Information Fusion 106 (2024) 102301

Contents lists available at ScienceDirect

Information Fusion

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

# Explainable Artificial Intelligence (XAI) 2.0: A manifesto of open challenges and interdisciplinary research directions

Luca Longo [1,2,*], Mario Brcic [3], Federico Cabitza [4,5], Jaesik Choi [6,7], Roberto Confalonieri [8], Javier Del Ser [9,10,11], Riccardo Guidotti [12], Yoichi Hayashi [13], Francisco Herrera [11], Andreas Holzinger [14], Richard Jiang [15], Hassan Khosravi [16], Freddy Lecue [17], Gianclaudio Malgieri [18], Andrés Páez [19,20], Wojciech Samek [21,22,23], Johannes Schneider [24], Timo Speith [25,26], Simone Stumpf [27]

[1] *The Artificial Intelligence and Cognitive Load Research Lab, Ireland*
[2] *School of Computer Science, Technological University Dublin, Republic of Ireland*
[3] *University of Zagreb, Faculty of Electrical Engineering and Computing, Croatia*
[4] *University of Milano-Bicocca, Milan, Italy*
[5] *IRCCS Ospedale Galeazzi Sant'Ambrogio, Milan, Italy*
[6] *Kim Jaechul Graduate School of AI, Korea Advanced Institute of Science and Technology (KAIST), Republic of Korea*
[7] *INEEJI Corporation, Republic of Korea*
[8] *Department of Mathematics, University of Padua, Italy*
[9] *TECNALIA, Basque Research & Technology Alliance (BRTA), Derio, Spain*
[10] *University of the Basque Country (UPV/EHU), Bilbao, Spain*
[11] *Department of Computer Science and Artificial Intelligence, DaSCI Andalusian Institute in Data Science and Computational Intelligence, University of Granada, Granada, Spain*
[12] *University of Pisa, Pisa, Italy*
[13] *Department of Computer Science, Meiji University, Tokyo, Japan*
[14] *Human-Centered AI Lab, University of Natural Resources and Life Sciences Vienna, Austria*
[15] *School of Computing and Communications, Lancaster University, UK*
[16] *The University of Queensland, Brisbane, Australia*
[17] *National Institute for Research in Digital Science and Technology (INRIA), Sophia Antipolis, France*
[18] *eLaw Center for Law and Digital Technologies, Leiden University, Netherlands*
[19] *Department of Philosophy, Universidad de los Andes, Bogotá, Colombia*
[20] *Center for Research and Formation in Artificial Intelligence (CinfonIA), Universidad de los Andes, Bogotá, Colombia*
[21] *Technical University of Berlin, Berlin, Germany*
[22] *Fraunhofer Heinrich Hertz Institute, Berlin, Germany*
[23] *Berlin Institute for the Foundations of Learning and Data (BIFOLD), Berlin, Germany*
[24] *Department of Information Systems and Computer Science, University of Liechtenstein, Liechtenstein*
[25] *Department of Philosophy, University of Bayreuth, Bayreuth, Germany*
[26] *Center for Perspicuous Computing, Saarland University, Saarbrücken, Germany*
[27] *School of Computing Science, University of Glasgow, United Kingdom*

## ARTICLE INFO

*Keywords:*
Explainable artificial intelligence
XAI
Interpretability
Manifesto
Open challenges
Interdisciplinarity
Ethical AI
Large language models
Trustworthy AI
Responsible AI

## ABSTRACT

Understanding black box models has become paramount as systems based on opaque Artificial Intelligence (AI) continue to flourish in diverse real-world applications. In response, Explainable AI (XAI) has emerged as a field of research with practical and ethical benefits across various domains. This paper highlights the advancements in XAI and its application in real-world scenarios and addresses the ongoing challenges within XAI, emphasizing the need for broader perspectives and collaborative efforts. We bring together experts from diverse fields to identify open problems, striving to synchronize research agendas and accelerate XAI in practical applications. By fostering collaborative discussion and interdisciplinary cooperation, we aim to propel XAI forward, contributing to its continued success. We aim to develop a comprehensive proposal for advancing XAI. To achieve this goal, we present a manifesto of 28 open problems categorized into nine categories. These

* Corresponding author at: School of Computer Science, Technological University Dublin, Republic of Ireland.
*E-mail address:* luca.longo@tudublin.ie (L. Longo).

https://doi.org/10.1016/j.inffus.2024.102301
Received 2 November 2023; Received in revised form 22 January 2024; Accepted 12 February 2024
Available online 15 February 2024

Generative AI
Multi-faceted explanations
Concept-based explanations
Causality
Actionable XAI
Falsifiability

challenges encapsulate the complexities and nuances of XAI and offer a road map for future research. For each problem, we provide promising research directions in the hope of harnessing the collective intelligence of interested stakeholders.

## 1. Introduction

The field of Explainable AI (XAI) has grown significantly over the past few years. It has evolved from being a niche research topic within the larger field of Artificial Intelligence (AI) [1–3] to becoming a highly active field of research, with a large number of theoretical contributions, empirical studies, and reviews being proposed every year [4,5]. Furthermore, XAI has evolved into an exceedingly multidisciplinary, interdisciplinary, and transdisciplinary field. Among others, XAI is now a research topic in a broad range of disciplines outside of computer science, such as engineering, chemistry, biology, education, psychology, neuroscience, and philosophy among others [6–8]. The growth of XAI can be attributed to the increasing proliferation of AI. In recent years, AI has been successfully used in many real-world applications due to its ability to learn and automatically extract patterns from complex and non-linear data. In particular, Machine Learning (ML) and Deep Learning (DL) techniques have been used for classification, forecasting, prediction, recommendation, and data generation. The success of these techniques and their application in critical areas, such as finance [9] and healthcare [10], among others, has made it necessary to understand these models' underlying mechanisms and their often opaque outputs. XAI has emerged as a response to this demand, as it seeks to develop methods for explaining AI systems and their outputs. In other words, increasing use of AI-based systems, especially in critical areas, has made XAI an area of study with a significant practical and ethical value.

Despite much progress in XAI in the last years, many questions and problems require further analysis, reflection and exploration. For instance, explainability is considered one of the four ethical principles for trustworthy AI, together with respect for human autonomy, the prevention of harm, and fairness [11]. However, the exact connection between XAI and the other requirements for trustworthy AI is still not fully clarified, thereby hindering a proper assessment of the overall impact of XAI on the design of trustworthy AI systems. The work in [12] is a recent, notable example of this growing concern. With its striking title (*"Dear XAI Community, We Need to Talk!"*), insights offered in this work highlight and discuss several misconceptions in current XAI research. What has become widely acknowledged is that (i) XAI alone is not enough for trustworthiness and that (ii) there is a need to shed light on the connection of XAI with the other requirements of trustworthy AI. Recently, some studies have started to look into these matters. For instance, there is work that examines how explainability relates to trust and trustworthiness [13,14]. Similarly, authors have scrutinized the relationship between explainability and robustness (as a proxy for harm prevention) by unveiling the existence of a tight relationship between adversarial examples and explanations [15]. Finally, the literature can also discuss specific misconceptions and downsides related to XAI. An example is a simple argument on how Shapley values for explainability can produce misleading info regarding the relative importance of features [16].

As research on XAI matures, the community is starting to critically reflect on the path built so far and on the rationales that are given in the literature to motivate XAI. An early example of such a critical reflection, published with the title *"Explainable AI: Beware of Inmates Running the Asylum",* highlights how research on explainability often fails to incorporate insights from research disciplines outside of computer science [17]. Several contributions have discussed the relationship between explainability and transparency. These suggest that explanations should be tailored to the model's audience consuming its outputs, being audience a notion introduced in [18], including developers, designers, regulators or users from society. This notion was recently analysed together with three levels of transparency: algorithmic, interactive and social [19]. This has not only added an additional dimension to the complexity of XAI but also revitalized the need to adjust explanations to the audience and consider domain-specific knowledge in the process of crafting them. In a similar vein, [6] emphasizes the need for more interdisciplinary collaboration in XAI, stating that this is crucial to achieving practical explanations for all stakeholders involved with and affected by AI systems. This work also finds that there are too few empirical studies on the effectiveness of explanations. Another exciting reflection was posed in [20], where it is suggested that XAI often prioritizes comprehensibility over providing comprehensive and faithful explanations, resulting in explanations that might not align with how the AI system truly made a prediction. However, it is also argued that this trade-off is generally acceptable unless comprehensive explanations are urgently needed, thus advocating for "satisficing" explanations for automated decisions in most cases.

These studies and reflections highlight the maturity reached in XAI research, signifying that it is open to debates about its development and usefulness, but also making recognizable that open problems are often viewed through isolated perspectives and narrow viewpoints [4,6,17]. Therefore, the community urgently needs a global vision and a discussion about how to move forward in the future development of XAI. A broader, multidisciplinary approach that draws on the expertise of researchers across different fields can bring about advances towards a new paradigm in XAI research, which we coin here as XAI 2.0. This research manuscript addresses this need by bringing together a wide range of experts to collaboratively identify and tackle open problems in XAI research. Our deliberations and proposals aim to place XAI at the centre of the current debates on AI regulation, trustworthy AI, and auditability of AI systems, identifying directions that could catalyze XAI in real-world applications, thus enthroning it as a fundamental piece in AI governance. Considering this background, this work focuses on synchronizing the research agendas of scholars working in the field. The goal is to form a proposal open to discussion, a sincere attempt at fostering a debate around XAI and what research should address in the future. By doing so, we hope to offer new insights and perspectives on, for instance, developing and improving methods and applying existing methods in novel domains. Through this collaborative effort, we seek to advance the field of XAI and contribute to its continued growth and success. In particular, we seek to propose a *manifesto* that comprises several propositions governing scientific research in the field of XAI. This article has come about through a peculiar synthesis to achieve said goal. Various experts from different disciplines, including philosophy, psychology, HCI, and computer science, were brought together to get different perspectives on XAI. This significant effort has resulted in 28 problems with their challenges, which we have divided into nine categories of two to four problems.

Overall, the structure of this paper is as follows. Section 2 introduces basic definitions and briefly examines XAI literature reviews published to date to provide a brief report on recent work in this field. To highlight the benefits of XAI for people, businesses, institutions, and society, this section also presents a variety of advances of XAI techniques and methods, along with a selection of their applications in real-world settings. Subsequently, the article's core follows in Section 3 by describing 28 problems in XAI, the challenges associated with solving them, and our suggestions for possible solutions. Finally, Section 4 summarizes our manifesto, offering a roadmap for future research.

## 2. Concepts, advances and applications of XAI research

This section showcases that research in XAI is alive and functional. To demonstrate this fact, basic definitions and recent reviews on XAI are presented in Section 2.1. Subsequently, Section 2.2 focuses on synthesizing the main breakthroughs in XAI, demonstrating its enormous potential. Finally, the large and increasing number of applications of XAI methods, techniques, and tools and their utility in real-world scenarios is summarized in Section 2.3.

