. However, this should not be the only way for explainability for non-image or non-text data for different reasons. For example, existing explanation approaches for image or text data need to be adjusted to be used with graph data [72], and the outcomes which are clearly interpretable from explanation approaches for images (e.g., saliency maps) might need expert knowledge to be understood for time series data [73].

Additionally, there is a need to develop new approaches for explaining the information that exists with non-image or non-text data, e.g., structural information for graph data [72] and multivariate time series data of variable length [63].

Finally, with the advent of AI systems that use various types of data, explainability approaches that can handle such heterogeneity of information are more promising [43]. For example, such systems can simulate clinicians' diagnostic processes in the medical domain where both images and physical parameters are utilized to make decisions [59]. Thus, they can enhance the diagnostic effectiveness of the systems as well as explain phenomena more thoroughly [59].

5.1.9. Challenges in the existing XAI models/methods

There are some challenges in the existing XAI models/methods that have been discussed in the literature. Starting with scalability, which is a challenge that exists in explainable models as discussed in [32]. For example, each case requiring an explanation entails creating a local model using LIME explainable model [88]. Scalability can be an issue when there is a huge number of cases for which prediction and explanation are needed. Likewise, when computing Shapley values [89], all combinations of variables must be considered when computing variable contributions. Therefore, such computations can be costly for problems that have lots of variables.

Feature dependence presents problems in attribution and extrapolation [42]. If features are correlated, attribution of importance and features effects becomes challenging. For sensitivity analyses that permute features, when the permuted feature has some dependence on another feature, the association breaks, resulting in data points outside the distribution, which could cause misleading explanations. In [8], the authors discussed some limitations with heatmaps explanations. Heatmaps explanations visualize what features are relevant for making predictions. There is, however, a lack of clarity regarding their relationship (e.g., their importance for the predictions either individually or in combination). Low abstraction levels of explanations are another limitation. Heatmaps highlight that specific pixels are significant without indicating how the relevance values relate to abstract concepts in the image, such as objects or scenes. The model's behavior can be explained in more abstract, more easily understood ways by meta-explanations that combine evidence from

low-level heatmaps. Therefore, further research is needed on meta-explanations.

Model-based (i.e., ante-hoc models) and post-hoc explainability models have some challenges, as have been discussed in [90]. When model-based methods cannot predict with reasonable accuracy, practitioners start the search for more accurate models. Therefore, one way to increase the usage of model-based methods is to develop new modeling methods that maintain the model's interpretability and render more accurate predictions [90]. The availability of such methods is needed and useful especially when implemented in clinical applications to benefit from the predictive power of advanced DL models and interpretability [54]. More details about this direction are provided in [25]. Further, according to [90], for model-based methods, there is a need to develop more tools for feature engineering. Simpler but accurate model-based methods can be built when the input features used with these methods are more informative and meaningful. Two categories of works can help achieve that: improve tools for exploratory data analysis and improve unsupervised techniques. The former helps to understand the data, and domain knowledge could help to identify helpful features. The latter is needed because unsupervised techniques are often used to identify relevant structures automatically, so advances in unsupervised techniques may result in better features.

The authors in [90] have also discussed some challenges for post-hoc explainability models. According to the authors, it is challenging to determine what format or combination of formats will adequately describe the model's behavior. Furthermore, there is uncertainty over whether the current explanation methods are adequate to capture a model's behavior or if novel methods are still needed. Another challenge is if post-hoc explanations methods identify learned relationships by the model that practitioners know to be incorrect, is it possible that practitioners fix these relationships learned and increase the predictive accuracy? Further research in post-hoc explanations can help exploit prior knowledge to improve the predictive accuracy of the models.

Finally, some research directions have been suggested in [75] to deal with some challenges associated with perturbation-based methods. For example, the need to find an optimal scope of perturbations of the inputs as sampling all perturbations is not possible (i.e., combinatorial complexity explosion problem). The development of cross-domain applications of perturbations methods would also benefit from empirical studies comparing perturbations to different data types [75].

5.1.10. Contrastive and counterfactual explanations

Contrastive explanations describe why one event occurred but not another, while counterfactual explanations describe what is needed to produce a contrastive output with minimal changes in the input [76]. Questions in the contrastive form "Why x and not y?" and questions of the counterfactual form "What if?" and "What would happen if?" [60].

In a recent survey of contrastive and counterfactual explanations [76], it has been found that contrastive and counterfactual explanations help improve the interaction between humans and machines and personalize the explanation of algorithms. A further important point as observed by [76] that one of the significant barriers towards a fair assessment of new frameworks is the lack of standardization of evaluation methods. The theoretical frameworks are also found inadequate for applying to XAI as a result of the disconnect between the philosophical accounts of counterfactual explanation to scientific modeling as well as ML-related concepts. Furthermore, it has been found that different domains of science define counterfactual explanations differently, as do the approaches used to solve specific tasks.

In the light of possible research directions on this point, it has been suggested in [76] the importance of including end-users in the evaluation of generated explanations since these explanations are designed to be user-oriented. In addition, since contrastive and counterfactual explanations address causal and non-causal relationships, new horizons open to the XAI community by unifying causal and non-causal explanatory engines within a contractually-driven framework. Furthermore, bringing together researchers from the humanities and the computational sciences could contribute to the further development of contrastive and counterfactual explanations generation. The work in [39] also highlighted that some existing contrastive explanation models cannot be applied to visual reasoning tasks. Visual reasoning involves solving problems about visual information [39]. Therefore, researchers in this area may contribute to finding answers for the reason(s) behind the why-not question and how to generate text or visual explanations for visual reasoning tasks [39].

5.1.11. Communicating uncertainties

Communicating uncertainty is an important research direction because it can help to inform the users about the underlying uncertainties in the model and explanations. According to [24], there are already inherent uncertainties in ML models; and model refinement efforts by developers may introduce new uncertainties (e.g., overfitting). Furthermore, some explanation methods such as permutation feature importance and Shapley value give explanations without measuring the uncertainty implied by the explanations [42].

Quantifying uncertainty is an open research topic [24]. However, some works exist towards quantifying uncertainty in areas such as feature importance, layer-wise relevance propagation, and Shapley values [42]. For example, the work in [91] aimed at quantifying the importance of a variable averaging across the entire population of interest by introducing a method for estimating the Shapley population variable importance measure. The obtained results in a simulation study showed that the method has good finite sample performance. Additionally, the results from an in-hospital mortality prediction task showed that the method yielded similar estimates of variable importance using different machine learning algorithms. Recently, the SHApley eFfects via random Forests (SHAFF) was proposed to estimate Shapley effects for measuring variable importance based on the random forests algorithm [92]. Unlike Monte-Carlo sampling and training one model for each selected subset of variable in [91], SHAFF improved Monte-Carlo sampling by utilizing importance sampling as a means to focus on the most relevant subsets of variables identified by the forest. Additionally, SHAFF allows fast and accurate estimates of the conditional expectations for any variable subset because of the utilized projected random forest algorithm. Through several experiments, it was found that SHAFF offers practical performance improvements over many existing Shapley algorithms.

The uncertainty surrounding ML models can take many forms and occur throughout the ML life phases [24]. Therefore, in order to make progress, it is needed to become more rigorous in studying and reliably quantifying uncertainties at the model's various phases and with the explanation methods and communicate them to the users, then users can respond accordingly [24, 42].

5.1.12. Time constraints

Time is an essential factor in producing explanations and in interpretations. Some explanations must be produced promptly to let the user react to the decision [50]. Producing explanations efficiently can save computing resources, thereby making it useful

for industrial use or in environments with limited computing capability [50]. In some situations (e.g., plant operation application), the provided explanations need to be understood quickly to help the end-user to make a decision [5]. On the other hand, in some situations (e.g., scientific applications), users would likely be willing to devote considerable time to understanding the provided explanation [5]. Therefore, time is an essential factor considering the situation, available resources, and end-users.

5.1.13. Natural language generation

Explaining in natural language needs to be accurate, useful, and easy to understand [93]. Furthermore, in order to produce good quality explanations, the generated explanations need to be tailored to a specific purpose and audience, be narrative and structured, and communicate uncertainty and data quality that could affect the system's output [93].

Four challenges that are crucial in generating good quality explanations have been discussed in [93]:

- Evaluation challenge: The need for developing inexpensive but reliable ways of evaluating the quality of the generated explanation based on a range of rigor levels (e.g., scrutability, trust, etc.). A recent taxonomy for existing automated evaluation metrics for natural language generation can be found in [94].
- Vague Language challenge: Using vague terms in explanations is much easier to understand for humans because they think in qualitative terms [95]. However, how can vague language be used in explanations, such that the user does not interpret it in a way that will lead to a misunderstanding of the situation? In addition, setting the priority of messages based on features and concepts that the user is aware of would be helpful. Furthermore, the phrasing and terminology used should be intuitive to users.
- Narrative challenge: Explaining symbolic reasoning narratively is more straightforward to comprehend than numbers and probabilities [96]. Therefore, we need to develop algorithms for creating narrative explanations to present the reasoning.
- Communicating data quality challenge: Techniques should be developed to keep users informed when data problems affect results. We have discussed this issue in detail in Communicating Data Quality Section.

Another challenge has been discussed in [59]. In some medical domains, it could be necessary for AI systems to generate long textual coherent reports to mimic the behavior of doctors. The challenge here is that after generating a few coherent sentences, language generation models usually start producing seemingly random words that have no connection to previously generated words. One of the solutions to this problem would be to use transformer networks [97] as language model decoders, which can capture word relationships in a longer sentence. In order to evaluate the generated reports, it is essential to compare them with human-generated reports. However, since human-generated reports are usually free-text reports (i.e., not following any specific template), it is important to first eliminate unnecessary information for the final diagnosis from human-generated reports and then conduct the comparison.

5.1.14. Analyzing assumption-free black-box models, not assumption-based data models

The author in [11] has discussed that knowledge can be extracted from the data using ML models that can be interpreted. With proper training, the interpretable ML model can help in identifying features' importance and relationships, and approximate reality to a reasonable degree [11].

The author added that we should focus more on analyzing assumption-free black-box AI models than analyzing assumption-based data models. That is because making assumptions about the data (i.e., distribution assumptions) is questionable. Further, assumption-based data models in many domains are typically less predictive than black-box AI models (i.e., generalization) when having lots of quality data, which is available due to digitization. Therefore, the author has argued that there should be a development of all the tools that statistics offer for answering questions (e.g., hypothesis tests, correlation measures, and interaction measures) and rewriting them for black-box AI models. Some relevant works to these already exist [11]. For example, in a linear model, the coefficients quantify the effects of an individual feature on the result. The partial dependent plot [98] represents this idea in a more generalized form [11].

5.1.15. Reproducibility

In a recent review of XAI models based on electronic health records, it has been found that research reproducibility was not stressed well in the reviewed literature, though it is paramount [9]. In order to facilitate comparisons between new ideas and existing works, researchers should use open data, describe the methodology and infrastructure they used, and share their code [9]. In addition, it has been suggested that publication venues should establish reproducibility standards that authors must follow as part of their publication process [9].

5.1.16. The economics of explanations

Research into the economic perspective of XAI is sparse, but it is essential [1]. With the pressures of social and ethical concerns about trusting black-box AI models, XAI has the potential to drive a real business value [1]. XAI, however, comes at a cost [99].

Recently, the work in [100] identified costs of explanations in seven main categories (1) costs of explanations design, (2) costs of creating and storing audit logs, (3) costs of trade secrets violation (e.g., the forced disclosure of source code), (4) costs of slowing down innovation (e.g., increasing time-to-market), (5) costs of reducing decisional flexibility if the future situation does not justify the previous explanation, (6) cost of conflict with security and privacy matters, and (7) costs of using less efficient models for their interpretability. Therefore, costs associated with algorithmic explanations should be incurred when the benefits of the explanations outweigh the costs [99].

Cost estimation is one of the issues that should be addressed by encouraging economic interpretations. Other issues include algorithms proprietary, revealing trade secrets and predicting XAI market evolution [1].

5.2. Challenges and research directions of XAI based on the ML life cycle's phases

In this section, we reported the challenges and research directions in XAI based on three ML life cycle phases. The selected papers that discussed these challenges and research directions are mentioned in Table 5. As shown in Fig. 8, the majority of these challenges and research directions have received much attention recently (dark blue colors in 2020 and 2021).