### *2.1. Basic definitions and existing reviews on XAI*

In the XAI debate, there is, unfortunately, no explicit agreement on the meaning of many of the terms used. A discussion of this as a category of problems is presented later, see Section 3.3. Nevertheless, the most important concepts are presented to have a common basis for this article, trying not to go beyond the established use of terms. In general, the goal of *explainability* is to make certain aspects of a system understandable for humans, being these developers, designers, regulators or users from general society [21–25]. In *XAI*, these aspects can be a single prediction of an AI model (*local* explainability) or the AI model as a whole (*global* explainability) [4]. Some scholars also include further aspects, such as the data used for training the model (see, for instance, [26]). However, only the distinction between local and global can be seen as consensual in the debate. Another important distinction is the one between directly training explainable models (*ante-hoc* explainability) and explaining a (plausibly opaque) model after it was trained (*post-hoc* explainability) [4]. *Ante-hoc* explainability is sometimes also called *(intrinsic) interpretability* or *transparent model design*. However, as the terms 'interpretability' and 'transparency' are used with differing meanings in the literature [19] (for instance, they are sometimes used as synonyms for 'comprehensibility' but sometimes also for other concepts), we will avoid these terms in what follows. A final distinction that is important to be made concerns only post-hoc explainability. Here, a distinction is made between methods that work independently of the underlying model (model-agnostic explainability methods) and those that only work for certain models or model classes (model-specific explainability methods) [4].

The above terms can be helpful in understanding the discussions throughout this paper. Expanded lists with further terms and concepts can be found in [4,5,18,27,28]. Furthermore, it is worth mentioning that explanations can help improve several desirable system and model properties, such as performance and robustness (for discussions, see [29,30]). To complement this short introduction of terms, Table 1 briefly describes a selection of reviews on XAI that summarize different contributions made over the years. While the table cannot be exhaustive due to the abundance of publications in XAI, it still aims to illustrate the different perspectives in the literature, from the theoretical development of XAI to the application of XAI methods to problems in different application domains.

### *2.2. XAI trends, advances, and breakthroughs*

The primary goal of explanations is to make a model understandable or comprehensible to its stakeholders. To this end, several methods have been introduced in the last few years to explain the decisions of complex AI systems in many application domains. These have been reviewed extensively over the years (see Table 1). Synthesizing explanations for AI systems has been shown to have the potential to solve several technical and societal problems. Explanations can facilitate understanding how learning from data has occurred, for instance, via feature attribution methods. Furthermore, explanations can reveal how a model can be exploited to improve its performance. They can also support and improve human confidence in the output of a given model. Explanations may reveal the existence of hidden biases in the training data, learned during model training, that negatively impact a model's generalization when predicting unseen data [51]. Other purposes for demanding explanations include data stream settings, where they can be used to characterize what a model observes over time. This can serve as a knowledge base to detect non-stationarities in the task being solved and, thus, concept deviations [52]. Similarly, wrongly annotated data instances in large-scale databases can be identified by computing a measure of disagreement between the explanations issued for a model. Application opportunities such as these may also arise in vertical federated learning, where aggregation policies can be adjusted by examining commonalities among local models during update rounds [53]. Explanations can also drive pruning and model compression strategies, linking irrelevant concepts to specific neurons that can hence be removed from a neural network [54].

#### *2.2.1. Attribution methods*

A lot of work exists on explaining the decisions of a classifier with *attribution methods* [44]. For instance, model agnostic attribution methods such as Local Interpretable Model-Agnostic Explanations (LIME) [55], Shapley Additive Explanation (SHAP) [56], and many others can contribute to the explanation of DL models by computing the importance of input features [57,58]. Furthermore, saliency maps built by attribution methods such as network gradients, Deconvolutional Neural Networks (DeConvNet), Layer-Wise Relevance Propagation (LRP), Pattern Attribution, and Randomized Input Sampling for Explanation (RISE) can identify relevant inputs for the decisions of classification or regression tasks. In the image or text domain, attribution explanations are intuitive and often perceived as easy to understand by the human receiver. For instance, one immediately understands that a classifier might not work correctly if it classifies horse images not by looking at the horse itself but by focusing on a copyright watermark, often present in this category's images. Such misbehaving classifiers have been termed 'Clever Hans' predictors [59] or 'Short-Cuts' [60]. However, identifying such misbehaviour or understanding the meaning of an attribution-based explanation can be significantly more difficult in other domains [61]. For instance, an attribution map computed on a multivariate time series signal or a complex biological sequence can be significantly more challenging to understand for the human receiver; that means the 'interpretation gap' is much more significant than in the horse example. Moreover, even in the image domain, attribution maps only indicate where the relevant information is located, but it is still up to the human to assign meaning to this information. For example, when an attribution map highlights the teeth of a 20 years old person as an indicator for the prediction of the class 'young adult', it does not convey whether the white colour of the teeth is the crucial cue for the prediction or the fact that the person smiles [62].

#### *2.2.2. Ante-hoc explainable models*

Explainability in contexts like finance often has a unique flavour. In this domain, information is mainly presented as tabular or temporal data. Here, traditional ML techniques are often adopted, especially techniques based on Decision Trees (DTs) [63]. The benefit of these techniques is, among others, that they are supposed to lead to ante-hoc explainable models. Some scholars argue that using a black box model, usually derived by applying DL methods, only marginally improves the performance of classical AI methods [64] (however, see [65] for a critical discussion). Accordingly, models that are ante-hoc explainable, such as DTs [66], are preferred for many applications [67]. For this reason, another recent development within XAI is that of rule-based approaches and rule extraction methods, building on their long history within AI. For example, using symbolic rules to derive knowledge is still popular today [68]. Although these methods can improve the overall performance of XAI systems by synthesizing compelling explanations, they are still primarily ignored when prioritizing ante-hoc explainability. One reason might be the low coverage and specificity of the generated trees or rules. In methods based on rule extraction, an opaque 'black box' model is typically trained first and then used to construct

**Table 1**
A short overview of different reviews related to XAI, including their respective contributions.

| Author(s) | Source | Year | Contribution |
|---|---|---|---|
| Ali et al. | [8] | 2023 | An overview of current research and trends in XAI as well as a taxonomy that incorporates four axes of XAI: data explainability, model explainability, post-hoc explainability, and explanation assessment. The authors highlight the connection of XAI to trustworthiness principles, user viewpoints, AI applications, and governmental perspectives. |
| Bodria et al. | [31] | 2023 | An overview of XAI methods (some of them benchmarked), categorized based on the type of explanation they produce. |
| Schwalbe & Finzel | [5] | 2023 | A meta-review of surveys on XAI's methods and concepts. along with a comprehensive taxonomy of the whole field. |
| Weber et al. | [30] | 2023 | A review of XAI methods used to improve ML models, discussing their advantages and drawbacks. |
| Guidotti et al. | [32] | 2022 | A literature review on counterfactual explanations and how to find them |
| Machlev et al. | [33] | 2022 | An overview of current challenges, applications, and future opportunities of XAI for energy and power systems. |
| Mei et al. | [34] | 2022 | A survey on how genetic programming can be used for XAI. |
| Minh et al. | [35] | 2022 | A review of XAI concepts, surveys, and methods, highlighting opportunities and challenges. A taxonomy was proposed for XAI methods with three categories: pre-modelling and post-modelling, explainability, and interpretable models. |
| Speith | [4] | 2022 | A meta-review of taxonomies of XAI methods. The author proposes a new taxonomy incorporating the reviewed ones and suggests creating a database of XAI methods with their properties and a decision tree to help choose fitting methods. |
| Theissler et al. | [36] | 2022 | A literature review on Explainable AI for time series classification. |
| Yang et al. | [37] | 2022 | A mini-review on XAI methods focusing on applications in medicine. |
| Zini & Awad | [38] | 2022 | A survey of XAI methods for natural language processing and their evaluation approaches. |
| Antoniadi et al. | [39] | 2021 | An overview of current challenges, applications, and future opportunities of XAI for clinical decision support systems. |
| Chazette et al. | [29] | 2021 | A systematic review of explainability definitions and the impacts its adoption might have on various system properties. The results are a definition for explainability, as well as a model and a knowledge catalogue of its impacts. |
| Heuillet et al. | [40] | 2021 | An overview of current challenges, methods, and future opportunities of XAI in reinforcement learning. |
| Langer et al. | [6] | 2021 | A review of the goals that XAI is supposed to fulfil. The authors propose a conceptual model to guide interdisciplinary XAI research and highlight the importance of considering the needs of different stakeholders involved with AI systems. |
| Markus et al. | [41] | 2021 | A survey of the role of XAI in creating trustworthy AI for health care, focusing on terminology and evaluation strategies. |
| Mohensi et al. | [42] | 2021 | A multi-disciplinary survey on XAI, focusing on the design and evaluation of explainable AI systems. |
| Rojat et al. | [43] | 2021 | A review of XAI methods for time series data, and an illustration of the type of explanations and the impact they produce. |
| Samek et al. | [44] | 2021 | A review of post-hoc XAI methods, focusing on theoretical foundations, evaluation, and best-practice recommendations. |
| Vilone & Longo | [27] | 2021 | A review of theories, notions, and the evaluation approaches for XAI methods, classified in a hierarchical system. |
| Vilone & Longo | [45] | 2021 | A review of XAI methods that classify them according to their output formats. |
| Confalonieri et al. | [3] | 2021 | A literature of XAI from a historical perspective of traditional and current approaches being developed. |
| Zhou et al. | [46] | 2021 | A survey on methods and metrics for evaluating the quality of XAI methods. |
| Barredo Arrieta et al. | [18] | 2020 | An overview of current research and trends in XAI as well as a taxonomy of existing XAI methods. The authors discuss the implications of XAI towards designing responsible AI systems. |
| Tjoa & Guan | [47] | 2020 | A review of XAI methods. The authors create a categorization of methods that they transfer to the medical field. |
| Carvalho et al. | [48] | 2019 | A general review of XAI, focusing on its societal impact and on metrics for evaluating XAI methods. |
| Adadi & Berrada | [49] | 2018 | An early review of XAI methods, focusing on distilling a taxonomy of methods, providing background on the topic, and identifying open questions and research directions. |
| Guidotti et al. | [50] | 2018 | A review of methods for opening the black-boxes, classifying them into four categories: transparent box design, model explanation, outcome explanation, and model inspection. |

an ante-hoc explainable 'white box' model, such as a rule-based model or a DT. However, limiting the complexity of a DT while achieving a high accuracy via rule extraction is an open problem [69]. Despite these limitations, the use of ante-hoc explainable models instead of black-box models whenever possible is recommended. Even for tasks such as time series forecasting and image analysis, a preliminary data engineering process that includes feature extraction and selection can help use DTs and rule-based models [70,71].

### 2.2.3. New kinds of approaches

Recent approaches have shown potential for resolving problems of older approaches, even if more research must be performed to confirm this [72,73]. One such approach has integrated attention-based explanations into a neural architecture to achieve an efficient computation of tabular data and to increase its comprehensibility [74]. Results are encouraging, but explanations remain highly subject to the inner variability of attention when transformer architectures are used. In that respect, the attention mechanisms could be heavily exploited with a variety of established techniques, including attention flow and rollout [75], LRP adaptation [76], or attention memory [77, 78]. Such techniques are promising in enhancing explanations for complex models but the explanation properties need to be further investigated, especially concerning stability, robustness, and fidelity [61, 79,80]. Connected to the use of rules as a means for enabling the explainability of AI systems, another new trend within XAI is the use of argumentation [81–83]. In particular, computational argumentation can be helpful to explain all the steps towards a rational decision, as well as enabling reasoning under uncertainty to find solutions with conflictual pieces of information [84–86]. In this context, rules are seen as arguments, and their interaction is seen as a conflict that can be resolved with argumentation semantics [87]. Typically, computational argumentation implements non-monotonic reasoning, a type of reasoning where conclusions can be retracted in the light of new evidence [88–90]. This formalism is appealing within XAI because it mirrors one common way in which human reasoning works [83].

### *2.3. Applications of XAI methods*

XAI methods have been widely applied in several fields, including finance, education, environmental science and agriculture, and medicine and health care. This section describes some of the many applications of XAI methods. The goal is to provide stakeholders with illustrations and case studies.

#### *2.3.1. Medicine, health-care, and bioinformatics*

The inferences produced by AI-based systems, such as Clinical Decision Support Systems, are often used by doctors and clinicians to inform decision-making, communicate diagnoses to patients, and choose treatment decisions. However, it is essential to adequately trust an AI-supported medical decision, as, for example, a wrong diagnosis can significantly impact patients. In this regard, understanding AI-supported decisions can help to calibrate trust and reliance. For this reason, many XAI methods such as LIME, SHAP, and Anchors have been applied in Electronic Medical Records, COVID-19 identification, chronic kidney disease, and fungal or bloodstream infections [91]. In these high-stakes scenarios, there is evidence that AI-based systems can have superior diagnostic capabilities than human experts [92]. Thus, the explainability of these systems is not only a technological issue but boils down to medical, legal, ethical, and societal questions that need careful consideration [93]. As in other domains of application, it is essential to connect XAI with requirements for trustworthy AI in this area. In this regard, a comprehensive survey of trustworthiness requirements for a practical case study in AI for healthcare can be found in [94]. This work supports the compliance of AI-based systems with regulation in high-risk scenarios as considered in the EU AI Act.[1]

#### *2.3.2. Finance*

In finance, institutions such as banks and investment firms leverage AI to automate their processes, reduce costs, improve service security, and, generally, gain a competitive advantage. AI algorithms are used at scale to predict credit risk, detect fraud, and diagnose investment portfolios for optimization purposes. Applying AI often requires transparency and explainability in these contexts for legal reasons. This requirement is particularly significant in the customer banking sector, where banks must comply with strict regulations such as the USA Equal Credit Opportunity Act (ECOA) or the USA Fair Housing Act (FHA) to expose adverse action codes and provide clear explanations for their decisions. Similar guidelines and law enforcement are present in Europe, guided by the General Data Protection Regulation (GDPR) law of the European Union. For example, if a customer's loan application is denied, the bank must be able to provide a clear and understandable reason for this. When adopting AI algorithms, it becomes increasingly difficult for banks to provide stable and trustworthy explanations [95,96]. In other words, it becomes increasingly complex to justify the inferences of AI models, both with simpler ante-hoc explainable models [97–99], and even more with complex models [72–74]. This lack of explainability can put banks at risk of regulatory penalties and erode customer trust. In investment banking, the demand for XAI is driven by the need to ensure the robustness and stability of AI systems [100], which could be subjected to extreme market conditions and unexpected events. If an AI system makes problematic inferences to validate information, it could lead to disastrous outcomes.

#### *2.3.3. Environmental science and agriculture*

Another area of application of AI that has benefited from adopting XAI methods is the intelligent analysis, modelling, and management of agricultural and forest ecosystems, an essential task for securing our planet for future generations. For example, forest carbon stock is a critical metric for climate research and management, as forests play a vital role in sequestering atmospheric carbon dioxide. In this context, drones can be deployed for data collection, and ML techniques can be used for estimating forest carbon storage [101]. Forest inventory also plays a crucial role in forest engineering, as it provides critical information on forest characteristics, such as tree species, size, and density, which can inform forest management decisions [102,103]. In these life-critical environments, sensor-based technology is employed to collect data, which is often high-dimensional and heterogeneous, and then AI-based models are trained on it. However, data is often poor in quality, thus leading to models that lack robustness. Furthermore, even if such models are robust, there are still challenges in terms of tracing and understanding their inferences and ascertaining the causal factors that underlie them. Even the slightest perturbations in the input data can dramatically affect a model's output, leading to entirely different inferences and thus undermining the trustworthiness of such models [104,105]. Additionally, in these naturalistic environments, a challenge for forest engineering is the development of methods for uncertainty quantification and propagation. AI methods for developing forest inventory models are subject to various sources of uncertainty, including measurement error, spatial variability, and model misspecification. It is, therefore, crucial to analyse the robustness of AI methods—for instance, through explainability—and enhance it for the produced models and their inferences [106,107].

#### *2.3.4. Education*

AI in Education (AIED) focuses on developing AI-powered educational technologies to aid students, instructors, and educational institutions [108–110] in their teaching and learning activities. On the one hand, for students, AIED has focused on developing models [111] and adaptive systems that can identify learners' strengths and weaknesses across a variety of topics, leading to customized instructions and resources that align with their learning needs [112]. These are, for example, focused on improving their meta-cognitive processes of self-monitoring, reflection, and planning [113]. On the other hand, for instructors, AIED tools can act as intelligent teaching assistants [114], help them orchestrate the classroom [115], grade assessments [116], and answer student queries [117], minimizing students dropout [118]. The most recent example of an application of AIED includes using some Large Language Model (LLM) capable of generating new textual content based on human input prompts. This can be used to write essays, produce software code, or generate educational content such as multiple-choice questions or work examples with step-by-step solutions. A growing concern is that students, instructors, and educators lose control of AI-based technologies as they fail to determine how these work, why they produce specific outputs, and what impact they may have. In particular, AIED tools such as educational recommender

[1] EU AI Act: https://artificialintelligenceact.eu/the-act/, accessed on January 5th, 2024.

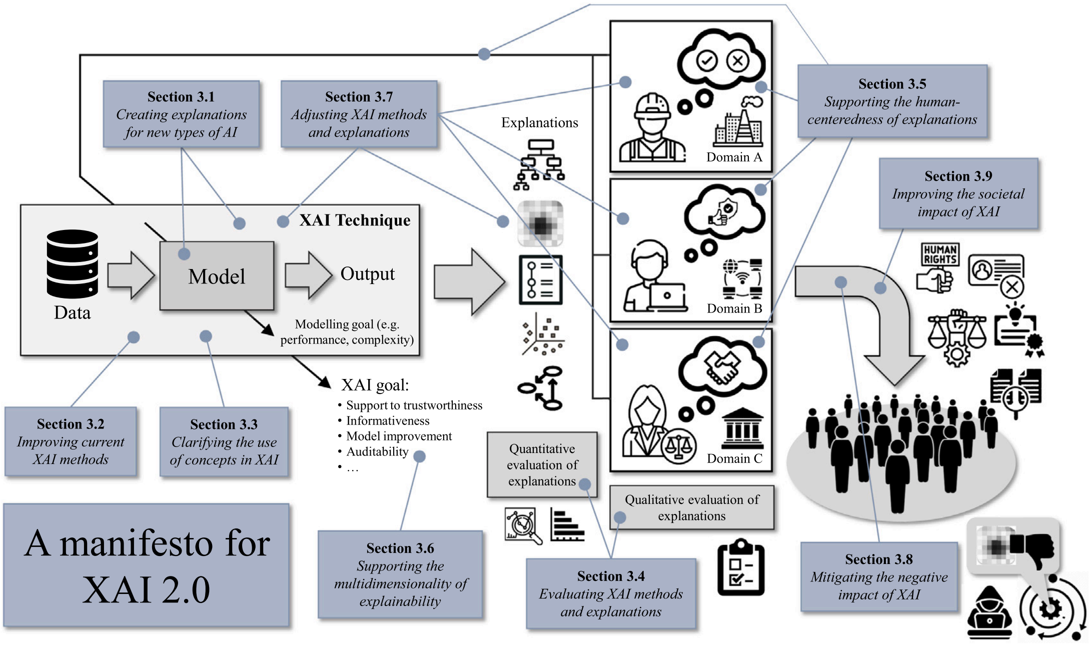
Section 3.1
Creating explanations for new types of AI
Section 3.7
Adjusting XAI methods and explanations
Explanations
Section 3.5
Supporting the human-centeredness of explanations
Domain A
XAI Technique
Data
Model
Output
Modelling goal (e.g. performance, complexity)
Section 3.9
Improving the societal impact of XAI
HUMAN RIGHTS
Domain B
XAI goal:
• Support to trustworthiness
• Informativeness
• Model improvement
• Auditability
• …
Section 3.2
Improving current XAI methods
Section 3.3
Clarifying the use of concepts in XAI
Quantitative evaluation of explanations
Domain C
Qualitative evaluation of explanations
A manifesto for XAI 2.0
Section 3.6
Supporting the multidimensionality of explainability
Section 3.4
Evaluating XAI methods and explanations
Section 3.8
Mitigating the negative impact of XAI

**Fig. 1.** A manifesto for eXplainable Artificial Intelligence (XAI): High-level challenges.

systems are increasingly used to automate and personalize learning activities [119]. These tools pose various concerns about their use in high-stakes decisions, including fairness, accountability, explainability, and ethics [120–122]. The impact of these technologies on students' agency and self-regulated learning is a growing concern, as the lack of explainability and feedback can make it difficult for instructors and learners to calibrate their trust in AI-based inferential systems and understand their current state of learning and the benefits derived from engaging with a particular educational resource [123].

## 3. Challenges and research directions

Despite the many advances, breakthroughs, and potential applications of XAI methods, more research is required to address open problems in the field. For example, it is still unclear how XAI methods should be evaluated, how different terms should be used in the debate, or how, strictly, XAI is related to trustworthiness. Many surveys tackling some of these aspects of XAI exist and keep appearing in conference proceedings and journals [4,8] (see Table 1 for an overview). However, they are somewhat scattered, often specific to an application domain or focused on specific methods. Against this backdrop, this section aims at extracting and synthesizing the diverse challenges in XAI that motivate the formation of a manifesto. Overall, we identified 28 problems, which we have grouped into nine high-level categories – our manifesto – as depicted in Fig. 1. These problems and the related challenges are often interconnected; thus, they may, in principle, belong to multiple categories.

### *3.1. Creating explanations for new types of AI*

The ever-evolving landscape of AI introduces novel types of models, such as generative models or distributed and collaborative learning algorithms, each with its unique set of properties. Against this backdrop, this category of challenges describes the intricacies of creating explanations for these new types of AIs.

#### *3.1.1. Creating explanations for generative and large language models*

Generative AI models, such as those employed for diffusion denoising [124,125] or the family of GPT models for large-scale language generation [126], are disrupting many sectors. These models deliver exceptional performance due to their immense scale. With billions, and in some cases, nearly trillions of parameters, their sheer size poses a significant challenge to existing XAI methods [38]. In particular, these methods grapple with the high-dimensional nature of such models, both in terms of computational complexity and in extracting learned concepts. For instance, one obstacle related to the latter point lies in the polysemantic nature of the neurons in generative models (that is, one neuron can represent several concepts), which is thought to arise from a superposition of multiple independent features [127]. XAI methods have been mostly limited to classification and regression problems. Accordingly, entirely new approaches have to be developed for generative models. In particular, self-supervised or neural generative models such as Variational Autoencoders (VAEs) and Generative Adversarial Networks (GANs) are becoming more popular. For instance, examining the latent spaces they learn and synthesizing their explanations is very challenging. Another challenge, particularly for LLMs, concerns scaling laws. Neural scaling laws are functional relationships that relate variables to a neural network, such as the number of layers in its architecture and its achieved accuracy after training. Such laws govern the aggregate capabilities of LLMs, yet a precise understanding of individual task-level implications of these laws remains elusive, as they appear to manifest unpredictably. Whether scaling laws can be used to infer the quality of the artefacts or concepts learned by LLMs is an open issue.