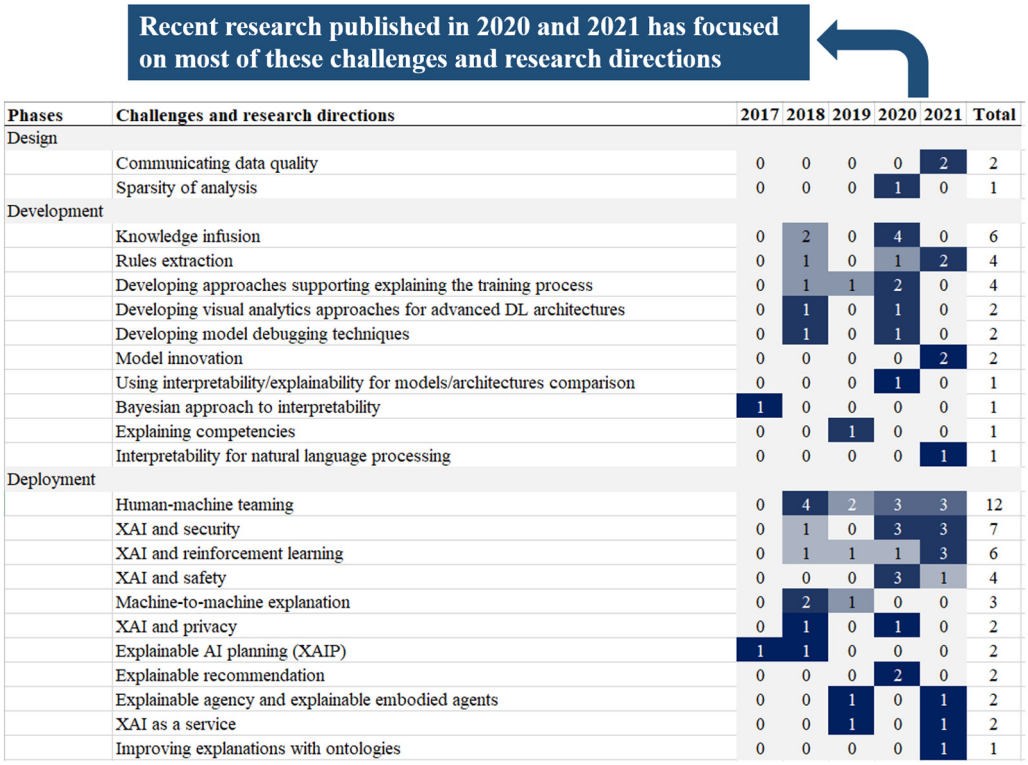
5.2.1. Challenges and research directions of XAI in the design phase

This phase is focused on the processes needed before starting training machine learning models on the given data (e.g., data collection and understanding). Challenges and Research Directions of XAI discussed in this section are shown in Fig. 9

Table 5
A summary of the selected papers, categorized by phases as well as challenges and research directions.

Phases	Challenges and research directions	Papers
Design	Communicating data quality	[18,20]
	Sparsity of analysis	[101][a]
Development	Knowledge infusion	[47,52,64,102,103], [23][a]
	Rules extraction	[38,39,69,104]
	Developing approaches supporting explaining the training process	[24,34,64,102]
	Developing visual analytics approaches for advanced DL architectures	[24,102]
	Developing model debugging techniques	[50,103]
	Model innovation	[65,105]
	Using interpretability/explainability for models/architectures comparison	[24]
	Bayesian approach to interpretability	[22]
	Explaining competencies	[34]
	Interpretability for natural language processing	[56]
Deployment	Human-machine teaming	[1,24,34,36,49,52,69,102,106–108], [61][a]
	XAI and security	[2,36,53,58,64,105,109]
	XAI and reinforcement learning	[10,19,110], [66,75,101][a]
	XAI and safety	[2,53,106,111]
	Machine-to-machine explanation	[1,35,112]
	XAI and privacy	[44,113]
	Explainable AI planning (XAIP)	[1,33][a]
	Explainable recommendation	[106,114]
	Explainable agency and explainable embodied agents	[57,115]
	XAI as a service	[11,53]
	Improving explanations with ontologies	[18]

[a]A non-peer-reviewed paper from arXiv.



Phases	Challenges and research directions	2017	2018	2019	2020	2021	Total
Design							
	Communicating data quality	0	0	0	0	2	2
	Sparsity of analysis	0	0	0	1	0	1
Development							
	Knowledge infusion	0	2	0	4	0	6
	Rules extraction	0	1	0	1	2	4
	Developing approaches supporting explaining the training process	0	1	1	2	0	4
	Developing visual analytics approaches for advanced DL architectures	0	1	0	1	0	2
	Developing model debugging techniques	0	1	0	1	0	2
	Model innovation	0	0	0	0	2	2
	Using interpretability/explainability for models/architectures comparison	0	0	0	1	0	1
	Bayesian approach to interpretability	1	0	0	0	0	1
	Explaining competencies	0	0	1	0	0	1
	Interpretability for natural language processing	0	0	0	0	1	1
Deployment							
	Human-machine teaming	0	4	2	3	3	12
	XAI and security	0	1	0	3	3	7
	XAI and reinforcement learning	0	1	1	1	3	6
	XAI and safety	0	0	0	3	1	4
	Machine-to-machine explanation	0	2	1	0	0	3
	XAI and privacy	0	1	0	1	0	2
	Explainable AI planning (XAIP)	1	1	0	0	0	2
	Explainable recommendation	0	0	0	2	0	2
	Explainable agency and explainable embodied agents	0	0	1	0	1	2
	XAI as a service	0	0	1	0	1	2
	Improving explanations with ontologies	0	0	0	0	1	1

Fig. 8. Number of research papers per year for each challenge and research direction categorized by phases. Method: Publication year was used for peer-reviewed papers published before 2022, otherwise, arXiv's publication year was considered. As can be seen from this figure, the majority of these challenges and research directions have received much attention recently (dark blue colors in 2020 and 2021).

5.2.1.1. Communicating data quality. The provided explanations for the AI system or its outcomes depend on the data used to build the system. Data bias, data incompleteness, and data incorrectness are issues that affect the quality of the data. Training AI systems using low-quality data will be reflected in their outcomes [93]. For example, an AI system developed for lung cancer risk prediction using data from Americans may not accurately estimate risks for a resident of Delhi due to the differences in polluted environments in which they are living at [93]. So, what can be of high quality for a particular purpose can be of low quality for another [20]. Reducing system accuracy is not the only consequence of building an AI system using low-quality data; producing unfair decisions and degrading the explainability of the AI system are other possible consequences.

With this in mind, it has been suggested to be aware of how data was collected and any limitations associated with the collected data [78]. Further, it has been highlighted the importance of clarifying any data issues that can reduce accuracy when

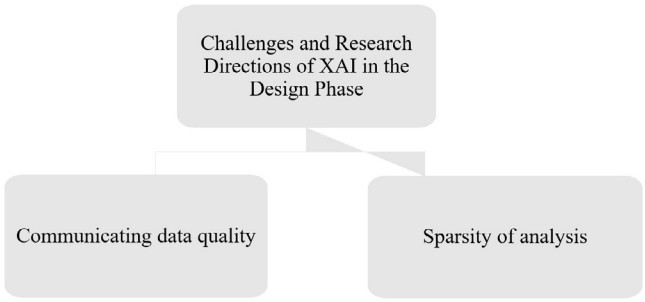


Fig. 9. Challenges and Research Directions of XAI in the Design Phase.

producing explanations [93]. However, how can we communicate data quality to users to let them know how the results are influenced by the data used.

In [116], the authors discussed several issues that arise when producing explanations for AI models that use imputation of missing data. They recommended disclaimers accompanied by the derived explanations and educated end-users about the risks involved in incorrect explanations. Even though it is good to come up with appropriate disclaimers, we believe that future studies should be undertaken to develop a practical and measurable way to communicate data quality to users. Proposing dimensions of data quality could be the basis for that. We recommend starting with the following questions which are inspired by the work in [117]:

- Which essential dimensions of data quality are wanted?
- What are the definitions of those dimensions? and how to measure them?
- How to deal with them to improve the AI model and hence its explanations?
- How to communicate them (and highlight any possible risks)?

According to [18], there is a variety of data quality dimensions such as completeness, accuracy, and consistency. For an extensive list of dimensions of data quality that occur in information systems, the reader may refer to the research paper in [117]. The fairness dimension can also be included, which may include demographic parity differences. It is essential to highlight that the way that can be used to communicate data quality can vary based on the type of users.

5.2.1.2. Sparsity of analysis. Interpreting and validating the reasoning behind an ML model may require examining many visualizations, which is a challenging task for the user, especially if there are a vast number of samples for such examination [101]. Therefore, the number of visualizations that a user has to analyze should be as small as possible to reduce the sparsity of the analysis [101]. A way to achieve that can be by developing novel methods to identify a meaningful subset of the entire dataset to interpret; then, by using this meaningful subset, it is needed to come up with an interpretation of the relationship between various samples and various subsets [101].

5.2.2. Challenges and research directions of XAI in the development phase

There are three main types of learning in ML: supervised, unsupervised, and reinforcement learning. In supervised learning, a learning algorithm is used to train an ML model to capture patterns in the training data that map inputs to outputs. With unsupervised learning, which is used when only the input data is available, an ML model is trained to describe or extract relationships in the training data. For reinforcement learning, an ML model is trained to make decisions in a dynamic environment to perform a task to maximize a reward function. In the following subsections, we discuss the challenges and research directions during the development of ML models (e.g., model training and model understating) (see Fig. 10).

5.2.2.1. Knowledge infusion. A promising research direction is incorporating human domain knowledge into the learning process (e.g., to capture desired patterns in the data). According to [52], understanding how experts analyze images and which regions of the image are essential to reaching a decision could be helpful to come up with novel model architectures that mimic that process. Furthermore, our explanations can be better interpretable and more informative if we use more domain/task-specific terms [23].

Recently, the work in [118] highlights various ways of incorporating approaches for medical domain knowledge with DL models such as transfer learning, curriculum learning, decision level fusion, and feature level fusion. According to that survey, it was seen that with appropriate integrating methods, different kinds of domain knowledge could be utilized to improve the effectiveness of DL models. A review focused on knowledge-aware methods for XAI is given by [47]. Based on the knowledge source, two categories are identified: knowledge methods and knowledge-based methods. Unstructured data is used as a knowledge source in knowledge methods, while knowledge-based methods use structured knowledge to build explanations. According to [47], when we use external domain knowledge, we are able to produce explanations that identify important features and why they matter. As concluded in that survey, many questions remain unanswered regarding utilizing external knowledge effectively. For instance, in a vast knowledge space, how can relevant knowledge be obtained or retrieved? To demonstrate this point, let us take the Human-in-the-loop approach as an example [47]. Typically, a user has a wide range of knowledge in multiple domains; thus, the XAI system must ensure that the knowledge provided to the user is desirable.

Recent works in the knowledge that can be incorporated during training ML are given in [103,119,120]. In [119], a one-shot learning technique was presented for incorporating knowledge about object categories, which may be obtained from previously learned models, to predict new objects when very few examples are available from a given class. Another work in [120] has shown how a knowledge graph is integrated into DL using knowledge-infused learning and presented examples of how to utilize knowledge-infused learning towards interpretability and explainability in education and healthcare. The work in [103] has mentioned that the middle-to-end learning of neural networks with weak supervision via human–computer interaction is believed to be a fundamental research direction in the future.

Based on all that, it can be seen that using XAI to explain the outcomes of the models (e.g., pointing out which regions of the image were used to reach the decision) can help to understand better what was learned from the incorporated human knowledge. Thus, it would help to adjust the way used in incorporating the knowledge or come up with innovations in model architectures. Furthermore, it could be used to confirm whether a model follows the injected knowledge and rules, especially with critical applications, e.g., autonomous driving model [102]. Therefore, more research is needed to investigate how experts can interact with ML models to understand them and improve their abilities, which would be a promising direction in which XAI can contribute.