*Solution ideas.* Mechanistic interpretability [128,129] is a promising approach to gaining deeper insights into generative models' functioning and scaling laws. Basically, the idea of mechanistic interpretability is to reverse-engineer neural networks (for instance, in terms of parameters) to find out what the model is actually doing. For instance, one goal of mechanistic interpretability is to find meaningful algorithms in the weights of neural networks [130]. Insights gained in this way can be useful to understand grokking mechanics (that is, a sudden spike in generalization even though the network has more parameters than training data) [131] and the ability to solve problems recursively [132]. In particular, mechanistic interpretability has shown promising results at small model scales and for toy problems. Researchers from institutions and companies like MIT, OpenAI, DeepMind, and Anthropic pursue mechanistic interpretability as an approach that attempts to reverse-engineer the learned representations and algorithms of trained models using causality-based methods. Piecewise linear activation functions have been used to partition the activation space into polytope-shaped monosemantic regions [133], and sparse autoencoders have been successfully used for the mono-semanticity of Deep Neural Network (DNN) models [127]. There are also challenges in mechanistic interpretability, such as disentangling multiple algorithm implementations and finding unknown algorithms [134]. Furthermore, there are preliminary results from vision models that scaling does not help the mechanistic interpretability of models [135], calling for designing models for mechanistic interpretability. A potential complement to mechanistic interpretability may be information geometry [136], which can help analyse high-dimensional spaces involved in processing LLMs. Furthermore, constraints may have to be imposed on the training and functioning of LLMs to ensure safety and explainability [135,137]. Such constraints could be directly part of the automated optimization (learning) process or indirectly used through a human-in-the-loop approach. An example of a promising direction can be found in [138], which introduces a training procedure that encourages modularity and ante-hoc explainability by discouraging non-local connections between neurons through local L1 regularization with swaps of neuron locations. Finally, whether these methods can be scaled to relevant models, problem sizes, and complexities remains to be seen.

#### 3.1.2. Creating explanations for distributed and collaborative learning

An ML setup that has garnered much attention lately, especially in privacy and security-critical applications, is distributed and collaborative learning (as in federated learning, among others). Distributed learning algorithms involve multiple computing nodes or devices collaborating to solve an ML problem. These nodes typically train models on their local datasets and periodically share information with other nodes to improve a global model [139]. Collaborative learning is a broader concept where multiple entities work together, sharing knowledge or resources to achieve a common learning goal. Explainability is also relevant in such scenarios to create explanations after the collaborative learning process [140]. However, designing XAI methods for collaborative learning scenarios without compromising sensitive data is challenging. Furthermore, models in such scenarios are updated iteratively and asynchronously, making it hard to trace changes and understand the impact of each node's contribution to the global model. To make matters worse, the nodes in distributed systems can have different data sets and model architectures, and as the number of nodes increases, explaining the decisions of a globally trained model becomes more complex.

*Solution ideas.* One avenue to create explanations for distributed and collaborative learning scenarios is to exploit the very architecture of this approach. Accordingly, nodes would generate explanations locally based on their updates to the model. These local explanations can be aggregated and summarized without sharing sensitive data, providing insights into the global model's behaviour. Another approach would be to use Multi-Party Computation (MPC) known from cryptography [141]. MPC protocols could be utilized to enable collaborative computation of explanations without directly sharing sensitive data among nodes. This ensures that no single node can access complete information while still contributing to the explanation process. Finally, another possible solution would be implementing differential privacy mechanisms to generate explanations to minimize the risk of revealing sensitive information. This could mean that the perturbation made to derive explanations is restricted to protect individual data.

### 3.2. Improving current XAI methods

A spectrum of challenges arises when considering current XAI methods. Many of these have long-known disadvantages that need to be overcome, as described below.

#### 3.2.1. Augmenting and improving attribution methods

One major branch of XAI methods relies on pixel attribution with heatmaps or saliency masks [142], one of the most prominent classes of XAI methods used for computer vision tasks. Such methods are often based on perturbations (that is, varying the input to look for changes in the output) [55,143] or gradients [144,145]. Despite the great success of these methods to, for instance, detect biases and flaws in the learned prediction strategies (so-called 'Clever Hans Effect' [59], see Section 2.2), attribution-based explanation methods also have limitations. For instance, saliency masks on the level of pixels are often unsuited for laypersons [61]. A major technical limitation of attribution methods is their sensitivity to (1) internal hyper-parameter tuning and customization (such as baselines), (2) the format chosen for their results, and (3) assumptions regarding the model under exploitation. For example, the results of model-agnostic attribution methods, including LIME and SHAP, can change based on the range of input perturbation. Similarly, many gradient-based methods require setting a proper sampling interval. Finally, relevance propagation methods (a type of gradient-based attribution method) such as LRP [145] have to be adjusted to the layer of the DNN. Additionally, some methods have issues with computational efficiency, requiring many passes for calculating attributions.

*Solution ideas.* One idea to solve the problems of attribution methods is to combine them with other XAI methods to obtain a portfolio approach that compensates for the weak properties of the individual approaches. The methods in the portfolio could negotiate, like in a market, to coordinate and converge to a majority view or, even better, to a list of hypotheses with their plausibilities based on the votes of each portfolio participant. Mechanistic interpretability is an orthogonal approach with different characteristics that could complement these approaches effectively. Likewise, ante-hoc explainable models can be used for attributions, avoiding some of the sensitivity and efficiency issues.

#### 3.2.2. Augmenting and improving concept-based learning algorithms

Concept-based learning algorithms are a large and increasingly popular class of methods that can be used for both post-hoc explainability and creating ante-hoc explainable models. The idea of concept-based learning algorithms is to explain a model's predictions in terms of human-understandable attributes or abstractions [146]. Several such algorithms have been proposed over the years to directly learn features that describe 'prototypical concepts' or 'prototypes' present in individual inputs to the model, including ProtoPNet [147], ProtoTree [148], ProtoPShare [149], Concept Bottleneck Models [150], Concept Activation Vectors [146], Concept Embedding Models [151] or Concept Atlases [152]. Neuro-symbolic learning, the symbiosis between connectionist and concept-based symbolic learning, has recently also gained momentum [153–155]. Hybridizing knowledge graphs (KGs) with learning algorithms also fall within the landscape of approaches used to map knowledge encoded in the parameters of a model with human-understandable concepts and the interrelationships among them [156]. Unfortunately, use cases proposed to showcase how these approaches explain their decisions are limited, very narrow, and assume a priori knowledge about the concepts that can be discriminative for the task at hand. This assumption may imprint a significant inductive bias in their explanation-producing process, not appropriately generalizing when explaining distributionally novel inputs concerning the training data. Most methods explaining concepts rely on a given dataset of human-defined concepts [157], which, however, might not be available for a specific domain and must be collected at high costs. Furthermore, even if a dataset is available, there is a considerable risk that the user-defined concepts are incomplete or inaccurate, leading to poor or biased explanations. Furthermore, the continuous proposal of new datasets for concept learning, including Clevr [158] and Clevrer [159], Kandinsky Patterns [160], or Closure [161], sheds evidence on the need for eliciting local explanations that can be formulated in terms of concepts and their spatial distribution.

*Solution ideas.* One line of research to solve the above problems explores the potential for genetically evolvable connections between identifiable concepts in the input data using object recognition models and evolutionary programming solvers [34]. This hybridization could offer the advantage of employing symbolic classifiers that are ante-hoc explainable and well-suited for handling datasets that encapsulate discriminative, concept-wise compositional information. Additionally, a growing demand exists to expand hybrid approaches that unite KGs with concept-based learning methods. This expansion aims to enable the discovery of relevant concepts, attributes, and relationships that extend beyond the confines of specific use cases or domains, as discussed by Lecue et al. [79].

#### 3.2.3. Removing artefacts in synthesis-based explanations

Generating explanations through synthesis is a promising direction to advance the field of XAI. The idea of this approach is to synthesize examples from the training set of a model that contribute to the prediction of a particular class or to visualize the features that a neural network layer has learned. While a user is unlikely to understand a neural network's layer activations of specific classes directly, the understanding may differ when considering examples of those classes. The synthesis of such examples, however, is often noisy. For instance, a synthesized image may contain artefacts. It is unclear whether this noise is due to the synthesis process itself or is, de facto, part of a concept learned by the model. For example, while a GAN architecture can synthesize an image representing the pattern that activates a neuron most strongly, this image might have various artefacts that make it appear somewhat distorted. This might happen due to shortcomings of the GAN models, which means these artefacts must be present to activate the neuron strongly. Two existing methods for synthesis in the literature are a decoder for layer activation [162] and a GAN for single neurons [163]. Unfortunately, the mere synthesis of inputs is insufficient for understanding concepts. There are few works on using generative models for explanations, including the work of [162] and the chapter concept vectors in [157].

*Solution ideas.* To minimize artefacts, state-of-the-art models and recent popular techniques in DL [164], especially diffusion models, could be leveraged [124]. However, even state-of-the-art generative models do not ensure the absence of artefacts. Thus, to verify any distortions due to the synthesis, one idea is to compute a reconstruction of the original input serving as a Ref. [165]. This reference stems from a separate model with the same architecture as the decoder synthesizing inputs from the model to explain. Subsequently, a user can compare the original input, the synthesized image from layer activations of the model to explain (that means, what the classifier 'sees'), and the reference, allowing them to identify distortions due to the synthesis process. If it can be seen that the original image and the reference are reasonably similar, then distortions might be considered minor. However, a classifier might not rely on certain concepts associated with the input. Therefore, while the comparison with a reference might be considered a valid approach, it is still tedious for the lay user and non-trivial to apply beyond autoencoders.

#### 3.2.4. Creating robust explanations

The fragility of posthoc XAI methods to small perturbations at the model's input and the known inconsistency in synthesized explanations for a given input [166] (see Section 3.2.3 above) highlight the challenge of creating compelling explanations. This is frequently advocated as a requirement for calibrating human trust and building acceptance of a model being audited. Several works have advocated the idea of exploiting explanations beyond just explaining decisions [30,167], for instance, also to improve models. However, the susceptibility of explanations to the XAI technique under consideration detracts from the explanation's robustness, jeopardizing the reliable application of explanations to improve a model. Methodologies for delivering robust explanations under different circumstances are investigated in several recent works [100,168,169]. A satisfying solution, however, does not yet exist. The difficulty lies predominantly in the fact that the model itself must be robust for a robust explanation.

*Solution ideas.* As a first step towards robust explanations, evaluations on benchmarks should be done to identify common biases of an XAI method and to define ways to mitigate them. Furthermore, robust explanations could be created by aggregating explanations (see also the solution idea in Section 3.2.1). For example, a proposal exists to blend uncertainty quantification and XAI [170]. Other research has emphasized the robustness of the AI model itself, for instance, in the form of explanations that inform about model inversion or extraction attacks [171,172]. In a similar vein, the recently proposed "reveal to revise" framework enables practitioners to iteratively identify, mitigate, and (re-)evaluate spurious model behaviour with a minimal amount of human interaction [173]. Finally, ante-hoc explainable models have the advantage of the explanation and the model being interdependent so that there is no loss of robustness when generating an explanation (the robustness of the model itself must still be guaranteed, however).

### 3.3. Clarifying the use of concepts in XAI

As a multidisciplinary research area, another category of challenges for XAI is the disparate and unclear use of terms.

#### 3.3.1. Elucidating the main concepts

In research on XAI, there is a conceptual ambiguity regarding various terms, such as explainability, interpretability, transparency, understanding, explicability, perspicuity, and intelligibility. This represents a challenge in XAI, as the lack of clear and consistent definitions of terms can hinder progress in developing practical and valuable XAI systems. Some researchers use terms like explainability and interpretability synonymously [23,26], while others draw significant distinctions between them [64]. These differences pose problems for applied research and interdisciplinary collaboration. Discussions about clarifying terms in the XAI field tend to take two distinct approaches. On the one hand, some contend that attempts to define the terms in question are futile, impossible, counterproductive or unnecessary, and previous definitions of explainability have failed and, in general, the whole endeavour of finding definitions is doomed to failure (for example, [174–176]). On the other hand, some attempt to provide explicit definitions, intending to differentiate between the various terms employed (for example, [27,177,178]).

*Solution ideas.* As the lack of a clear and consistent definition of terms related to explainability can hinder progress in developing practical and valuable XAI systems, the communication challenges should be addressed holistically rather than perpetuated by ambiguity in the use of terms. Against this background, it seems desirable to join the latter of the above camps and strive for a uniform use of different terms. A minimal solution of this kind would be for authors to always clarify, in their articles, what they mean by certain concepts. A more desirable solution, however, would be to define the various terms once and for all. In this line of thought, meaningful definitions can only be found if already existing ways of usage are considered. Creating entirely new usages of the various terms is more likely to contribute to conceptual confusion than to resolve it. The first step in coining a generally applicable definition of the terms is to identify their current usage and create an overview and comparison of them. For instance, some work identifies relevant notions [27], but limited work exists in comparing them. The merit of the various proposed definitions must be determined as the next step. For this purpose, quality criteria should be established (see, for instance, [177]), which can be consulted to evaluate each proposed definition. In summary, seeking consensus around the terminology in use by the community is constantly required, along with a continuous effort towards developing the technology associated with each concept, which conforms to one of the core objectives of this manifesto.

#### 3.3.2. Clarifying the relationship between XAI and trustworthiness

A similar conceptual challenge exists concerning trustworthiness. Properties like safety, fairness, and accountability are often mentioned for meeting regulatory actions focusing on the trustworthiness of AI. For instance, the Ethics Guidelines for Trustworthy AI, issued by the EU High-Level Expert Group on AI, listed seven requirements for AI-based models and systems to be seen as trustworthy [11,179]: human agency and oversight, technical robustness and safety, privacy awareness and data governance, transparency, diversity, non-discrimination and fairness, societal and environmental well-being, and accountability. While XAI has the potential to help with most of these [6], it is taken to help with one of them primarily: transparency. However, even this relationship is unclear, as many sources contain contradicting statements. In these sources, it is possible to observe various claims about the relationship between trustworthiness and XAI: trustworthiness is seen as a central goal of XAI [18], but XAI is also claimed to be a part of trustworthiness [180]. XAI is purported to change the belief in the trustworthiness of a system [181], while it should also support the trustworthy integration of systems [182]. These are just a few examples, and in other articles, it is also possible to find completely different relationships (see [13] for an overview). One reason for this divergence is that there is no uniform way of using terms like trustworthiness (and other terms in XAI, see Section 3.3.1). For this reason, as long as it is not clarified what each term describes and what property it expresses, it will not be possible to specify the relationship between XAI and trustworthiness (and other desirable properties such as fairness and safety).

*Solution ideas.* The relationship between XAI and trustworthiness is widely discussed [13]. To clarify the relationship between XAI and trustworthiness, more needs to be learnt about the trustworthiness of AI systems. Here, one could build on results from philosophy and psychology [13,179], which have been researching the concept of the trustworthiness of humans or organizations. In general, we must distinguish between trustworthiness as a property of an AI system and the technical requirements required for an AI system to be trustworthy. As for the latter issue, XAI is identified as one of the seven requirements for trustworthy AI [11,179]. Against this background, XAI must contribute towards achieving trustworthiness in connection with the rest of the requirements for this purpose. Steps in this direction can be found in reviews like [8,29,30], which highlight methods that leverage XAI to guide and improve ML models or clarify the impact of explainability on other system properties. This suggests that the use of XAI may contribute to accomplishing other trustworthiness requirements. Concerning the prerequisites for an AI system to be trustworthy, we highlight the report recently published by the UC Berkeley Centre for Long-Term Cybersecurity (CLTC) [183]. This report aims to help organizations develop and deploy more trustworthy AI technologies, including 150 properties related to one of the seven “characteristics of trustworthines” defined in the NIST AI RMF[2]: valid and reliable, safe, secure and resilient, accountable and transparent, explainable and interpretable, privacy-enhanced, and fair with harmful biases managed. Another essential aspect for XAI is AI governance [184] and the need for governance measures linked to managing AI risks. These new scenarios pose essential challenges for the design, development, and safe deployment of AI systems [179]. In the current debate, XAI is identified as a vital ethical principle and technology to decrease the uncertainty and concerns about AI systems in society.

#### 3.3.3. Finding a useful account of understanding

Another challenge to conceptual clarity is finding a valuable account of understanding. An obstacle to providing such an account of understanding in XAI is the lack of conceptual clarity about what understanding itself is. In philosophy, there are at least three different approaches to this problem. The more traditional view asserts that understanding logically depends on explanation: only *true* explanations can provide understanding [185,186]. The other end of the spectrum is occupied by philosophers who entertain alternative paths to understanding, even if they present distorted or false accounts of their targets [187,188]. Finally, intermediate views exist, allowing that some, but not all, of the pieces of information used to provide understanding can be false [189–191]. There is no consensus regarding which of these views is more adequate in the context of AI explanations. For example, while [192] sides with the traditional view, [24] adopts a more pragmatic stance. Another obstacle arises when the understanding provided by XAI methods, focusing on singular predictions, differs from the type of understanding offered by proxy or surrogate models that provide a global account of the target AI model. There might be different underlying cognitive processes and abilities involved in each case. Prima facie, the explanation of singular predictions provides a type of understanding that epistemologists call ‘understanding-why’ [193], while proxy or surrogate models provide ‘objectual explanations’ [194] of their targets. The relation between the two types of understanding requires clarification both from a philosophical and psychological point of view. In addition, a third type of understanding depends entirely on the functional correlations between inputs and outputs [195]. Functional understanding might be sufficient for most users in many cases of human–computer interaction.

*Solution ideas.* Solving the problem of a useful account of understanding in XAI potentially requires a two-pronged approach. On the one hand, conceptual clarity is required. Several recent papers [22,24,29,192,196–201] have focused on the relation between explanation and understanding in AI. The conceptual map of this specific problem is now quite clear. Still, future developments will have to respond to new psychological evidence about human–computer interaction and to the development of new XAI methods. On the other hand, empirical work on understanding is essential. For a long time, XAI researchers have tried to ensure that the methods they develop are comprehensible to their peers, a phenomenon referred to as “inmates running the asylum” [17, p. 36]. The proposed and endorsed alternative is to incorporate results from psychology and philosophy to XAI [6,23,202,203]. Existing theories of how people formulate questions and select and evaluate answers should inform the discussion [23].

### 3.4. Evaluating XAI methods and explanations

Evaluation is essential to developing and deploying XAI systems. However, evaluating XAI methods is a complex task, and no gold standard exists on what makes for a good explanation [61]. It is vital to mention that this is an objective that closely relates to the perception of humans in the evaluation, as later discussed in this section, and also linked to the human-centeredness of explanations, as discussed in Section 3.5. While there is some overlap, these sections are aimed towards two distinct objectives: evaluating explanations versus adapting explanations to humans.

#### 3.4.1. Facilitating human evaluation of explanations

One problem concerning evaluating the XAI methods is that they often lack user studies. Current evaluation approaches typically only analyse specific properties of the XAI methods themselves without accounting for the interaction with the final user [27,32,80,204,205]. For instance, a survey of user studies has shown that only 36 out of 127 research works employing counterfactual explainers adopted a human evaluation approach, and only 7% of them tested alternative approaches [206]. Individual differences in understanding, prior

[2] NIST AI RMF, https://www.nist.gov/itl/ai-risk-management-framework (Last access: January 5th, 2024).

knowledge, and the cognitive load required to comprehend explanations further challenge evaluating XAI methods. It is not difficult to compare different forms and types of explanations to determine the most effective. Additionally, users are typically ‘passive recipients’ of explanations, and the actual usage or exploitation of such explanations is barely tested. For specific properties, no approaches at all that test for them [80]. While some studies evaluated the impact of synthesized explanations of AI systems on humans when compared to the scenario where no explanations were provided [207–209], there is a need for more (and more systematic) work on the topic.

*Solution ideas.* Establishing a solid foundation for XAI must be grounded in empirical research involving users. Achieving this demands a collaborative, interdisciplinary approach, uniting ML experts with researchers from HCI, psychology, and the social sciences. Valuable insights can be gleaned from the collective body of knowledge in these domains, leveraging their expertise in conducting user studies [17,23]. To streamline the evaluation process, it is imperative to establish standardized frameworks encompassing every stage, from formulating hypotheses to data collection, analysis, and utilizing online questionnaires. With this robust methodology in place, the research community can then embark on the crucial task of developing heuristics, principles, and patterns that enable the design of effective XAI systems for real-world applications. This comprehensive approach ensures that XAI benefits from theoretical foundations and is shaped by empirical user-centric research, ultimately enhancing its practical utility.

#### 3.4.2. *Creating an evaluation framework for XAI methods*

Several works address the evaluation of XAI methods. For instance, Hoffman et al. [210] integrate extensive literature and various psychometric assessments to introduce critical concepts for measuring the quality of an XAI system. Similarly, Vilone and Longo [27] aggregate evaluation approaches for XAI methods from several scientific studies via a hierarchical system. Furthermore, Van der Lee et al. [211] define a list of steps and best practices for conducting evaluations in the context of generated text. An analysis of these works reveals that evaluating the goodness and effectiveness of explanations is a prerequisite for calibrating trust in AI. However, standardized methods and metrics are lacking for evaluating XAI systems. In other words, despite the broad interest in the design of XAI methods [18,31,48–50], it is still unclear how to compare the results of different evaluations and establish a common understanding of how to evaluate explanations. What is missing is a set of evaluation metrics for explainability that are generally applicable across studies, contexts, and settings.

*Solution ideas.* There are already some promising approaches in the literature to solve this problem. In a recent survey on the evaluation of XAI, for instance, the authors identify several conceptual properties that should be considered to assess the quality of an explanation, and they propose quantitative evaluation methods to evaluate an explanation [212]. Recently, a survey-based methodology for guiding the human evaluation of explanations was proposed in [213]. This methodology amalgamates leading practices from existing literature and is implemented as an operational framework that assists researchers throughout the evaluation process, encompassing hypothesis formulation, online user study implementation and deployment, and analysis and interpretation of collected data. Furthermore, the recently developed XAI evaluation framework Quantus [214] implements over 30 evaluation metrics from six categories. Frameworks such as Quantus allow for evaluating and comparing explanations in a standardized and reproducible manner. Furthermore, publicly available XAI evaluation datasets with ground truth information, such as CLEVR-XAI [215], allow for objective evaluations. In the future, artefacts like these need to be extended to more application areas, especially outside of computer vision.

#### 3.4.3. *Overcoming limitations of studies with humans*

Evaluating XAI methods with humans has limitations. Often, the number of participants that can be put together in a study does not represent the general population. Thus, a study’s results may be prone to bias and errors and may not generalize well [216]. Overall, the evaluation of XAI methods in studies with humans is prone to issues such as poor reproducibility and inappropriate statistical analyses, resulting in no solid evidence for their usefulness [207–209,217–220].

*Solution ideas.* A potential solution involves augmenting human studies with synthetic data and virtual participants. Researchers can address the issue of limited sample representativeness by creating synthetic datasets that span a wide range of demographic characteristics, behaviours, and preferences. These synthetic datasets can simulate diverse user profiles and scenarios, enabling more robust and extensive evaluations of XAI methods. Additionally, virtual participants, based on AI-driven agents or personas, can be incorporated into studies to provide a broader range of user interactions and perspectives. Standardized methodologies and statistical analyses must be employed to enhance the reproducibility and rigour of XAI evaluations. Researchers should adopt transparent reporting practices and adhere to well-defined evaluation protocols, ensuring that the evidence generated from these studies is solid and dependable. Another approach would be to create sample explanations or schemes for explanations against which generated explanations are checked. While the samples would still need to be tested in studies first, this could alleviate the overall need for studies with humans.