5.2.2.2. Rules extraction. Historically, the need for explanations dates back to the early works in explaining expert systems and Bayesian networks [4]. Rule extraction from ML models has been studied for a long time [121–123]. However, there is still

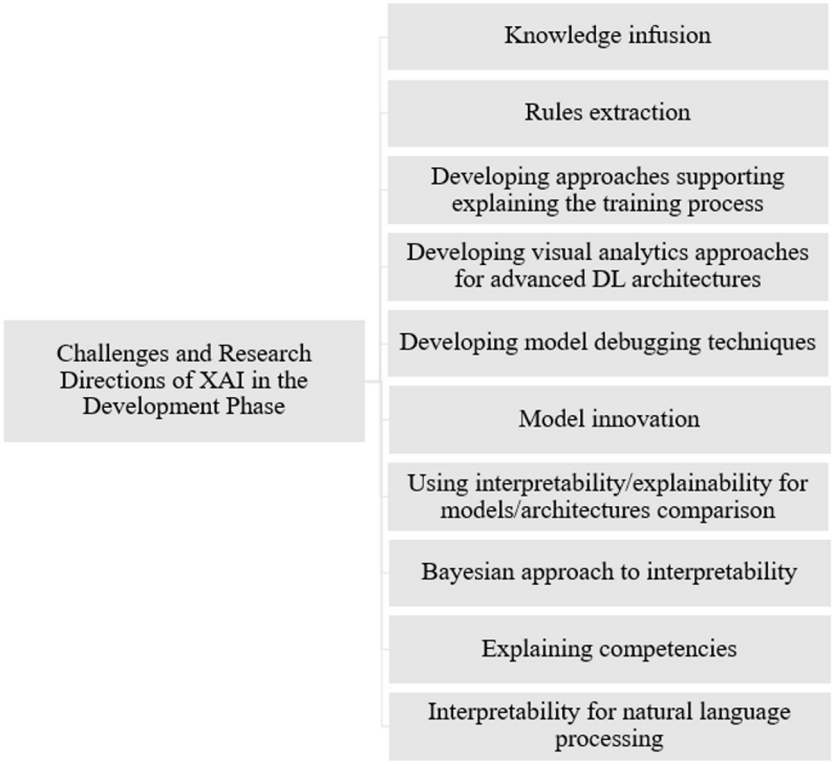


Fig. 10. Challenges and Research Directions of XAI in the Development Phase.

increasing interest in utilizing rule extraction for explainability/interpretability [104,124]. Therefore, to discover methods that may work for explainability/interpretability, we should revisit the past research works [69].

According to [104,123], there are three main approaches for rule extraction: (1) Decomposition approach based on the principle that the rules are extracted at the neuron level, such as visualizing a network's structure, (2) Pedagogical approach that extracts rules that map inputs directly to outputs regardless of their underlying structure, such as computing gradient, (3) Eclectics approach, which is the combination of both decompositional and pedagogical approaches.

Further research for rules extraction is needed, which has been discussed in [104]. First, visualize neural networks' internal structure. Through visualizing each activated weight connection/neuron/filter from input to output, one can understand how the network works internally and produce the output from the input. Second, transform a complex neural network into an interpretable structure by pruning unimportant or aggregating connections with similar functions. By doing this, the overfitting issue can be reduced, and the model's structure becomes easier to interpret. Third, explore the correspondence between inputs and outputs, for example, by modifying the inputs and observing their effects on the output. Fourth, calculate the gradient of the output to the inputs to know their contributions.

In visual reasoning tasks, logic rule extraction and representation are challenging tasks [39]. For rule extraction, predicates and arguments may show multi-granularity concepts, which causes difficulty in the extraction task. For rule representation, there is a need to go beyond implementing discriminative rule representations at the syntactic level, which is the current research focus, to the semantic level. Therefore, further work needs to be done to address the challenges associated with logic rule extraction and representation for interpretable visual reasoning. Finally, it has been suggested to combine the best of DL and fuzzy logic towards an enhanced interpretability [38].

5.2.2.3. Developing approaches supporting explaining the training process. Training ML models, especially DL, is a lengthy process that usually takes hours to days to finish, mainly because of the large datasets used to train the models [102]. Therefore, researchers and practitioners have contributed to developing systems that could help steer the training process and develop better models.

Examples of progressive visual analytics systems are cited in [102]. For example, DeepEyes [125] is an example of a progressive visual analytics system that enables advanced analysis of DNN models during training. The system can identify stable layers, identify degenerated filters that are worthless, identify inputs that are not processed by any filter in the network, reasons the size of a layer, and helps to decide whether more layers are needed or eliminate unnecessary layers. DGMTracker is another example [126] which is developed for better understanding and diagnosing the training process of deep generative models (DGMs). In addition, big tech companies such as Google and Amazon have developed toolkits to debug and improve the performance of ML models such as Tensor-Board[3] and SageMaker Debugger.[4]

Future studies to deal with this challenge are therefore recommended in order to develop XAI approaches supporting the online training monitoring to get insights that could help to steer the training process by the experts, which could help in developing better models and minimizing time and resources [24,102].

5.2.2.4. Developing visual analytics approaches for advanced DL architectures. While visual analytic approaches have been developed for basic DL architectures (e.g., CNNs and RNNs), advanced DL architectures pose several challenges for visual analytic and information visualization communities due to their large number of layers, the complexity of network design for each layer, and

[3] https://www.tensorflow.org/tensorboard
[4] https://aws.amazon.com/sagemaker/debugger/

the highly connected structure between layers [102]. Therefore, developing efficient visual analytics approaches for such architectures in order to increase their interpretability as well as the explainability of their results is needed [24,102]. The work in [127] provided two recent works on using variations of class activation maps (CAMs) to explain the generated results from an ensemble of advanced DL architectures (e.g., ResNet) [128,129]. Therefore, it is expected that several visual analytics approaches will be developed for advanced DL architectures due to the wide applications of the advanced DL architectures.

5.2.2.5. Developing model debugging techniques. The model is already trained at this stage, and we want to discover any problems that can limit its predictions. The debugging of ML models is paramount for promoting trust in the processes and predictions, which could result in creating new applications [103,130], e.g., visual applications for CNN. A variety of debugging techniques exists, including model assertion, security audit, variants of residual analysis and residual explanation, and unit tests [130]. According to [50], understanding what causes errors in the model can form the foundation for developing interpretable explanations. The next step is developing more model debugging techniques and combining them with explanatory techniques to provide insight into the model's behavior, enhance its performance, and promote trust [130].

5.2.2.6. Model innovation. By explaining DL models, we can gain a deeper understanding of their internal structure which can lead to the emergence of new models (e.g., ZFNet [131]) [105]. Therefore, in the future, the development of explanation methods for DL and new DL models are expected to complement each other [105].

Another research area is developing new hybrid models where the expressiveness of opaque models is combined with the apparent semantics of transparent models (e.g., combining a neural network with a linear regression) [65]. This research area can be helpful for bridging the gap between opaque and transparent models and could help in developing highly efficient interpretable models [65].

5.2.2.7. Using interpretability/explainability for models/architectures comparison. It is widely known that the performance of ML models/architectures varies from one dataset/task to another [24]. Usually, error performance metrics are used for the comparison to choose the suitable model/architecture for the given dataset/task and to decide how to combine models/architectures for better performance [24]. However, even if the models may have the same performance, they can use different features to reach the decisions [14]. Therefore, the interpretability/explainability of models can be helpful for models/architectures comparison [14]. It could even be said that the better we understand models' behavior and why they fail in some situations, the more we can use those insights to enhance them [14]. In the future, it is expected that explanations will be an essential part of a more extensive optimization process to achieve some goals such as improving a model's performance or reducing its complexity [8]. Further, XAI can be utilized in models/architectures comparison. For example, the works in [132–134] show some recent works using visual explanations for model comparison.

5.2.2.8. Bayesian approach to interpretability. The work in [22] has discussed that there exist elements in DL and Bayesian reasoning that complement each other. Comparing Bayesian reasoning with DL, Bayesian reasoning offers a unified framework for model development, inference, prediction, and decision-making. Furthermore, uncertainty and variability of outcomes are explicitly accounted for. In addition, the framework has an "Occam's Razor" effect that penalizes overcomplicated models, which makes it robust to model overfitting. However, to ensure computational tractability, Bayesian reasoning is typically limited to conjugate and linear models.

In a recent survey on Bayesian DL (BDL) [135] this complement observation has been exploited, and a general framework for BDL within a uniform probabilistic framework has been proposed. Further research is needed to be done to exploit this complement observation because it could improve model transparency and functionality [22].

5.2.2.9. Explaining competencies. There is a need for users to gain a deeper understanding of the competencies of the AI system, which includes knowing what competencies it possesses, how its competencies can be measured, as well as whether or not it has blind spots (i.e., classes of solutions it never finds) [34]. Through knowledge and competency research, XAI could play a significant role in society. Besides explaining to individuals, there are other roles including leveraging existing knowledge for further knowledge discovery and applications and teaching both agents and humans [34].

5.2.2.10. Interpretability for natural language processing. There are many ways to categorize XAI methods. One standard category is categorizing XAI methods as local or global methods. The local methods explain a single decision from the model, while the global methods explain the entire model [11]. The authors in [56] suggested adding class explanation methods to this category. Such methods are focused on explaining the entire output-class, hence the name [56]. An example of a class explanation method is summarizing a model focusing on one class only [56]. There is still no attention given to these types of methods by researchers (not only for NLP), therefore, these methods should be investigated further [56]. It has been also suggested in [56] to pay attention to particularly developing more explanation methods for sequence-to-sequence models which have numerous real-world applications such as machine translation, question answering, and text summarization.

5.2.3. Challenges and research directions of XAI in the deployment phase

The following subsections are dedicated to challenges and research directions during the deployment of AI systems. The deployment phase starts with deploying ML solutions until we stop using the solutions (or maybe after that). Challenges and Research Directions of XAI discussed for this phase are shown in Fig. 11.

5.2.3.1. Human-machine teaming. Most provided explanations for AI systems are typically static and carry one message per explanation [69]. Explanations alone do not translate to understanding [1]. Therefore, for a better understanding of the system, users should be able to explore the system via interactive explanations, which is a promising research direction to advance the XAI field [1,69] as the majority of current XAI libraries lack user interactivity and personalization of the explanations [107].

Even though there are already some works in this research direction as has been reported in [69], much work is still needed to tailor interfaces to different audiences, exploit interactivity, and choose appropriate interactions for better visualization designs [24,69]. Various works have also been suggested to go beyond static explanations and enhance human-machine teaming. In [52], open-ended visual question answering (VQA) has been suggested to be used rather than providing a report with too many details. Here, an user queries (or make follow-up questions), and the system answers. Achieving that would provide better interaction between the system and the expert user. In another work [102], it has been mentioned that generative models

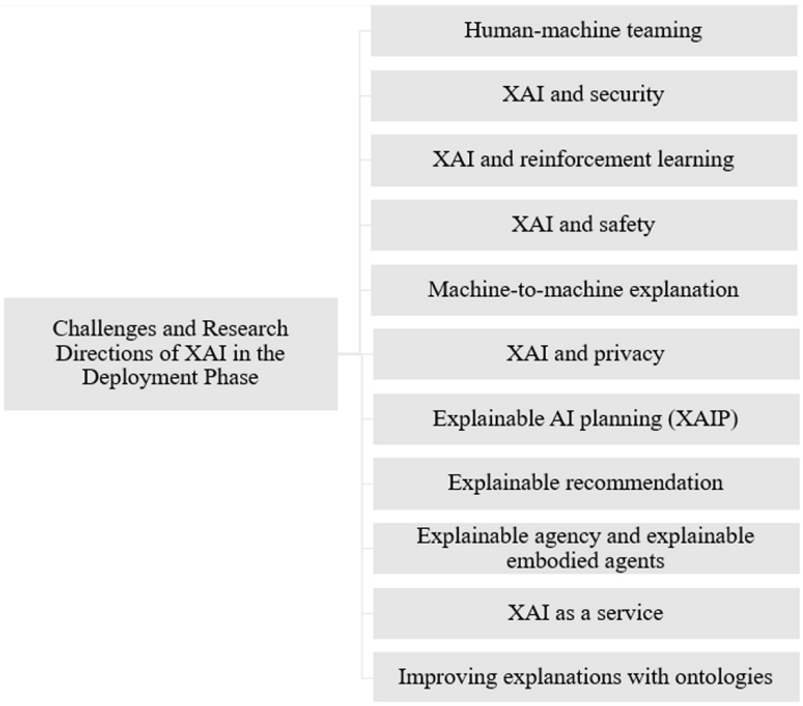


Fig. 11. Challenges and Research Directions of XAI in the Deployment Phase.

can allow for interactive DL steering because they allow for multiple answers. They highlighted that developing new DL models capable of adapting to various user inputs and generating outputs accordingly as well as developing visualization-based interfaces that enable effective interaction with DL systems are promising research areas in the future.

In [69], rather than providing static explanations, the authors have suggested building on existing intelligibility work for context-aware systems (e.g., design space explorations, conceptual models for implicit interaction, and intelligible interfaces for various scenarios and using a variety of modalities). Additionally, they have highlighted a research area that is effectively interacting with AI augmentation tools. In [136], an XAI system was proposed which involves a group of robots that attempt to infer human values (i.e., the importance of various goals) learned using a cooperative communication model while providing explanations of their decision-making process to the users. Through experiments, it was found that it is possible to achieve real-time mutual understanding between humans and robots in complex cooperative tasks through a learning model based on bidirectional communication. The work in [108] stressed the importance and benefits of multimodal interactive explanations with the users for better understanding and traceability of model working and its decisions, better user satisfaction and trust, improving transparency in the decision-making process, and acting as an enabler to make advances in human–computer interaction field. In [1], it has been emphasized the importance of bridging HCI empirical studies with human sciences theories to make explainability models more human-centered models. In this way, adaptive explainable models would emerge by providing context-aware explanations that could be adapted to any changes in the parameters of their environment, such as user profile (e.g., expertise level, domain knowledge, cultural background, interests, and preferences) and the explanation request setting (e.g., justification).

The authors in [24] have mentioned that extracting, visualizing, and keeping track of the history of interaction data between users and systems can allow users to undo certain actions and examine them interactively would help to address some common challenges (e.g., hyperparameter exploration). Finally, the authors in [106] have highlighted that user-friendliness and intelligent interface modalities need to take into account the type of explanations that meet users' goals and needs. For example, the system can ask for feedback from the users to know how good was the provided explanations (e.g., "explain more", "redundant explanation", or "different explanation"). Such interaction can help to improve future explanations.

Taken together, it seems that different ways are needed to enhance human-machine teaming. Approaching HCI and other related studies can contribute to making explainability models more human-centered. In addition, humans can provide feedback on the provided explanations, which can help in improving future explanations.

5.2.3.2. XAI and security. Two main concerns have been discussed for XAI and security: confidentiality and adversarial attacks [2,36,105,109]. For the confidentiality concern, several aspects of a model may possess the property of confidentiality [2]. As an example given by [2], think of a company invested in a multi-year research project to develop an AI model. The model's synthesized knowledge may be regarded as confidential, and hence if only inputs and outputs are made available, one may compromise this knowledge [137]. The work in [138] presented the first results on how to protect private content from automatic recognition models. Further research is recommended to develop XAI tools that explain ML models while maintaining models' confidentiality [2].

Turning now to the adversarial attacks concern, the information revealed by XAI can be utilized in generating efficient adversarial attacks to cause security violations, confusing the model and causing it to produce a specific output, and manipulation of explanations [2,109]. In adversarial ML, three types of security violations can be caused by attackers using adversarial examples [139]: integrity attacks (i.e., the system identifies intrusion points as normal), availability attacks (i.e., the system makes multiple classification errors, making it practically useless), and privacy violation (i.e., violating the privacy of system

users). Attackers can do such security violations because an AI model can be built based on training data influenced by them, or they might send carefully crafted inputs to the model and see its results [139]. According to [105], existing solutions to handle perturbations still suffer from some issues, including instabilities and lack of variability. Therefore, it is necessary to develop new methods to handle perturbations more robustly [36,105]. Additionally, a challenge ahead for XAI is to come up with provable guarantees that the provided explanations for the predictions are robust to many external distortion types [53].

The information uncovered by XAI can also be utilized in developing techniques for protecting private data, e.g., utilizing generative models to explain data-driven decisions [2]. Two recent research directions have been highlighted in this context [2]: using generative models as an attribution method to show a direct relationship between a particular output and its input variables [140]. The second is creating counterfactuals through generative models [141]. It is expected that generative models will play an essential role in scenarios requiring understandable machine decisions [2].

5.2.3.3. XAI and reinforcement learning. The use of DL by reinforcement learning (RL) has been applied successfully to many areas [101]. Through the explicit modeling of the interaction between models and environments, RL can directly address some of the interpretability objectives [19]. Despite that, unexplained or non-understandable behavior makes it difficult to users to trust RL agents in a real environment, especially when it comes to human safety or failure costs [101]. Additionally, we lack a clear understanding of why an RL agent decides to perform an action and what it learns during training [101]. RL's interpretability can help in exploring various approaches to solving problems [101]. For instance, understanding why the RL AlphaFold system [142] is capable of making accurate predictions can assist bioinformatics scientists in understanding and improving the existing techniques in protein structures to speed produce better treatment before new outbreaks happen [101].

Recently, the work in [110] highlighted several issues that need to be addressed and potential research directions in the area of XAI for RL. The authors find that the selected studies used "toy" examples or case studies that were intentionally limited in scope mainly to prevent the combinatorial explosion problem in the number of combinations of states and actions. Therefore, more focus on real-world applications has been suggested. It has also been mentioned that there is a lack of new algorithms in the area. Therefore, the design of RL algorithms with an emphasis on explainability is essential. Symbolic representations can be utilized so RL agents can inherently be explained and verified. Another issue is highlighted, which is the lack of user testing with the existing approaches, which is in line with what was mentioned in [67]. As for the complexity of the provided explanations, it has been found that the current focus is presenting explanations for users with a background in AI. Therefore, it has been suggested to conduct further research to present the explanations for those who might interact with the agents, which may have no background in AI. For example, providing more visceral explanations, e.g., annotations in a virtual environment. Additionally, enriching visualization techniques by considering the temporal dimensions of RL and multi-modal forms of visualization, e.g., virtual or augmented reality. Lastly, it has been emphasized the importance of open-source code sharing for the academic community.

Another interesting point for consideration has been highlighted in [10], which is learning from explanations. The work in [143] provides a starting point, which presents an agent who trained to simulate the Mario Bros. game using explanations instead of prior play logs.

Finally, various RL techniques such as hierarchical, multi-goal, multi-objective, and intrinsically motivated learning have been suggested to be used in goal-driven explanation and emotion-aware XAI [75]. Further, event-based and expectation-based explanations can be investigated to increase the usage of RL in human-agent mixed application domains [75].

5.2.3.4. XAI and safety. Trust and acceptance are benefits of explainability/interpretability [106]. However, focusing on benefits without considering the potential risks may have severe consequences (e.g., relying too much or too little on the advice provided by the prescription recommendation system) [106]. Several studies have been conducted to evaluate the safety of processes that depend on model outputs because erroneous outputs can lead to harmful consequences in some domains [2]. Therefore, possible risks must be at the top priority when designing the presented explanations [106].

Many techniques have been proposed to minimize the risk and uncertainty of adverse effects of decisions made using model outputs [2]. As an example, the model's output confidence technique can examine the extent of uncertainty resulting from a lack of knowledge regarding the inputs and the corresponding output confidence of the model to notify the user and cause them to reject the output produced by the model [2]. In order to achieve this, explaining what region of the inputs was used by the model to arrive at the outcome can be used for separating out such uncertainty that may exist within the input domain [2]. Additionally, as has been suggested in [106], it is important to develop explanations that evolve with time, keeping in mind past explanations for long-term interactions with end-users and identifying ways to minimize risks. Developing evaluation metrics and questionnaires would be essential to integrate the user-centric aspects of explanations as well as evaluating error-proneness and any possible risks [106]. Finally, in [111], some major challenges have been discussed, including developing distance metrics that more closely reflect human perception, improvement to robustness by designing a set of measurable metrics for comparing the robustness of black-box AI models across various architectures, verification completeness using various verification techniques, scalable verification with tighter bounds, and unifying formulation of interpretability. It is good to note that the utilization of formal verification methods has been suggested as a potential step to move forward towards establishing a truly safe and trustworthy model [53].

5.2.3.5. Machine-to-machine explanation. A promising area of research is enabling machine-to-machine communication and understanding [112]. Furthermore, it is an important research area because of the increasing adoption of the Internet of Things (IoT) in different industries. A growing body of research has begun exploring how multiple agents can efficiently cooperate and exploring the difference between explanations intended for humans and those intended for machines [35,112].

According to [35], future explainable approaches are likely to provide both human and machine explanations, especially adaptive explainable approaches [1]. For machine explanations, complex structures that are beyond the comprehension of humans may be developed [112]. However, how is it possible to measure the success of "transfer of understanding" between agents? The work in [112] has suggested a metric for that, which is measuring the improvement of agent B's performance on a particular task, or set of tasks, as a result of the information obtained from agent A - though it will be crucial to determine some key details, such as the bandwidth constraints and already existing knowledge with agent A.

Based on what has been mentioned above, it is expected that much work is going to be done on how to construct machine

explanations, how to communicate these explanations, and which metrics we need to measure as a success of the transfer of understanding between agents and how to measure them. With more research into how machines communicate/explain themselves, we will be able to understand intelligence better and create intelligent machines [11].

5.2.3.6. XAI and privacy. When individuals are affected by automated decision-making systems, two rights conflict: the right to privacy and the right to an explanation [144]. At this stage, it could be a demand to disclose the raw training data and thus violate the privacy rights of the individuals from whom the raw training data came [144]. Another legal challenge has been discussed in [44], which is the right to be forgotten [145]. By this right, individuals can claim to delete specific data so that they cannot be traced by a third party [44]. Data preservation is another related issue because to use XAI to justify a decision reached by automated decision-making, the raw data used for training must be kept, at least until we stop using the AI solution.

One of the key challenges is establishing trust in the handling of personal data, particularly in cases where the algorithms used are challenging to understand [113]. This can pose a significant risk for acceptance to end-users and experts alike [113]. For example, end-users need to trust that their personal information is secured and protected as well as that only their consented data is used, while experts need to trust that their input is not altered later [113].

Anonymization of data can be used to obscure the identity of people. However, privacy cannot always be protected by anonymization [144]. According to [144], the more information in a data set, the greater the risk of de-anonymization, even if the information is not immediately visible. Asserting that anonymization helps conceal who supplied the data to train the automated decision-making system might be comforting for the individuals whom the training data came from, but this does not the case with individuals who are entitled to an explanation of the results produced by the system [144].

In order to address some issues with anonymization techniques, it is recommended that further research should be undertaken in privacy-aware ML, which is the intersection between ML and security areas [113]. XAI can play an essential role in this matter because to develop new techniques to ensure privacy and security, it will be essential to learn more about the inner workings of the system they are meant to protect [113]. In addition, in the future, to promote the acceptance of AI and increase privacy protection, XAI needs to provide information on how the personal data of a particular individual was utilized in a data analysis workflow [44]. However, according to [144], what if it is needed to review the data of many individuals and they may not have consented to review their data in litigation. In such cases, a path to review data for which individuals have not consented would be demanded, but it would be difficult to find such a path [144].

Data sharing is another related issue because AI is used as a data-driven method and therefore any requested explanations depend on data used to build AI systems. Data sharing in this context means making raw data available to be used by other partners [113]. According to [113], the implementation of watermarking or fingerprinting are a typical reactive technique used to deal with this issue. Federated learning can be a possible solution to avoid raw data sharing. Federated learning allows building ML models using raw data distributed across multiple devices or servers [146,147]. Even though the data never leaves the user's device, increasing the number of clients involved in a collaborative model makes it more susceptible to inference attacks intended to infer sensitive information from training data [147,148]. Possible research directions to deal with privacy challenges of federated learning have been discussed in [147] such as privacy-preserving security assurance, defining optimal bounds of noise ratio, and proposing granular and adaptive privacy solutions.

5.2.3.7. Explainable AI planning (XAIP). Existing literature focuses mainly on explainability in ML, though similar challenges apply to other areas in AI as well [1]. AI planning is an example of such an area that is important in applications where learning is not an option [33]. Recent years have seen increased interest in research on explainable AI planning (XAIP) [149]. XAIP includes a variety of topics from epistemic logic to ML, and techniques including domain analysis, plan generation, and goal recognition [149]. There are, however, some major trends that have emerged, such as plan explanations, contrastive explanations, human factors, and model reconciliation [149].

Recently, the work in [33] has explored the explainability opportunities that arise in AI planning. They have provided some of the questions requiring explanation. They also have described initial results and a roadmap towards achieving the goal of generating effective explanations. Additionally, they have suggested several future directions in both plan explanations and executions. Temporal planning, for instance, can open up interesting choices regarding the order of achieving (sub)goals. It is also interesting to consider whether giving the planner extra time to plan would improve the performance. In addition, one of the challenges in plan execution is explaining what has been observed at the execution time that prompts the planner to make a specific choice. As with XAI, it is crucial to have a good metric for XAIP that defines what constitutes a good explanation. Finally, it is imperative that the existing works on XAIP be reconsidered and leveraged so that XAIP will be more effective and efficient when used in critical domains.

5.2.3.8. Explainable recommendation. Explainable recommendation aims to build models that produce high quality recommendations as well as provide intuitive explanations that can help to enhance the transparency, persuasiveness, effectiveness, trustworthiness, and satisfaction of recommendation systems [114].

The work in [114] conducted a comprehensive survey of explainable recommendations, and they discussed potential future directions to promote explainable recommendations. With regards to the methodology perspective, it has been suggested that (1) further research is needed to make deep models explainable for recommendations because we still do not fully understand what makes something recommended versus other options, (2) develop knowledge-enhanced explainable recommendation which allows the system to make recommendations based on domain knowledge, e.g., combine graph embedding learning with recommendation models, (3) use heterogeneous information for explainability such as multi-modal explanations, transfer learning over heterogeneous information sources, information retrieval and recommendation cross-domain explanations, and the impact that specific information modalities have on user receptiveness on the explanations, (4) develop context-aware explainable recommendations, (5) aggregate different explanations, (6) integrate symbolic reasoning and ML to make recommendations and explainability better by advancing collaborative filtering to collaborative reasoning, (7) further research is needed to help machines explain themselves using natural language, and (8) with the evolution of conversational recommendations powered by smart agent devices, users may ask "why" questions to get explanations when a recommendation does not make sense. Therefore, it is essential to answer the "why" in conversations which could help to improve system efficiency, transparency, and trustworthiness.