### 3.5. *Supporting the human-centeredness of explanations*

One class of challenge in XAI lies in providing explanations adapted explicitly to the humans receiving them.

#### 3.5.1. *Creating human-understandable explanations*

In his seminal paper about explanations in AI and social sciences, Miller points out that explanations should be social, contrastive, and selective to be understandable to humans [23]. Confalonieri et al. discussed further properties for explanations, including integrating symbolic knowledge and statistical approaches to explainability [3,221]. Unfortunately, many current XAI methods do not have these properties. In particular, many XAI methods provide explanations that do not extrapolate beyond the domain of their input data. A clear example of this phenomenon is the manifold number of gradient-based attribution methods [144,145,222], all yielding explanations in the form of visual heatmaps quantifying the relative importance of every pixel of the input image to the prediction issued by the model. Many contributions assume that such heatmaps are enough for explainability simply because a ‘narrative’ can be built to relate pixels to concepts that emerge from intuition. There are, however, several problems with this assumption. First, the intuitions in question are often from experts [17], and the presentation of explanations in pixel attributions may not be comprehensible to laypersons [61]. Second, in more complex scenarios, crafting a narrative can become challenging, especially when discriminating between classes relies on intricate distributions of concepts within an image [223,224], or other semantically defined relations among the entities to which these concepts belong. Third, these narratives are sometimes elaborate guesses at best. Assume a saliency map, serving as an explanation, highlights coarsely a person’s face to classify it as a human. It is unclear whether the underlying classifier used features such as the shape of the face, the skin colour of the face, or characteristics of the face such as mouth and lips, or a combination thereof to make its inference.

*Solution ideas.* Audiences without technical background are often concerned with *concepts*, not with data. For instance, in a classifier discriminating between 'dogs' and 'cats', it is significantly more informative for many people to state that 'the shape of whiskers' is a discriminative concept in the images rather than the relevance of isolated pixels as dictated by a gradient-based attribution technique. In this line of thought, concept-based XAI methods (such as concept-based learning algorithms, see Section 3.2.2) explain individual predictions not as pixel-wise attributions but in terms of semantically meaningful concepts (for example, 'eye', 'red stripe', 'tyre') represented by hidden-layer elements of the neural network. Often, concept-based explanations can be enriched by reference samples from the training dataset. Combining local XAI methods with global XAI methods might lead to semantically richer and more human-understandable explanations. This so-called 'glocal' approach was taken in concept relevance propagation, an upgrade to LRP, to simultaneously identify concepts learned by the model (global) and match them to each input (local) [152]. Enriching explanations with explicit knowledge can enact scenarios in which formal and common-sense reasoning can be used to create explanations that are closer to how humans think. In this line of thought, computational argumentation techniques could be exploited to generate explanations that can mimic the way humans reason under uncertainty [82–85,225–228]. Another possible solution to create human-understandable explanations is to map explanations to a more comprehensible domain. For instance, one approach to providing more comprehensible explanations on time series data has been recently explored in [229]. In this context, the explanation is firstly computed on the time domain, which is the domain of the operation of the model. Then, the solution is mapped through an invertible layer where explanations can be computed in different spaces. Future research should investigate meaningful invertible mappings, for example, using autoencoders [230] for this and other domains.

#### 3.5.2. Facilitating explainability with concept-based explanations

Humans and AI systems make decisions differently. In particular, AI systems, especially those based on DL, often rely on features that humans can grasp. On the other hand, humans use concepts that are coarse-grained representations of reality [231,232]. This difference is often not taken into account when it comes to creating explanations. For example, prominent explainability methods such as LIME or SHAP rely on feature attributions that might reveal little about how an AI model works [57,58]. Concept-based XAI methods go beyond attribution and aim to express human-understandable concepts as part of the explanation that must first be synthesized from the model to be explained. One benefit of concept-based explanations is that they can aid the insertion of expert knowledge in the learning process of a model, allowing users to impose explicit domain-driven constraints defined as concepts, attributes, and predicates (for example, in so-called Logic Tensor Networks [233]). However, explanations based on human-understandable concepts are still in early development. In particular, concept-based explanations are primarily elaborated only for classification or regression models, leaving aside other problems and models for which concept-based explanations could be helpful. This could be the case for reinforcement learning, in which explanations should inform about how the agent's interaction with *concepts* existing in the environment produces a series of actions that fulfil the formulated task [40]. Furthermore, limited work investigates XAI methods that aim at synthesizing human-understandable concepts in concrete applications. While some concepts are universal, such as 'every car has tires and tires are round', others are more subjective or differ among stakeholders and cultures and depend on domain knowledge, that is, knowledge related to training data [234]. Accordingly, a method that is generalizable and applicable across diverse areas and contexts is needed, as one might be interested in using concepts in a personalized way to explain.

*Solution ideas.* Creating concept-based XAI requires a multi-faceted approach considering a broad range of sub-problems. It begins with finding reliable ways to extract and identify relevant concepts from data or AI models. For this first step, employing techniques from natural language processing, semantic analysis, and domain-specific knowledge can assist in systematically pinpointing concepts. This systematic identification lays the foundation for offering insights rooted in comprehensible terms. Here, concept-based learning algorithms could be a fruitful way forward (see Section 3.2.2). Next, the concepts must be personalized to tailor to the individual consuming them. Allowing users to define their concepts would be one way to ensure personalization. Interdisciplinary collaboration and continuous feedback loops could refine these concepts, making them more meaningful and comprehensible. A supplementary avenue could be to organize concepts within a hierarchical structure. This structure could be useful for delivering explanations that can be provided at different levels of granularity. This hierarchy may allow users (or the XAI methods) to select explanations that match specific needs. Technical challenges include identifying and minimizing the inaccuracies of synthesized concept-based explanations (see Section 3.2.3), which could be tackled by introducing quality metrics for concept-based explanations. Likewise, applying concept synthesis in different domains and applications is another sub-problem. This might be solved by personalization, as described above.

#### 3.5.3. Addressing explanations divorced from reality

The complexity of information flows in increasingly complex AI systems can result in what we call a 'reality drift'. As AI systems become smarter, their decision-making becomes more intricate. AI systems might start using concepts impossible to convey to humans [235,236]. This means that the concepts humans use to understand the world might no longer suffice to describe reality in a meaningful and useful way [237]. Consequently, the workings of such systems would become necessarily incomprehensible to us, and the utility of explanations, which are increasingly divorced from reality, may be questionable. To bridge this gap, one might initially think that new concepts are needed that both humans and machines can use. However, there are differences in how humans and machines store and process information, making the success of this approach uncertain. In general, explanations provided by AI systems may seem plausible to humans but could be detached from actual reality. This raises important questions about the usefulness of explainability in ensuring AI safety, especially when dealing with highly complex AI systems that are hard to decipher [137,238].

*Solution ideas.* To address the gap between explanations and reality, one potential solution involves engaging society and implementing regulations that ensure that someone can be held accountable for the performance of AI systems, especially in critical situations.[3] To achieve this, it is crucial to ensure that explanations are falsifiable (see Section 3.8.2). Selecting explanation forms based on their falsifiability enables market and legal control over the types of AI systems used. Future systems should also tackle the uncertainty in modelling explanations by incorporating ontological information. There are three research directions to consider from here. The first direction explores the proof of the (non)existence of specific concept properties, such as gap size, robustness, simplicity, and estimability, to mention a few. The second direction focuses on developing adaptive ontology-generation methods to track evolving reality. These methods create adaptable and robust ontologies with computational properties that respect the limitations of human understanding. Basically, this approach would enhance the relevance of explainability in the context of reality drift. The third direction is sociological and deals with updating ontologies

[3] See, for instance, the EU AI Act: https://artificialintelligenceact.eu/the-act/ (Last access: January 5th, 2024).

within society after adaptations. In addition, when seeking adversarial robustness, it is preferable to establish protectorates at the highest possible level of abstraction in the ontology generation process for computational efficiency [69]. This comprehensive approach aims to improve the alignment of AI explainability with the dynamic nature of real-world scenarios.

### 3.5.4. *Uncovering causality for actionable explanations*

Causality is arguably among the most desired properties when constructing a model from data. In this regard, uncovering causal connections learned through a model via explanations is a fundamental hope associated with XAI [48,49,239]. However, off-the-shelf posthoc XAI methods fail to disentangle the correlation represented in the learned model from the causation between observed variables and predictions, making it questionable whether received explanations are suitable for guiding people's actions [240]. Explanations based on correlations can hinder decision-making when a model's outputs contain essential information for action, for instance, the probability of failure of a production facility in industrial forecasting. Actionable and action-guiding explanations derived from causal models are needed in the real world, significantly when decisions may affect people. To address this issue, counterfactual generation methods for ML methods have garnered attention [32]. Contrary to most XAI approaches, counterfactuals attempt to answer why a black box model leads to a particular prediction by helping users understand what would need to change at its output to achieve a desired result [241]. In this answer, several desired properties should be met: proximity, plausibility, sparsity, diversity, and feasibility [32]. However, most works only regard a subset of these when producing counterfactuals, ignoring challenging issues. These include the provision of plausibility guarantees in highly complex data or generating diverse samples for largely parametric generative models prone to fall into single modalities. Furthermore, there are few causal approaches for XAI since finding causal relationships from observational data is extremely difficult to achieve [242].

*Solution ideas.* To tackle the need for actionable explanations, technological advancements in AI, such as large generative models, can open new opportunities in counterfactual explanations. One assumption is that such advancements can endow the produced counterfactuals with some desired properties for explanations such as proximity, plausibility, sparsity, diversity, and feasibility. This has been approached recently in [243], where counterfactuals are produced using an optimization problem formulated over conditional GANs comprising three different objectives: one related to plausibility, another one to sparsity, and a third one that relates to feasibility. With initial explorations of diffusion-based counterfactuals being reported in recent research [244, 245], questions such as how to sample-efficiently diversify adversarial outputs produced by these models will be interesting. Another direction worth exploring is how to connect causal graphs, relating each input of the model with its output, particularly in high-dimensional data. Most expert knowledge is represented in terms of entities and semantic relationships that inherently encode cause–effect links, as in knowledge bases. The goal in this context is to construct causal graphs automatically for models that do not necessarily operate on concepts or entities but instead on raw data. A potential solution is interfacing learning algorithms with symbolic knowledge about how the world behaves so that explanations for models grounded on such established causal links are endowed with the sought actionability.

## 3.6. *Supporting the multi-dimensionality of explainability*

Another class of challenges for XAI is that explanations are multi-dimensional. In other words, explainability is a concept which has multiple facets and spans a variety of disciplines.

### 3.6.1. *Creating multi-faceted explanations*

For regulatory purposes, explanations should depend on and incorporate information about requirements for trustworthy AI systems. In some cases, there is no reason to spend much resources and effort explaining a decision made by an AI model if such a model is inaccurate, lawful, or unfair. In this line of thought, there have recently been calls stating that different dimensions of trustworthiness (for example, safety, fairness, accountability) should not be shown separately or individually to the audience of a given model or AI-based artefact. For this reason, explanations should be offered to humans by not only *explaining* the functioning (that means, traditional explainability) but also by *justifying* the reliability of the inferences of an AI system (for example, concerning technical robustness, safety, lawfulness, and fairness). If these properties are not considered, explanations will fail to calibrate users' trust correctly. This issue is particularly acute in situations of concept drifts or uncertainty.