For the evaluation perspective, the authors in [114] have suggested the importance of developing reliable and easily implemented evaluation metrics for different evaluation perspectives

(i.e., user perspective and algorithm perspective). Additionally, evaluating explainable recommendation systems using user behavior perspectives may be beneficial as well. Lastly, it has been highlighted that explanations should have broader effects than just persuasion. For example, investigate how explanations can make the system more trustworthy, efficient, diverse, satisfying, and scrutable.

In [106], the authors have presented several research challenges in delivery methods and modalities in user experience. As mentioned in that paper, for the delivery method, the current focus in the literature is on providing an explanation to the users while they are working on a task or looking for recommendations. However, more focus should be done on the long-term retrieval of such explanations, for example, through a digital archive, and their implications for accountability, traceability, and users' trust and adoption. That could increase the adoption of intelligent human-agent systems in critical domains. Another challenge is designing autonomous delivery capable of considering the context and situation in which users may need explanations and suitable explanations for them. It is worth mentioning that privacy matters should be taken into account when deriving the recommendations.

It has also been highlighted in [106] that users' goals and needs would have to be met by user-friendly and intelligent interface modalities that provide appropriate explanations. Further, interaction with the system is needed and could help to improve future generated explanations. Finally, focusing on the benefits of explainability without considering the potential risks may have severe consequences. Therefore, when designing explanations, possible risks should be the first priority.

5.2.3.9. Explainable agency and explainable embodied agents. Explainable agency refers to a general capability in which autonomous agents must provide explanations for their decisions and the reasons leading to these decisions [150]. Based on the three explanation phases proposed in [151], the authors in [115] presents a research roadmap for the explainable agency.

The first phase is explanation generation which is intended to explain why an action/result was taken/achieved [115]. This phase of research focuses on the following key research directions: (1) there is a need to connect the internal AI mechanism of the agent/robot with the explanation generation module, (2) to produce dynamic explanations, new mechanisms are required for identifying relevant explanation elements, identifying its rationales, and combining these elements to form a coherent explanation.

The second phase is the explanation communication phase. Here, the focus is on what content end users will receive and how to present that content [151]. According to [115], explainable agents/robots may be deployed in a variety of environments. Therefore, in some cases, multimodal explanation presentations (e.g., visual, audio, and expressive) could be a useful explanation communication approach for enabling efficient explainable agency communication.

For the last phase, explanation reception, the focus is on the human's understanding of explanations. Some considerations should be taken into account to ensure an accurate reception [115]. It is important to develop metrics to measure the explanations' effectiveness and the users' reaction to the provided explanations. In addition, the agent/robot should maintain a model of user knowledge and keep updating it based on the evolution of user expertise and the user's perception of the State of Mind (SoM) of the agent/robot, i.e., an internal representation of how the agent/robot treats the outer world.

The work in [57] reviewed the works related to explainable embodied agents. Embodied agents can interact with humans using both verbal and non-verbal communicative behaviors [57]. Although these behaviors, the actions taken by agents are not necessarily understandable [57]. Therefore, there is an increasing interest in how to make embodied agents explainable [57]. According to [57], there are still unanswered questions in the literature on explainable embodied agents that need further investigation like what are the suitable models that can help to predict/track human expectations/beliefs about the goals and actions of an embodied agent? What is the efficient way to include the Human-in-the-loop approach when designing embodied agents with explainability? What are the impact of the environment and social cues embodiment in the selection of social cues used for explainability (e.g., speech, text, or movement)? How can we best objectively measure trust? and why there is a mixed impact of explainability on the efficiency of the human-agent interaction?

5.2.3.10. XAI as a service. There is an increasing trend in developing automated ML (AutoML) tools [11]. AutoML tool is an end-to-end pipeline starting with raw data and going all the way to a deployable ML model. Model-agnostic explanation methods are applicable to any ML model resulting from automated ML [11]. Similarly, we can automate the explanation step: calculate the importance of each feature, plot the partial dependence, construct a surrogate model, etc [11]. Further, at a more advanced level, Auto XAI can be further designed to extract collective variables and explain their terms, for example, extracting mathematical formulas used in the formation of the collective variables and then using these formulas to explain the generated predictions by ML [53].

Some existing AutoML tools provide automatically generated explanations, e.g., AutoML H2O [152] and MLJAR AutoML [153]. We expect that more Auto XAI tools will be available in the future, either incorporated with AutoML tools or as services. Since these would be services, so one can expect that these services would be developed to be of great help to a wide range of end-users (e.g., non-technical experts).

5.2.3.11. Improving explanations with ontologies. An ontology is defined as *"an explicit specification of a conceptualization"* [154]. The use of ontologies for representing knowledge of the relationships between data is helpful for understanding complex data structures [18]. Therefore, the use of ontologies can help to produce better explanations as found in [155,156].

The work in [18] has discussed some recent works of the literature on this topic such as [155,156]. In [155], Doctor XAI was introduced as a model-agnostic explainer that focused on explaining the diagnosis prediction task of Doctor AI [157], which is a black-box AI model that predicts the patient's next visit time. It was shown that taking advantage of the temporal dimension in the data and incorporating the domain knowledge into the ontology helped improve the explanations' quality. Another work in [156] showed that ontologies can enhance human comprehension of global post-hoc explanations, expressed in decision trees.

It should be noted that ontologies are thought of as contributing a lot to explaining AI systems because they provide a user's conceptualization of the domain, which could be used as a basis for explanations or debugging [158]. Towards that goal, new design patterns, new methodologies for creating ontologies that can support explainable systems, and new methods for defining the interplay between ontologies and AI techniques are needed [158]. Furthermore, it is essential to conduct several user studies to determine the benefits of combining ontologies with explanations [18].

6. What do we think?

XAI is a hot research direction with many existing original and survey papers published covering various aspects of XAI. We believe this will continue due to the need for XAI from regulatory, scientific, industrial, model developmental, and end-user and social perspectives (as discussed in Section 2).

In the literature, there seems to be no agreement on what "explainability" or "interpretability" mean. There are some examples in the selected papers that distinguish between the two terms, despite the fact that they are often used interchangeably in the literature. We think a distinction is needed towards more formalism for XAI. Therefore, we proposed a distinction between the two terms (as discussed in Section 3), in which explainability aims to satisfy a need by providing insights through explainability techniques for the targeted audience, whereas interpretability is more about how the provided insights can make sense for the targeted audience's domain knowledge to be able to reason/infer to support decision-making.

The reported challenges and research directions of XAI in the literature are scattered so placing them in a meaningful context was a challenge. Since XAI is meant to detect and prevent or at least mitigate bias that can occur along all phases of the ML pipeline, our taxonomy was developed based on ML life cycle's phases so it can help the readers to better understand the type of challenges and research directions of XAI in general and in each ML life cycle's phases.

Even though the reported challenges and research directions are presented individually in 39 points (as discussed in Section 5), they can overlap and combine based on researchers' backgrounds and interests, resulting in new research opportunities where XAI can play an important role. Researchers also can use these points (or some of them) as abstract ideas and think about how these challenges can be further explored in their domains. There are many domains in which XAI can promote ML algorithms' adoption, but medicine is one of the most essential. Below are some points, which are derived based on the discussed challenges and research directions in Section 5), that need further exploration:

- How to determine if any newly proposed approach is better at explaining ML models compared to other existing ML models considering that there is no agreement on what "explainability" or "interpretability" mean and the lack of formalized rigorous evaluation metrics?
- Since it is crucial to tailor explanations based on user experience and expertise, how one can define the meaning of a quality explanation, and how we can measure the degree of explanation quality?
- Are the explanations provided by the existing XAI methods tailored to different users (e.g., radiologists, medical image analysis scientists) or only designed for users with ML backgrounds?
- How different types of data and explanations can be composited in a framework to help generate several types of explanations so that they are comprehensive and diverse to support clinicians to care for patients?
- How to better communicate uncertainty to inform the clinicians about the underlying uncertainties in the model and explanations?
- What data quality dimensions can be used to communicate data quality and how these dimensions can be measured?
- How can XAI help medical experts to interact with ML models to understand what was learned from the incorporated human domain knowledge and improve their abilities? Can this help to come up with innovations in model architectures?
- How can XAI establish trust in handling personal data in medicine considering the right to privacy and the right to an explanation?

These are some questions on how some of the discussed challenges and research directions in our meta-survey can be considered in medicine for further research works (which can also be tailored to other domains).

7. Conclusions

In this systematic meta-survey paper, we presented two main contributions to the literature on XAI. First, we proposed an attempt to present a distinction between explainability and interpretability terms. Second, we shed light on the significant challenges and future research directions of XAI resulting from the selected 73 papers, which guide future exploration in the XAI area.

The discussion is divided into two main themes. On the first theme, we discussed general challenges and research directions in XAI. As the second theme, we have discussed the challenges and research directions of XAI based on ML life cycle phases.

For the first theme, we have highlighted the following points: (1) The importance of working towards more formalism (e.g., establishing systematic definitions and the formalization and quantification of explanations and performance metrics), (2) The importance of tailoring explanations based on user experience and expertise, (3) The role that XAI can play in fostering trustworthy AI, (4) The value of multidisciplinary research collaborations in offering new insights for explainable methods, (5) The interpretability vs. performance trade-off, (6) The value of explainability methods composition for more powerful explanations, (7) The value of causal, contrastive, and counterfactual explanations for better human understanding, (8) The importance to put much efforts to explain other data types (e.g., sequences, graphs, and spatio-temporal data), (9) The challenges in the existing XAI models/methods, (10) The value of communicating uncertainty to the users to know about the underlying uncertainties in the model and explanations, (11) The value of time as an essential factor in producing explanations and in interpretation. Time matters when considering the situation, available resources, and end users, (12) The challenges of generating good quality explanations from a natural language generation perspective, (13) the advantages of analyzing models rather than data, (14) The imperative of establishing reproducibility standards for XAI models to facilitate comparisons between new ideas and existing works, and (15) The importance to know when it is reasonable to incur additional costs for explanations.

For the second theme, during the design phase, it is important to communicate data quality to users, which can vary based on the type of users. The quality of data used to train AI systems can reduce their performance as well as cause unfair decisions and deteriorate the explainability of the AI system. Therefore, developing a practical and measurable way to communicate data quality to users is essential. In addition, there is a need to reduce the challenge of the sparsity of the analysis that a user has to analyze if there are a huge number of samples.

For the development phase, we have highlighted that XAI can help explain how the included human knowledge has contributed to the outcomes of the models. Thus, it would help with changing the way utilized in integrating the knowledge or come with innovations in model architectures. Other research directions are utilizing rule extraction for explainability/interpretability, developing XAI approaches for explaining the training process, developing visual analytic approaches for advanced DL architectures, developing model debugging techniques and combining them

with explanatory techniques, using XAI approaches to gain a deeper understanding of the internal structure of the models and then develop new models, using interpretability/explainability for models/architectures comparison, utilizing Bayesian approach to interpretability, explaining the competencies of the AI systems, and interpretability for natural language processing.

With regards to the deployment phase, we have discussed human-machine teaming, XAI and security, some issues and research directions in the area of XAI for reinforcement learning, XAI and safety, machine-to-machine explanation, the two rights conflict (i.e., privacy and explanation). In addition, we pointed out the need to focus on explainability in other AI areas (e.g., explainable AI planning, explainable recommendations, and explainable agency and explainable embodied agents). Finally, pointing out to the prominence of Auto XAI as a service, and the potential of improving explanations with ontologies.

Finally, this meta-survey has three limitations. First, because we cannot ensure that the selected keywords are complete, we could miss some very recent papers. Second, to avoid listing the challenges and future research directions per each paper, we come up with the reported 39 points, which are the results of combining what was reported in the selected papers based on the authors' points of view. Third, we believe that more challenges and future research directions can be added where XAI can play an important role in some domains, such as IoT, 5G, and digital forensics. However, related surveys did not exist at the time of writing this meta-survey.

CRediT authorship contribution statement

Waddah Saeed: Conceptualization, Methodology, Investigation, Writing – original draft, Writing – review & editing. **Christian Omlin:** Conceptualization, Writing – review & editing, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

Acknowledgment

Open access funding provided by University of Agder.

References

[1] A. Adadi, M. Berrada, Peeking inside the black-box: A survey on explainable artificial intelligence (XAI), IEEE Access 6 (2018) 52138–52160, http://dx.doi.org/10.1109/ACCESS.2018.2870052.

[2] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, F. Herrera, Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI, Inf. Fusion 58 (2020) 82–115, http://dx.doi.org/10.1016/j.inffus.2019.12.012.

[3] D. Gunning, Broad Agency Announcement Explainable Artificial Intelligence (XAI), Technical report, 2016.

[4] O. Biran, C. Cotton, Explanation and justification in machine learning: A survey, in: IJCAI-17 Workshop on Explainable AI, Vol. 8, XAI, (1) 2017, pp. 8–13.

[5] F. Doshi-Velez, B. Kim, Towards a rigorous science of interpretable machine learning, 2017, arXiv preprint arXiv:1702.08608.

[6] K. Gade, S.C. Geyik, K. Kenthapadi, V. Mithal, A. Taly, Explainable AI in industry: Practical challenges and lessons learned: Implications tutorial, in: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, in: FAT* '20, Association for Computing Machinery, New York, NY, USA, 2020, p. 699, http://dx.doi.org/10.1145/3351095.3375664.

[7] B. Goodman, S. Flaxman, European union regulations on algorithmic decision-making and a "right to explanation", AI Mag. 38 (3) (2017) 50–57, http://dx.doi.org/10.1609/aimag.v38i3.2741.

[8] W. Samek, K.-R. Müller, Towards explainable artificial intelligence, in: Explainable AI: Interpreting, Explaining and Visualizing Deep Learning, Springer International Publishing, Cham, 2019, pp. 5–22, http://dx.doi.org/10.1007/978-3-030-28954-6_1.

[9] S.N. Payrovnaziri, Z. Chen, P. Rengifo-Moreno, T. Miller, J. Bian, J.H. Chen, X. Liu, Z. He, Explainable artificial intelligence models using real-world electronic health record data: a systematic scoping review, J. Am. Med. Inform. Assoc. 27 (7) (2020) 1173–1185, http://dx.doi.org/10.1093/jamia/ocaa053.

[10] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, D. Pedreschi, A survey of methods for explaining black box models, ACM Comput. Surv. 51 (5) (2018) http://dx.doi.org/10.1145/3236009.

[11] C. Molnar, Interpretable Machine Learning, 2019, https://christophm.github.io/interpretable-ml-book/.

[12] L. Veiber, K. Allix, Y. Arslan, T.F. Bissyandé, J. Klein, Challenges towards production-ready explainable machine learning, in: 2020 {USENIX} Conference on Operational Machine Learning, OpML 20, 2020.

[13] R. Confalonieri, L. Coba, B. Wagner, T.R. Besold, A historical perspective of explainable artificial intelligence, WIREs Data Min. Knowl. Discov. 11 (1) (2021) e1391, http://dx.doi.org/10.1002/widm.1391.

[14] W. Samek, T. Wiegand, K.-R. Müller, Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models, 2017, arXiv preprint arXiv:1708.08296.

[15] L. Arras, F. Horn, G. Montavon, K.-R. Müller, W. Samek, " What is relevant in a text document?": An interpretable machine learning approach, PLoS One 12 (8) (2017) e0181142.

[16] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, R. Fergus, Intriguing properties of neural networks, 2013, arXiv preprint arXiv:1312.6199.

[17] A. Nguyen, J. Yosinski, J. Clune, Deep neural networks are easily fooled: High confidence predictions for unrecognizable images, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR, IEEE, 2015, pp. 427–436.

[18] N. Burkart, M.F. Huber, A survey on the explainability of supervised machine learning, J. Artificial Intelligence Res. 70 (2021) 245–317, http://dx.doi.org/10.1613/jair.1.12228.

[19] Z.C. Lipton, The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery, Queue 16 (3) (2018) 31–57, http://dx.doi.org/10.1145/3236386.3241340.

[20] A.F. Markus, J.A. Kors, P.R. Rijnbeek, The role of explainability in creating trustworthy artificial intelligence for health care: A comprehensive survey of the terminology, design choices, and evaluation strategies, J. Biomed. Inform. 113 (2021) 103655, http://dx.doi.org/10.1016/j.jbi.2020.103655.

[21] Z. Akata, D. Balliet, M. de Rijke, F. Dignum, V. Dignum, G. Eiben, A. Fokkens, D. Grossi, K. Hindriks, H. Hoos, H. Hung, C. Jonker, C. Monz, M. Neerincx, F. Oliehoek, H. Prakken, S. Schlobach, L. van der Gaag, F. van Harmelen, H. van Hoof, B. van Riemsdijk, A. van Wynsberghe, R. Verbrugge, B. Verheij, P. Vossen, M. Welling, A research agenda for hybrid intelligence: Augmenting human intellect with collaborative, adaptive, responsible, and explainable artificial intelligence, Computer 53 (8) (2020) 18–28, http://dx.doi.org/10.1109/MC.2020.2996587.

[22] S. Chakraborty, R. Tomsett, R. Raghavendra, D. Harborne, M. Alzantot, F. Cerutti, M. Srivastava, A. Preece, S. Julier, R.M. Rao, T.D. Kelley, D. Braines, M. Sensoy, C.J. Willis, P. Gurram, Interpretability of deep learning models: A survey of results, in: 2017 IEEE SmartWorld, Ubiquitous Intelligence Computing, Advanced Trusted Computed, Scalable Computing Communications, Cloud Big Data Computing, Internet of People and Smart City Innovation, SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI, 2017, pp. 1–6, http://dx.doi.org/10.1109/UIC-ATC.2017.8397411.

[23] Y. Zhang, P. Tiňo, A. Leonardis, K. Tang, A survey on neural network interpretability, 2020, arXiv preprint arXiv:2012.14261.

[24] A. Chatzimparmpas, R.M. Martins, I. Jusufi, A. Kerren, A survey of surveys on the use of visualization for interpreting machine learning models, Inf. Vis. 19 (3) (2020) 207–233, http://dx.doi.org/10.1177/1473871620904671.

[25] C. Rudin, Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, Nat. Mach. Intell. 1 (5) (2019) 206–215.

[26] D.V. Carvalho, E.M. Pereira, J.S. Cardoso, Machine learning interpretability: A survey on methods and metrics, Electronics 8 (8) (2019) http://dx.doi.org/10.3390/electronics8080832.

[27] J. Zhou, A.H. Gandomi, F. Chen, A. Holzinger, Evaluating the quality of machine learning explanations: A survey on methods and metrics, Electronics 10 (5) (2021) 593.

[28] S. Keele, et al., Guidelines for Performing Systematic Literature Reviews in Software Engineering, Tech. rep., Citeseer, 2007.

[29] S. Salehi, A. Selamat, H. Fujita, Systematic mapping study on granular computing, Knowl.-Based Syst. 80 (2015) 78–97.

[30] G. Murtaza, L. Shuib, A.W. Abdul Wahab, G. Mujtaba, H.F. Nweke, M.A. Al-garadi, F. Zulfiqar, G. Raza, N.A. Azmi, Deep learning-based breast cancer classification through medical imaging modalities: state of the art and research challenges, Artif. Intell. Rev. 53 (3) (2020) 1655–1720.
[31] A. Qazi, H. Fayaz, A. Wadi, R.G. Raj, N. Rahim, W.A. Khan, The artificial neural network for solar radiation prediction and designing solar systems: a systematic literature review, J. Clean. Prod. 104 (2015) 1–12.
[32] M. Ahmad, A. Teredesai, C. Eckert, Interpretable machine learning in healthcare, in: 2018 IEEE International Conference on Healthcare Informatics, ICHI, IEEE Computer Society, Los Alamitos, CA, USA, 2018, p. 447, http://dx.doi.org/10.1109/ICHI.2018.00095.
[33] M. Fox, D. Long, D. Magazzeni, Explainable planning, 2017, arXiv preprint arXiv:1709.10256.
[34] D. Gunning, M. Stefik, J. Choi, T. Miller, S. Stumpf, G.-Z. Yang, XAI—Explainable artificial intelligence, Science Robotics 4 (37) (2019).
[35] A. Preece, Asking 'Why' in AI: Explainability of intelligent systems – perspectives and challenges, in: Intelligent Systems in Accounting, Finance and Management, Vol. 25, No. 2, 2018, pp. 63–72, http://dx.doi.org/10.1002/isaf.1422.
[36] G. Ras, M. van Gerven, P. Haselager, Explanation methods in deep learning: Users, values, concerns and challenges, in: Explainable and Interpretable Models in Computer Vision and Machine Learning, Springer International Publishing, Cham, 2018, pp. 19–36, http://dx.doi.org/10.1007/978-3-319-98131-4_2.
[37] G. Schwalbe, B. Finzel, A comprehensive taxonomy for explainable artificial intelligence: A systematic survey of surveys on methods and concepts, 2021, arXiv e-prints arXiv–2105.
[38] F.-L. Fan, J. Xiong, M. Li, G. Wang, On interpretability of artificial neural networks: A survey, IEEE Trans. Radiat. Plasma Med. Sci. (2021) http://dx.doi.org/10.1109/TRPMS.2021.3066428.
[39] F. He, Y. Wang, X. Miao, X. Sun, Interpretable visual reasoning: A survey, Image Vis. Comput. 112 (2021) 104194, http://dx.doi.org/10.1016/j.imavis.2021.104194.
[40] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, A. Galstyan, A survey on bias and fairness in machine learning, ACM Comput. Surv. 54 (6) (2021) http://dx.doi.org/10.1145/3457607.
[41] F.K. Došilović, M. Brčić, N. Hlupić, Explainable artificial intelligence: A survey, in: 2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics, MIPRO, 2018, pp. 0210–0215, http://dx.doi.org/10.23919/MIPRO.2018.8400040.
[42] C. Molnar, G. Casalicchio, B. Bischl, Interpretable machine learning – A brief history, state-of-the-art and challenges, in: ECML PKDD 2020 Workshops, Springer International Publishing, Cham, 2020, pp. 417–431.
[43] M. Reyes, R. Meier, S. Pereira, C.A. Silva, F.-M. Dahlweid, H.v. Tengg-Kobligk, R.M. Summers, R. Wiest, On the interpretability of artificial intelligence in radiology: Challenges and opportunities, Radiol. Artif. Intell. 2 (3) (2020) e190043, http://dx.doi.org/10.1148/ryai.2020190043, PMID: 32510054.
[44] L. Longo, R. Goebel, F. Lecue, P. Kieseberg, A. Holzinger, Explainable artificial intelligence: Concepts, applications, research challenges and visions, in: Machine Learning and Knowledge Extraction, Springer International Publishing, Cham, 2020, pp. 1–16.
[45] M. Pocevičiūtė, G. Eilertsen, C. Lundström, Survey of XAI in digital pathology, in: Artificial Intelligence and Machine Learning for Digital Pathology: State-of-the-Art and Future Challenges, Springer International Publishing, Cham, 2020, pp. 56–88, http://dx.doi.org/10.1007/978-3-030-50402-1_4.
[46] J.-X. Mi, A.-D. Li, L.-F. Zhou, Review study of interpretation methods for future interpretable machine learning, IEEE Access 8 (2020) 191969–191985, http://dx.doi.org/10.1109/ACCESS.2020.3032756.
[47] X.-H. Li, C.C. Cao, Y. Shi, W. Bai, H. Gao, L. Qiu, C. Wang, Y. Gao, S. Zhang, X. Xue, L. Chen, A survey of data-driven and knowledge-aware explainable AI, IEEE Trans. Knowl. Data Eng. (2020) http://dx.doi.org/10.1109/TKDE.2020.2983930.
[48] I. Nunes, D. Jannach, A systematic review and taxonomy of explanations in decision support and recommender systems, User Model. User Adapt. Interact. 27 (3) (2017) 393–444.
[49] A. Seeliger, M. Pfaff, H. Krcmar, Semantic web technologies for explainable machine learning models: A literature review, in: PROFILES/SEMEX@ ISWC, Vol. 2465, 2019, pp. 1–16.
[50] G. Ras, N. Xie, M. van Gerven, D. Doran, Explainable deep learning: A field guide for the uninitiated, J. Artif. Int. Res. 73 (2022) http://dx.doi.org/10.1613/jair.1.13200.
[51] V. Buhrmester, D. Münch, M. Arens, Analysis of explainers of black box deep neural networks for computer vision: A survey, Mach. Learn. Knowl. Extr. 3 (4) (2021) 966–989.
[52] P. Messina, P. Pino, D. Parra, A. Soto, C. Besa, S. Uribe, M. Andía, C. Tejos, C. Prieto, D. Capurro, A survey on deep learning and explainability for automatic report generation from medical images, ACM Comput. Surv. 54 (10s) (2022) http://dx.doi.org/10.1145/3522747.
[53] W. Samek, G. Montavon, S. Lapuschkin, C.J. Anders, K.-R. Müller, Explaining deep neural networks and beyond: A review of methods and applications, Proc. IEEE 109 (3) (2021) 247–278, http://dx.doi.org/10.1109/JPROC.2021.3060483.
[54] M.A. Gulum, C.M. Trombley, M. Kantardzic, A review of explainable deep learning cancer detection models in medical imaging, Appl. Sci. 11 (10) (2021) http://dx.doi.org/10.3390/app11104573.
[55] A.M. Antoniadi, Y. Du, Y. Guendouz, L. Wei, C. Mazo, B.A. Becker, C. Mooney, Current challenges and future opportunities for XAI in machine learning-based clinical decision support systems: a systematic review, Appl. Sci. 11 (11) (2021) 5088.
[56] A. Madsen, S. Reddy, S. Chandar, Post-hoc interpretability for neural NLP: A survey, ACM Comput. Surv. (2022) http://dx.doi.org/10.1145/3546577.
[57] S. Wallkötter, S. Tulli, G. Castellano, A. Paiva, M. Chetouani, Explainable embodied agents through social cues: A review, J. Hum.-Robot Interact. 10 (3) (2021) http://dx.doi.org/10.1145/3457188.
[58] A. Rawal, J. Mccoy, D.B. Rawat, B. Sadler, R. Amant, Recent advances in trustworthy explainable artificial intelligence: Status, challenges and perspectives, IEEE Trans. Artif. Intell. 1 (01) (2021) 1.
[59] A. Lucieri, M.N. Bajwa, A. Dengel, S. Ahmed, Achievements and challenges in explaining deep learning based computer-aided diagnosis systems, 2020, arXiv preprint arXiv:2011.13169.
[60] S.T. Mueller, R.R. Hoffman, W. Clancey, A. Emrey, G. Klein, Explanation in human-AI systems: A literature meta-review, synopsis of key ideas and publications, and bibliography for explainable AI, 2019, arXiv preprint arXiv:1902.01876.
[61] S.R. Islam, W. Eberle, S.K. Ghafoor, M. Ahmed, Explainable artificial intelligence approaches: A survey, 2021, arXiv preprint arXiv:2101.09429.
[62] M. Naiseh, N. Jiang, J. Ma, R. Ali, Personalising explainable recommendations: Literature and conceptualisation, in: A. Rocha, H. Adeli, L.P. Reis, S. Costanzo, I. Orovic, F. Moreira (Eds.), Trends and Innovations in Information Systems and Technologies, Springer International Publishing, Cham, 2020, pp. 518–533.
[63] A. Kotriwala, B. Klöpper, M. Dix, G. Gopalakrishnan, D. Ziobro, A. Potschka, XAI for operations in the process industry-applications, theses, and research directions, in: AAAI Spring Symposium: Combining Machine Learning with Knowledge Engineering, 2021.
[64] K. Cheng, N. Wang, M. Li, Interpretability of deep learning: A survey, in: The International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery, Springer, 2020, pp. 475–486.
[65] V. Belle, I. Papantonis, Principles and practice of explainable machine learning, Front. Big Data (2021) 39.
[66] S. Atakishiyev, M. Salameh, H. Yao, R. Goebel, Explainable artificial intelligence for autonomous driving: A comprehensive overview and field guide for future research directions, 2021, arXiv preprint arXiv:2112.11561.
[67] T. Miller, Explanation in artificial intelligence: Insights from the social sciences, Artificial Intelligence 267 (2019) 1–38, http://dx.doi.org/10.1016/j.artint.2018.07.007.
[68] B. Kovalerchuk, M.A. Ahmad, A. Teredesai, Survey of explainable machine learning with visual and granular methods beyond quasi-explanations, in: Interpretable Artificial Intelligence: A Perspective of Granular Computing, Springer International Publishing, Cham, 2021, pp. 217–267, http://dx.doi.org/10.1007/978-3-030-64949-4_8.
[69] A. Abdul, J. Vermeulen, D. Wang, B.Y. Lim, M. Kankanhalli, Trends and trajectories for explainable, accountable and intelligible systems: An HCI research agenda, in: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, Association for Computing Machinery, New York, NY, USA, 2018, pp. 1–18.
[70] G. Vilone, L. Longo, Explainable artificial intelligence: a systematic review, 2020, arXiv preprint arXiv:2006.00093.
[71] M. Danilevsky, K. Qian, R. Aharonov, Y. Katsis, B. Kawas, P. Sen, A survey of the state of explainable AI for natural language processing, in: Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing, 2020, pp. 447–459.
[72] H. Yuan, H. Yu, S. Gui, S. Ji, Explainability in graph neural networks: A taxonomic survey, IEEE Trans. Pattern Anal. Mach. Intell. (2022) 1–19, http://dx.doi.org/10.1109/TPAMI.2022.3204236.
[73] T. Rojat, R. Puget, D. Filliat, J. Del Ser, R. Gelin, N. Díaz-Rodríguez, Explainable artificial intelligence (xai) on timeseries data: A survey, 2021, arXiv preprint arXiv:2104.00950.
[74] R. Moraffah, M. Karami, R. Guo, A. Raglin, H. Liu, Causal interpretability for machine learning - problems, methods and evaluation, SIGKDD Explor. Newsl. 22 (1) (2020) 18–33, http://dx.doi.org/10.1145/3400051.3400058.
[75] R. Dazeley, P. Vamplew, F. Cruz, Explainable reinforcement learning for broad-xai: A conceptual framework and survey, 2021, arXiv preprint arXiv:2108.09003.
[76] I. Stepin, J.M. Alonso, A. Catala, M. Pereira-Fariña, A survey of contrastive and counterfactual explanation generation methods for explainable artificial intelligence, IEEE Access 9 (2021) 11974–12001, http://dx.doi.org/10.1109/ACCESS.2021.3051315.