*Solution ideas.* One approach to such multi-faceted explanations could involve developing trustworthiness metrics that encapsulate safety, fairness, and accountability dimensions. XAI can, then, be tailored to the trustworthiness level of the AI system, ensuring that less trustworthy models provide extensive *justifications* for their decisions while highly trustworthy systems may offer simple *explanations*. Trustworthiness thresholds can be established, triggering detailed explanations when the system falls below a predefined trustworthiness level. Furthermore, dynamic explanations that adapt to context, such as concept drift or uncertainty, can ensure that users' trust remains calibrated. A user-centric approach, allowing customization of explanation depth, would empower users to align the system's explanations with their specific needs. Transparency in the trustworthiness assessment process may enhance user confidence, and continuous monitoring and reporting offer the capability to adapt explanations as trustworthiness metrics change. This comprehensive strategy ensures that trustworthiness considerations are integral to the XAI process, leading to multi-faceted explanations. A complementary way to tackle the multidimensionality of explanation concerns its operationalization, which should be performed as it happens with other psychological constructs such as 'intelligence' or 'cognitive load' [220]. A solution is to propose a novel, inclusive definition of explainability that is modellable and that can be seen as a foundation to support the next generation of empirical-based research in the field. Modelability here means that the definition should contain high-level classes of notions and concepts that can be individually modelled, operationalized, and investigated empirically. The primary rationale behind this solution is practical, as the aim is to provide scholars with an operational characterization of explainability that can be parsed into sub-components that, in turn, can be individually modelled. This should motivate using quantitative methods for more excellent reproducibility, replicability and falsifiability.

### 3.6.2. *Enabling interdisciplinary work in XAI*

XAI is an interdisciplinary research field [6,7]. For example, through the collaboration of philosophers and computer scientists, XAI is envisioned to ensure the ethical use of AI [7]. However, it is often difficult for researchers of different disciplines to engage in joint research in XAI [4]. There are several reasons for that. First, the rapid increase of publications in XAI makes it difficult for researchers to keep up even with research in their discipline, such that they often cannot spare to engage with research of other disciplines (which also has an overwhelming number of publications) [4]. Furthermore, the different disciplines involved in XAI may have their own established usage of specific terms [178]. This can lead to confusion and difficulty adapting to different usage in XAI. Eventually, for terms for which there is no typical usage, different disciplines may establish their meanings, further leading to confusion.

*Solution ideas.* To counteract the information overload caused by a rapid increase in publications, a centralized knowledge-sharing platform for XAI could be established. This platform would curate and categorize relevant research from various disciplines, making it more manageable for scholars to access and engage with research from other disciplines. A crucial aspect of this collaborative platform would involve the development of standardized terminology and glossaries that unify the usage of key terms across disciplines. This would reduce confusion arising from varying interpretations of terminology, ensuring that researchers can communicate effectively and harmoniously. These terms should be updated periodically to accommodate evolving interdisciplinary insights. Moreover, fostering regular cross-disciplinary dialogues and forums can promote mutual understanding among researchers from different backgrounds. Dedicated workshops, conferences, and seminars for interdisciplinary work in XAI could facilitate knowledge exchange and encourage the development of shared research goals and methodologies. Additionally, funding agencies and institutions should incentivize and prioritize interdisciplinary research by offering grants, awards, and recognition for collaborative projects. This would motivate researchers to actively engage in cross-disciplinary efforts in XAI.

## 3.7. Adjusting XAI methods and explanations

Another class of challenges in XAI is related to adjusting explanations. With the diverse range of applications of AI systems, XAI methods must produce explanations that fit diverse stakeholders, domains, and goals. However, there is not yet enough research addressing these concerns.

### 3.7.1. Adjusting explanations to different stakeholders

Many stakeholders can require an explanation during the development, evaluation, and use of an AI system [6]. Each stakeholder brings their attitudes, preferences, aptitudes, abilities, and previous experiences that influence the kind of explanation they require. Designing and tailoring appropriate explanations for each stakeholder type, both in terms of *content* and *format and presentation*, is an ongoing challenge. For example, the same objective facts must be explained and tailored to the stakeholders' respective interests and objectives in the business context. A business person is usually primarily interested in the bottom line impact of an AI system, a technical person is interested in the process and implementation validity, and a financial person is interested in the cash flow. Adding to that mix, the different educational backgrounds and language used necessarily call for very different explanations for each of the three actors.

*Solution ideas.* Future work should investigate new ways to enrich explanations semantically by combining different types of XAI methods and utilizing additional information sources (for example, training data, ontologies, and other modalities). Ideas from personalizing DL models [246] and, more specifically, creating personalized explanations [247] can be helpful. Explanations could also be made interactive. Humans should be able to refine explanations through interaction, as recently advocated in the reinforcement learning community through reinforcement learning from human feedback [248,249].

### 3.7.2. Adjusting explanations to different domains

The domain and context in which explanations are consumed are critical. For example, explanations for using a self-driving car must differ significantly from those in a clinical decision support system. Each domain brings different assumptions, environments, expectations, and stakes. In self-driving cars, the details about the passengers are not as important, but adherence to regulation is paramount. In contrast, in a clinical situation, the patient details are crucial, but regulation does not (directly) prescribe decisions. Making each explanation universally applicable, precise, and compact means omitting many details that pertain to a domain, certainly sacrificing the explanation's effectiveness. Instead, we take the domain as indispensable and build on it. This makes meaningful explanations dependent on the domain whose peculiarities and context are built. In this line of thought, research is starting to emerge focused on distinguishing between high-stakes and low-stakes domains [6,250,251]. However, the influence of the domain in which an AI is used has not been fully explored.

*Solution ideas.* Domain-specific explanation models should be developed to cater to the unique requirements of various application areas. These models should incorporate relevant knowledge, terminology, and context-specific reasoning to provide meaningful explanations. Furthermore, research efforts should prioritize the development of guidelines and standards for context-aware explanations. These guidelines would provide a structured approach for AI developers to assess the use of and determine the most suitable explanation strategy.

### 3.7.3. Adjusting explanations to different goals

Another fundamental challenge is to adjust explanations to what they should achieve when being presented to a stakeholder. For instance, data scientists might want to develop an accurate data-driven model; a regulator might want to assess the fairness of an AI-assisted loan offer; or a loan applicant might want to know the reason behind a rejection [6,236,252]. An underlying assumption is that XAI seeks to achieve these desiderata by improving the mental model that a stakeholder has of an underlying AI-system [6,239,253,254]. However, the understanding required for each desideratum might differ, requiring tailored explanations.

*Solution ideas.* Adjusting explanations to different goals might not be possible without factoring in the stakeholders who have these goals. Accordingly, one approach is to employ a stakeholder-centric explanation strategy, recognizing that different stakeholders have distinct goals and information needs. For data scientists aiming to improve model accuracy, explanations can focus on technical model details, feature importance, and model performance metrics. Regulators seeking to assess fairness may require explanations related to fairness metrics, compliance with regulations, and potential bias sources. Meanwhile, end-users, such as loan applicants, often require clear, user-friendly explanations regarding AI-driven decisions, allowing them to understand the reasons behind outcomes. This stakeholder-specific tailoring ensures that the goals pursued with explainability are met effectively.

## 3.8. Mitigating the negative impact of XAI

Although XAI has noble goals, it might also have negative impacts that must be avoided or mitigated.

### 3.8.1. Mitigating failed support by XAI

In some domains, especially in the medical domain [255], the ineffective support by XAI can sometimes be harmful. This has been associated with the so-called 'white-box paradox' [256,257], which urges not to take the value of the support delivered by XAI systems for granted. There are two possible cases: failed and misleading explanations. The first case might occur when the advice from an AI system is correct, but the associated explanation fails to inform the decision maker positively. This can happen because the explanation is inappropriate or wrong, to appear faulty, irrelevant, or unclear to users [256]. In this situation, users might not accept the correct advice because of inadequate explanations. The second case is perhaps even worse and paradoxical; it occurs when the inference or advice of an AI system is wrong, but the synthesized explanations have a sufficient persuasive force for convincing users that such advice is correct. In this situation, users are misled and thus potentially prone to mistakes [258].

*Solution ideas.* A first option would be to detect and label failure situations appropriately and reliably. Then, one possible course of action would be not to provide users with any XAI support if this is deemed detrimental or irrelevant in a given setting, for instance, in radiological settings (see [255,257]). Another approach to mitigate failed support by XAI is to challenge the *oracular* conception of AI support. This conception assumes that AI outputs are judged based on moral categories like right and wrong, with AI-generated explanations serving as aids to help humans determine whether to trust the outputs. However, AI systems were initially conceived as *generative* and *persuasive* technologies [259], not oracular ones. This oracular nature can be characterized as an *alethic* nature, which assumes that machines can, and should, always state the truth [260]. Relaxing the expectation of truthfulness is feasible, especially when dealing with probabilistic outputs or uncertainty estimates from AI systems. To this end, we could introduce a third type of explanation, namely a *perorative* explanation, alongside the two traditional types of explanations provided by XAI systems: *motivational* and *justificative* explanations. In legal terms, peroration refers to the conclusion of a speech or argument, where a speaker summarizes their main points and seeks to persuade an audience of their position. By providing a set of possible explanations for different AI-based inferences, including opposing and contradictory ones, XAI systems enhance accountability among human decision-makers. This approach can be likened to a judicial process, where opposing parties present evidence and arguments before an impartial judge make the final decision, offering a more balanced perspective than the oracular approach [260–262].

### 3.8.2. *Devising criteria for the falsifiability of explanations*

Explanations are often requested to clarify issues such as accountability [203,235,263,264]. However, explanations might be wrong. In such a case, parties that did not contribute to a mistake could be held accountable. Unfortunately, there is a lack of clarity regarding when an explanation is incorrect and under what conditions it becomes falsifiable. Falsifiability is a critical element in introducing a commitment to the explanations provided by AI systems and understanding the potential consequences that follow. Without clear criteria for falsifiability, benchmarks for the correctness of explanations cannot be established, and it becomes challenging to hold AI practitioners accountable for the accuracy of their explanations. In some cases, practitioners may rely too heavily on intuition rather than rigorous methods regarding explainability. Therefore, the question of establishing what ground truths for explainability in benchmarks are and how they were produced are open questions. As a more ambitious goal, we may ask about the discriminability between differing plausible explanations and their ordering concerning quality and acceptability.

*Solution ideas.* Establishing criteria for falsifiability in XAI could draw inspiration from the philosophy of science and related research fields. One potential solution lies in adopting the Popperian notion of falsifiability as a cornerstone of empirical science that can serve as a guiding principle [265]. Within this framework, XAI could systematically integrate hypothesis testing and experimentation to subject explanations to rigorous empirical examination. In this line of thought, some researchers have advocated for a framework that promotes falsifiable research in the field of explainability, emphasizing the need for precision and rigour in evaluating and validating explanations [266]. Additionally, insights from epistemology and cognitive science can inform the development of standardized protocols for evaluating the correctness of explanations, drawing parallels with how empirical claims in the sciences are subjected to rigorous scrutiny. Furthermore, interdisciplinary collaboration between computer scientists, philosophers of science, ethicists, and cognitive psychologists can facilitate the development of a comprehensive framework that incorporates not only empirical falsifiability but also ethical considerations and cognitive principles. By anchoring XAI practices in well-established principles from the philosophy of science and related disciplines, it can pave the way for more robust, accountable, and scientifically grounded explanations within AI systems.