[77] Y. Lu, R. Garcia, B. Hansen, M. Gleicher, R. Maciejewski, The state-of-the-art in predictive visual analytics, in: Computer Graphics Forum, Vol. 36, No. 3, Wiley Online Library, 2017, pp. 539–562.
[78] Explainable AI: The Basics, The Royal Society, 2019, URL https://royalsociety.org/-/media/policy/projects/explainable-ai/AI-and-interpretability-policy-briefing.pdf.
[79] D. Slack, S. Hilgard, E. Jia, S. Singh, H. Lakkaraju, Fooling lime and shap: Adversarial attacks on post hoc explanation methods, in: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, 2020, pp. 180–186.
[80] D. Saraswat, P. Bhattacharya, A. Verma, V.K. Prasad, S. Tanwar, G. Sharma, P.N. Bokoro, R. Sharma, Explainable AI for healthcare 5.0: Opportunities and challenges, IEEE Access (2022).
[81] R. Tomsett, A. Preece, D. Braines, F. Cerutti, S. Chakraborty, M. Srivastava, G. Pearson, L. Kaplan, Rapid trust calibration through interpretable and uncertainty-aware AI, Patterns 1 (4) (2020) 100049.
[82] A. Deeks, The judicial demand for explainable artificial intelligence, Columbia Law Rev. 119 (7) (2019) 1829–1850.
[83] D.H. Park, L.A. Hendricks, Z. Akata, A. Rohrbach, B. Schiele, T. Darrell, M. Rohrbach, Multimodal explanations: Justifying decisions and pointing to the evidence, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2018.
[84] S. Chen, Q. Zhao, REX: Reasoning-aware and grounded explanation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 15586–15595.
[85] D. Rajapaksha, C. Bergmeir, R.J. Hyndman, LoMEF: A framework to produce local explanations for global model time series forecasts, Int. J. Forecast. (2022) http://dx.doi.org/10.1016/j.ijforecast.2022.06.006.
[86] Y. Xie, S. Katariya, X. Tang, E. Huang, N. Rao, K. Subbian, S. Ji, Task-agnostic graph explanations, 2022, arXiv preprint arXiv:2202.08335.
[87] A. Dikshit, B. Pradhan, Interpretable and explainable AI (XAI) model for spatial drought prediction, Sci. Total Environ. 801 (2021) 149797, http://dx.doi.org/10.1016/j.scitotenv.2021.149797.
[88] M.T. Ribeiro, S. Singh, C. Guestrin, " Why should i trust you?" explaining the predictions of any classifier, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 1135–1144.
[89] E. Štrumbelj, I. Kononenko, Explaining prediction models and individual predictions with feature contributions, Knowl. Inf. Syst. 41 (3) (2014) 647–665.
[90] W.J. Murdoch, C. Singh, K. Kumbier, R. Abbasi-Asl, B. Yu, Definitions, methods, and applications in interpretable machine learning, Proc. Natl. Acad. Sci. 116 (44) (2019) 22071–22080, http://dx.doi.org/10.1073/pnas.1900654116.
[91] B. Williamson, J. Feng, Efficient nonparametric statistical inference on population feature importance using Shapley values, in: International Conference on Machine Learning, PMLR, 2020, pp. 10282–10291.
[92] C. Bénard, G. Biau, S. Da Veiga, E. Scornet, SHAFF: Fast and consistent shapley effect estimates via random forests, in: International Conference on Artificial Intelligence and Statistics, PMLR, 2022, pp. 5563–5582.
[93] E. Reiter, Natural language generation challenges for explainable AI, in: Proceedings of the 1st Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence, NL4XAI 2019, Association for Computational Linguistics, 2019, pp. 3–7, http://dx.doi.org/10.18653/v1/W19-8402.
[94] A.B. Sai, A.K. Mohankumar, M.M. Khapra, A survey of evaluation metrics used for NLG systems, ACM Comput. Surv. 55 (2) (2022) 1–39.
[95] K. Van Deemter, Not Exactly: In Praise of Vagueness, Oxford University Press, 2012.
[96] K. Daniel, Thinking, fast and slow, 2017.
[97] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in Neural Information Processing Systems, 2017, pp. 5998–6008.
[98] J.H. Friedman, Greedy function approximation: a gradient boosting machine, Ann. Statist. (2001) 1189–1232.
[99] F. Doshi-Velez, M. Kortz, R. Budish, C. Bavitz, S. Gershman, D. O'Brien, K. Scott, S. Schieber, J. Waldo, D. Weinberger, et al., Accountability of AI under the law: The role of explanation, 2017, arXiv preprint arXiv:1711.01134.
[100] V. Beaudouin, I. Bloch, D. Bounie, S. Clémençon, F. d'Alché Buc, J. Eagan, W. Maxwell, P. Mozharovskyi, J. Parekh, Flexible and context-specific AI explainability: a multidisciplinary approach, 2020, Available At SSRN 3559477.
[101] G. Dao, M. Lee, Demystifying deep neural networks through interpretation: A survey, 2020, arXiv preprint arXiv:2012.07119.
[102] J. Choo, S. Liu, Visual analytics for explainable deep learning, IEEE Comput. Graph. Appl. 38 (04) (2018) 84–92, http://dx.doi.org/10.1109/MCG.2018.042731661.
[103] Q. Zhang, S.-C. Zhu, Visual interpretability for deep learning: a survey, Front. Inf. Technol. Electron. Eng. 19 (2018) 27–39, http://dx.doi.org/10.1631/FITEE.1700808.
[104] C. He, M. Ma, P. Wang, Extract interpretability-accuracy balanced rules from artificial neural networks: A review, Neurocomputing 387 (2020) 346–358, http://dx.doi.org/10.1016/j.neucom.2020.01.036.
[105] Y. Liang, S. Li, C. Yan, M. Li, C. Jiang, Explaining the black-box model: A survey of local interpretation methods for deep neural networks, Neurocomputing 419 (2021) 168–182, http://dx.doi.org/10.1016/j.neucom.2020.08.011.
[106] M. Naiseh, N. Jiang, J. Ma, R. Ali, Explainable recommendations in intelligent systems: Delivery methods, modalities and risks, in: Research Challenges in Information Science, Springer International Publishing, Cham, 2020, pp. 212–228.
[107] J.M. Darias, B. Díaz-Agudo, J.A. Recio-García, A systematic review on model-agnostic XAI libraries, in: ICCBR Workshops, 2021, pp. 28–39.
[108] G. Joshi, R. Walambe, K. Kotecha, A review on explainability in multimodal deep neural nets, IEEE Access 9 (2021) 59800–59821, http://dx.doi.org/10.1109/ACCESS.2021.3070212.
[109] E. Tjoa, C. Guan, A survey on explainable artificial intelligence (XAI): Toward medical XAI, IEEE Trans. Neural Netw. Learn. Syst. (2020) 1–21, http://dx.doi.org/10.1109/TNNLS.2020.3027314.
[110] L. Wells, T. Bednarz, Explainable AI and reinforcement learning—A systematic review of current approaches and trends, Front. Artif. Intell. 4 (2021) 48, http://dx.doi.org/10.3389/frai.2021.550030.
[111] X. Huang, D. Kroening, W. Ruan, J. Sharp, Y. Sun, E. Thamo, M. Wu, X. Yi, A survey of safety and trustworthiness of deep neural networks: Verification, testing, adversarial attack and defence, and interpretability, Comp. Sci. Rev. 37 (2020) 100270, http://dx.doi.org/10.1016/j.cosrev.2020.100270.
[112] A. Weller, Transparency: Motivations and challenges, in: Explainable AI: Interpreting, Explaining and Visualizing Deep Learning, Springer International Publishing, Cham, 2019, pp. 23–40, http://dx.doi.org/10.1007/978-3-030-28954-6_2.
[113] A. Holzinger, P. Kieseberg, E. Weippl, A.M. Tjoa, Current advances, trends and challenges of machine learning and knowledge extraction: From machine learning to explainable AI, in: A. Holzinger, P. Kieseberg, A.M. Tjoa, E. Weippl (Eds.), Machine Learning and Knowledge Extraction, Springer International Publishing, Cham, 2018, pp. 1–8.
[114] Y. Zhang, X. Chen, Explainable recommendation: A survey and new perspectives, Found. Trends® Inform. Retr. 14 (1) (2020) 1–101, http://dx.doi.org/10.1561/1500000066.
[115] S. Anjomshoae, A. Najjar, D. Calvaresi, K. Främling, Explainable agents and robots: Results from a systematic literature review, in: Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems, AAMAS '19, International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 2019, pp. 1078–1088.
[116] M.A. Ahmad, C. Eckert, A. Teredesai, The challenge of imputation in explainable artificial intelligence models, in: Proceedings of the Workshop on Artificial Intelligence Safety, 2019, URL http://ceur-ws.org/Vol-2419/paper_26.pdf.
[117] A. Black, P. Nederpelt, Dimensions of Data Quality (DDQ), DAMA NL Foundation, 2020, URL http://www.dama-nl.org/wp-content/uploads/2020/09/DDQ-Dimensions-of-Data-Quality-Research-Paper-version-1.2-d.d.-3-Sept-2020.pdf.
[118] X. Xie, J. Niu, X. Liu, Z. Chen, S. Tang, S. Yu, A survey on incorporating domain knowledge into deep learning for medical image analysis, Med. Image Anal. 69 (2021) 101985, http://dx.doi.org/10.1016/j.media.2021.101985.
[119] L. Fe-Fei, Fergus, Perona, A Bayesian approach to unsupervised one-shot learning of object categories, in: Proceedings Ninth IEEE International Conference on Computer Vision, Vol. 2, IEEE, 2003, pp. 1134–1141, http://dx.doi.org/10.1109/ICCV.2003.1238476.
[120] M. Gaur, K. Faldu, A. Sheth, Semantics of the black-box: Can knowledge graphs help make deep learning systems more interpretable and explainable? IEEE Internet Comput. 25 (1) (2021) 51–59, http://dx.doi.org/10.1109/MIC.2020.3031769.
[121] G.G. Towell, J.W. Shavlik, Extracting refined rules from knowledge-based neural networks, Mach. Learn. 13 (1) (1993) 71–101.
[122] C.W. Omlin, C. Giles, Extraction of rules from discrete-time recurrent neural networks, Neural Netw. 9 (1) (1996) 41–52, http://dx.doi.org/10.1016/0893-6080(95)00086-0.
[123] R. Andrews, J. Diederich, A.B. Tickle, Survey and critique of techniques for extracting rules from trained artificial neural networks, Knowl.-Based Syst. 8 (6) (1995) 373–389, http://dx.doi.org/10.1016/0950-7051(96)81920-4, Knowledge-based neural networks.
[124] J.R. Zilke, E. Loza Mencía, F. Janssen, DeepRED – rule extraction from deep neural networks, in: Discovery Science, Springer International Publishing, Cham, 2016, pp. 457–473.
[125] N. Pezzotti, T. Höllt, J. Van Gemert, B.P. Lelieveldt, E. Eisemann, A. Vilanova, DeepEyes: Progressive visual analytics for designing deep neural networks, IEEE Trans. Vis. Comput. Graphics 24 (1) (2018) 98–108, http://dx.doi.org/10.1109/TVCG.2017.2744358.

[126] M. Liu, J. Shi, K. Cao, J. Zhu, S. Liu, Analyzing the training processes of deep generative models, IEEE Trans. Vis. Comput. Graphics 24 (1) (2018) 77–87, http://dx.doi.org/10.1109/TVCG.2017.2744938.
[127] B.H. van der Velden, H.J. Kuijf, K.G. Gilhuijs, M.A. Viergever, Explainable artificial intelligence (XAI) in deep learning-based medical image analysis, Med. Image Anal. 79 (2022) 102470, http://dx.doi.org/10.1016/j.media.2022.102470.
[128] H. Jiang, K. Yang, M. Gao, D. Zhang, H. Ma, W. Qian, An interpretable ensemble deep learning model for diabetic retinopathy disease classification, in: 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBC, 2019, pp. 2045–2048, http://dx.doi.org/10.1109/EMBC.2019.8857160.
[129] H. Lee, S. Yune, M. Mansouri, M. Kim, S.H. Tajmir, C.E. Guerrier, S.A. Ebert, S.R. Pomerantz, J.M. Romero, S. Kamalian, et al., An explainable deep-learning algorithm for the detection of acute intracranial haemorrhage from small datasets, Nat. Biomed. Eng. 3 (3) (2019) 173–182.
[130] P. Hall, N. Gill, N. Schmidt, Proposed guidelines for the responsible use of explainable machine learning, 2019, arXiv preprint arXiv:1906.03533.
[131] M.D. Zeiler, R. Fergus, Visualizing and understanding convolutional networks, in: D. Fleet, T. Pajdla, B. Schiele, T. Tuytelaars (Eds.), Computer Vision – ECCV 2014, Springer International Publishing, Cham, 2014, pp. 818–833.
[132] Y. Li, T. Fujiwara, Y.K. Choi, K.K. Kim, K.-L. Ma, A visual analytics system for multi-model comparison on clinical data predictions, Vis. Inform. 4 (2) (2020) 122–131.
[133] D.L. Arendt, N. Nur, Z. Huang, G. Fair, W. Dou, Parallel embeddings: a visualization technique for contrasting learned representations, in: Proceedings of the 25th International Conference on Intelligent User Interfaces, 2020, pp. 259–274.
[134] X. Xuan, X. Zhang, O.-H. Kwon, K.-L. Ma, VAC-CNN: A visual analytics system for comparative studies of deep convolutional neural networks, IEEE Trans. Vis. Comput. Graphics 28 (6) (2022) 2326–2337.
[135] H. Wang, D.-Y. Yeung, Towards Bayesian deep learning: A framework and some existing methods, IEEE Trans. Knowl. Data Eng. 28 (12) (2016) 3395–3408, http://dx.doi.org/10.1109/TKDE.2016.2606428.
[136] L. Yuan, X. Gao, Z. Zheng, M. Edmonds, Y.N. Wu, F. Rossano, H. Lu, Y. Zhu, S.-C. Zhu, In situ bidirectional human-robot value alignment, Science Robotics 7 (68) (2022) eabm4183, http://dx.doi.org/10.1126/scirobotics.abm4183.
[137] T. Orekondy, B. Schiele, M. Fritz, Knockoff nets: Stealing functionality of black-box models, in: Conference on Computer Vision and Pattern Recognition, 2019.
[138] S.J. Oh, M. Augustin, B. Schiele, M. Fritz, Towards reverse-engineering black-box neural networks, in: International Conference on Learning Representations, 2018.
[139] L. Huang, A.D. Joseph, B. Nelson, B.I. Rubinstein, J.D. Tygar, Adversarial machine learning, in: Proceedings of the 4th ACM Workshop on Security and Artificial Intelligence, AISec '11, Association for Computing Machinery, New York, NY, USA, 2011, pp. 43–58, http://dx.doi.org/10.1145/2046684.2046692.
[140] C.F. Baumgartner, L.M. Koch, K.C. Tezcan, J.X. Ang, E. Konukoglu, Visual feature attribution using wasserstein gans, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 8309–8319.
[141] S. Liu, B. Kailkhura, D. Loveland, Y. Han, Generative counterfactual introspection for explainable deep learning, in: 2019 IEEE Global Conference on Signal and Information Processing, GlobalSIP, 2019, pp. 1–5, http://dx.doi.org/10.1109/GlobalSIP45357.2019.8969491.
[142] A.W. Senior, R. Evans, J. Jumper, J. Kirkpatrick, L. Sifre, T. Green, C. Qin, A. Žídek, A.W. Nelson, A. Bridgland, et al., Improved protein structure prediction using potentials from deep learning, Nature 577 (7792) (2020) 706–710.
[143] S. Krening, B. Harrison, K.M. Feigh, C.L. Isbell, M. Riedl, A. Thomaz, Learning from explanations using sentiment and advice in RL, IEEE Trans. Cogn. Dev. Syst. 9 (1) (2017) 44–55, http://dx.doi.org/10.1109/TCDS.2016.2628365.
[144] T.D. Grant, D.J. Wischik, Show us the data: Privacy, explainability, and why the law can't have both, Geo. Wash. L. Rev. 88 (2020) 1350.
[145] E.F. Villaronga, P. Kieseberg, T. Li, Humans forget, machines remember: Artificial intelligence and the right to be forgotten, Comput. Law Secur. Rev. 34 (2) (2018) 304–313, http://dx.doi.org/10.1016/j.clsr.2017.08.007.
[146] B. McMahan, E. Moore, D. Ramage, S. Hampson, B.A. y Arcas, Communication-efficient learning of deep networks from decentralized data, in: Artificial Intelligence and Statistics, PMLR, 2017, pp. 1273–1282.
[147] O.A. Wahab, A. Mourad, H. Otrok, T. Taleb, Federated machine learning: Survey, multi-level classification, desirable criteria and future directions in communication and networking systems, IEEE Commun. Surv. Tutor. 23 (2) (2021) 1342–1397, http://dx.doi.org/10.1109/COMST.2021.3058573.
[148] J. Konečný, H. McMahan, F. Yu, P. Richtárik, A. Suresh, D. Bacon, Federated learning: Strategies for improving communication efficiency, 2016, arXiv preprint arXiv:1610.05492.
[149] J. Hoffmann, D. Magazzeni, Explainable AI planning (XAIP): Overview and the case of contrastive explanation (extended abstract), in: Reasoning Web. Explainable Artificial Intelligence: 15th International Summer School 2019, Bolzano, Italy, September 20–24, 2019, Tutorial Lectures, Springer International Publishing, Cham, 2019, pp. 277–282, http://dx.doi.org/10.1007/978-3-030-31423-1_9.
[150] P. Langley, B. Meadows, M. Sridharan, D. Choi, Explainable agency for intelligent autonomous systems, in: Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, AAAI '17, AAAI Press, 2017, pp. 4762–4763.
[151] M.A. Neerincx, J. van der Waa, F. Kaptein, J. van Diggelen, Using perceptual and cognitive explanations for enhanced human-agent team performance, in: D. Harris (Ed.), Engineering Psychology and Cognitive Ergonomics, Springer International Publishing, Cham, 2018, pp. 204–214.
[152] E. LeDell, S. Poirier, H2O autoML: Scalable automatic machine learning, in: 7th ICML Workshop on Automated Machine Learning, AutoML, 2020, URL https://www.automl.org/wp-content/uploads/2020/07/AutoML_2020_paper_61.pdf.
[153] A. Płońska, P. Płoński, MLJAR: State-of-the-art Automated Machine Learning Framework for Tabular Data. Version 0.10.3, MLJAR, Łapy, Poland, 2021, URL https://github.com/mljar/mljar-supervised.
[154] T.R. Gruber, A translation approach to portable ontology specifications, Knowl. Acquis. 5 (2) (1993) 199–220, http://dx.doi.org/10.1006/knac.1993.1008.
[155] C. Panigutti, A. Perotti, D. Pedreschi, Doctor XAI: An ontology-based approach to black-box sequential data classification explanations, in: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, in: FAT* '20, Association for Computing Machinery, New York, NY, USA, 2020, pp. 629–639, http://dx.doi.org/10.1145/3351095.3372855.
[156] R. Confalonieri, T. Weyde, T.R. Besold, F. Moscoso del Prado Martín, Using ontologies to enhance human understandability of global post-hoc explanations of black-box models, Artificial Intelligence 296 (2021) 103471, http://dx.doi.org/10.1016/j.artint.2021.103471.
[157] E. Choi, M.T. Bahadori, A. Schuetz, W.F. Stewart, J. Sun, Doctor AI: Predicting clinical events via recurrent neural networks, in: Proceedings of the 1st Machine Learning for Healthcare Conference, in: Proceedings of Machine Learning Research, vol. 56, PMLR, Northeastern University, Boston, MA, USA, 2016, pp. 301–318.
[158] T. Tudorache, Ontology engineering: Current state, challenges, and future directions, Semant. Web 11 (1) (2020) 125–138.