### 3.8.3. *Securing explanations from being abused by malicious human agents*

Explainability is an essential aspect of human coordination with machines [238,254,267,268]. This is especially true in the near term, where AI systems may not be competent enough for autonomous adversarial behaviour. XAI involves understanding how AI systems arrive at their inferences, decisions or recommendations and clearly explaining the logic and reasoning behind these outcomes. The effectiveness and adequacy of explainability as a tool for AI safety may be limited in specific scenarios [269]. For instance, AI systems in the hands of malicious human actors, can pose significant challenges to explainability [238] through manipulation and adversarial attacks. For example, employers may systematically discriminate against job applicants using socially misaligned ML models while serving borderline plausible explanations to avoid detection.

*Solution ideas.* The need for discriminating between different explanations ties directly to the falsifiability of explanations (see Section 3.8.2). Furthermore, concept-based explanations could help combat adversarial attacks, especially a recent form of such attacks that aim to trick both humans and classifiers [162]. For example, a malicious sample might be detectable by comparing a concept-based explanation of an adversarial sample with that of a non-adversarial sample. However, this is challenging because explanations can also be manipulated and used to trick or deceive [270]. Similarly, another application context includes forensic analysis, which aims to understand the concepts learned by a classifier [271]. Concept-based explanations could also be helpful for reflective learning from data [272], which means classifiers can be improved through processing explanations during training.

### 3.8.4. *Securing explanations from being abused by malicious superintelligent agents*

Explainability is an essential aspect of AI safety. Many of the challenges highlighted in the literature [137,273] and here already showcase fundamental limitations on the human ability to understand the behaviour of current AI systems. However, assuming no constraints on their design or physical limitations, future AI systems may become so competent that understanding them becomes fundamentally impossible. Exacerbating this issue, using formal verification to guarantee benign behaviour may not be viable due to unverifiability [274]. In independent domains where AI agents are non-adversarial, these issues are not of much concern. At worst, we are in a situation where cooperative AI agents fulfil our tasks for us, narrating comforting fairytales that make us content. However, when it comes to adversarial scenarios, the question is how much our assimilated explanatory concepts are adversarially robust through the existence of some computational protectorates that leave not many exposed loopholes. As the complexity and capabilities of AI agents increase, these agents may discover ways to deliberately fool people by exploiting the tension between the explanatory concepts that emerge from human capabilities, perception, and action and between those that complex agents can utilize. If such were the case, and humans rely only on explainability for safety, malignant gain by AI agents could be unbounded [137].

*Solution ideas.* Explainability can be an effective tool in ensuring the safety of AI systems, even in the long term, assuming that the problem of alignment between the technical capabilities of XAI methods and their application and utility for humans, in reality, is solved (see Section 3.5.3). The effectiveness and adequacy of explainability as a tool for AI safety may be limited in specific scenarios [269]. For this reason, explainability should be only one part of every safety toolkit as it has strengths and limitations that must be complemented in a portfolio of approaches. Work on building such a portfolio is welcome, as there is a growing need for it. This line of work is more long-term and can be solved only partially by constructive approaches. At the same time, the other part would be a restraint in building robust superintelligent systems without solid reasons to believe they are aligned with our values.

### 3.9. Improving the societal impact of XAI

Research on explainable AI and the derived methods, models and techniques used to create real-world applications can impact society.

#### 3.9.1. Facilitating originality attribution of AI-generated data and plagiarism detection

A special challenge for the explainability of novel generative models, which we think warrants to be mentioned separately, exists concerning originality attribution and plagiarism detection of AI-generated data. Concerning the problem of originality attribution, pieces of art produced by generative models have been recently taken to exhibit a similar level of creativity as humans. In particular, contests won with AI-generated art have stirred controversy concerning the intellectual property of the output of a model learned from third-party data [275,276]. Likewise, plagiarism detection is becoming central in LLMs that excel across different domains, for example, with ChatGPT [277]. The massive usage of these models to produce original textual content has disrupted the idea of plagiarism, as such content has been shown to easily evade mainstream tools for plagiarism detection. Thus, whether information biases the generative process, as, for example, with the prompt in a language-to-image stable diffusion model, is sufficiently original for intellectual property and author attribution claims remains an open question.

*Solution ideas.* Regarding originality attribution, a solution is reformulating the concept of authorship in these models both from the technical point of view and from the legal and regulatory perspectives. Explainability should play a part in future regulation, as explanations could reveal which instances or parts of the modelled data distribution are relevant for a given synthesized output of the model. Solutions should be devoted to understanding if generalization implies any form of plagiarism or whether it is a new form of inspiration, interfacing creative thoughts with original synthesized content. On the topic of plagiarism detection, efforts have been made recently to determine whether the content produced by AI models is artificially generated, proposing the inclusion of tailored tokens, for example, *watermarking* [278], in the produced content [279]. Explainability techniques will be relevant in determining which learning instances were more influential in producing a given outcome.

#### 3.9.2. Facilitating the right to be forgotten

Large-scale generative models require tons of data to fine-tune their trainable parameters, which often account for several hundreds of terabytes in size. Such a huge data substrate may clash with a fundamental right in data governance: the right to be ignored or forgotten by data-driven models. While the interest in the *machine unlearning* paradigm [280] has been on the rise [281], it is unclear how to efficiently ensure that data owned by a particular user is *unlearned* by a given large-scale generative model, so that it can be ensured that no instance like that of the user claiming their right to be forgotten will be produced when the model is queried.

*Solution ideas.* The right to be forgotten could be supported by similarity-based explanations and incrementally retraining the model to avoid sampling around the part of the subspaces close to the forbidden data. XAI can also play a pivotal role by explaining model decisions and revealing which data points influenced those decisions. This explainability can empower users to identify the data instances that relate to them, enabling them to exercise their right to be forgotten. Additionally, XAI can aid in auditing and verifying that the unlearning process is carried out effectively, reassuring users that their privacy rights are upheld. Humans should be allowed to verify that a generative model does not learn from them.

#### 3.9.3. Addressing the power imbalance between individuals and companies

A significant issue in XAI is that the efforts in facilitating more comprehension of AI systems often are not enough to mitigate or even address the problem of unfair AI systems that exacerbate the societal power imbalance between individuals and companies using AI systems [236,282–284]. In other words, explaining the logic of an algorithm might be essential to empowering individuals to understand how to react to unreasonable AI-driven systems, especially when those systems take automated decisions that can legally or similarly significantly affect individuals. Still, explainability is often hard to achieve in practice and limited in scope. The capability to understand ‘why’ a particular automated system followed a path from some inputs to some outputs may not be enough to empower individuals in case such a path was logically correct but legally or ethically disputable. Explanations are not enough if they are not accompanied by accountable systems of *contestability* [285] and by justificatory statements that could prove why the ‘path’ from inputs to outputs is not only logically correct but also non-discriminatory, non-manipulative, non-illegal, non-unfair [286,287]. Therefore, only acting at the level of the individual ‘reactions’ to the outputs of automated decision-making, including understandability, contestability, and justifiability, fails to completely address the main societal and ethical problems behind unfair and untrustworthy AI. The XAI community should shift its focus to tackle the power imbalance between AI developers or controllers and those affected by AI [288,289]. The power imbalance is a structural problem, but the way AI increases such an imbalance cannot be faced only by more explainability. There is a broader problem of under-representation, hidden discrimination, and lack of accountability [290]. Current XAI methods answer this issue, but they can address only a tiny part of the problem [291].

*Solution ideas.* To address the power imbalance between individuals and companies in the realm of XAI, a new approach to designing future AI systems via XAI methods can include participative design, where impacted stakeholders are invited into the decision-making process [290,292]. There are different modalities of participative approach to AI design, but an essential consideration is the participative impact assessment [293]. Vulnerable impacted stakeholders should be included, through an open and circular approach, in the key value-sensitive decisions in the AI design [294]. Following the example of environmental impact assessment or workers’ participation in business decision-making [295], there are different ways in which digital users, individually, in groups or through representatives, can participate in the data processing decision-making or the design of data-driven technologies.

## 4. A novel manifesto

We conclude this article by presenting a manifesto for XAI. This manifesto aims to define and briefly describe the open challenges scholars in the field face. It includes propositions governing independent scientific research. The Manifesto is a mechanism for shaping our shared visions about science in the field of XAI, and it is the outcome of the engagement of diverse expertise and different experiences by its authors.

1. **Creating Explanations for New Types of AI:** To create explanations for generative models (for example, LLMs) and for distributed and collaborative learning.
2. **Improving (and Augmenting) Current XAI Methods**: To augment and improve attribution methods and concept-based learning algorithms, remove artefacts in synthesis-based explanations, and create robust explanations.
3. **Clarifying the Use of Concepts in XAI:** To clarify the main concepts in XAI and its relationship to trustworthiness and to find a useful account of understanding.

4. **Evaluating XAI Methods and Explanations:** To facilitate the human evaluation of explanations, create an evaluation framework for XAI methods, and overcome limitations of studies with humans.
5. **Supporting the Human-Centeredness of Explanations:** To create human-understandable explanations, facilitate explainability with concept-based explanations, address explanations divorced from reality, and uncover causality for actionable explanations.
6. **Supporting the Multi-Dimensionality of Explainability:** To create multi-faceted explanations and enable interdisciplinary work in XAI.
7. **Adjusting XAI Methods and Explanations:** To adjust explanations to different stakeholders, domains, and goals.
8. **Mitigating the Negative Impact of XAI:** To adjust explanations to different stakeholders, devise criteria for the falsifiability of explanations, and secure explanations from being abused by malicious human or superintelligent agents.
9. **Improving the Societal Impact of XAI:** To facilitate the originality attribution of AI-generated data and plagiarism detection, support the right to be forgotten, and address the power imbalance between individuals and companies.

We believe working together as a community will lead to more productive and up-to-date work, increase reliability and enhance falsifiability. The spirit of close collaboration, even among scholars with different scientific backgrounds and focused on specific disciplines, along with the respect and the willingness to build on each other's work, will undoubtedly inspire more scholars to join us in advancing XAI as a field. As a conclusion and an invitation to reflect on our manifesto, we return to the criticisms concerning the partly unsuccessful development of XAI research that we highlighted in Section 1. Despite remarkable theoretical advances leading the incredible momentum of this field, we as a community should recognize that we still have a long way to go to improve and realize the practical usability of XAI from various perspectives, many of which have already been mentioned in this manifesto. Our community must work to ensure that XAI becomes an essential tool in the design of responsible AI systems driven by current regulations (for example, the EU AI Act). This manifesto is a genuine call for consensus and an exciting opportunity for shaping the future of AI-based systems for the benefit of human society.

## CRediT authorship contribution statement

**Luca Longo:** Conceptualization, Funding acquisition, Investigation, Methodology, Project administration, Validation, Writing – original draft, Writing – review & editing. **Mario Brcic:** Conceptualization, Investigation, Methodology, Validation, Writing – review & editing. **Federico Cabitza:** Conceptualization, Investigation, Methodology, Validation, Writing – review & editing. **Jaesik Choi:** Conceptualization, Investigation, Methodology, Validation, Writing – review & editing. **Roberto Confalonieri:** Conceptualization, Investigation, Methodology, Validation, Writing – review & editing. **Javier Del Ser:** Conceptualization, Investigation, Methodology, Validation, Writing – review & editing. **Riccardo Guidotti:** Conceptualization, Investigation, Methodology, Validation, Writing – review & editing. **Yoichi Hayashi:** Conceptualization, Investigation, Methodology, Validation, Writing – review & editing. **Francisco Herrera:** Conceptualization, Investigation, Methodology, Validation, Writing – review & editing. **Andreas Holzinger:** Conceptualization, Investigation, Methodology, Validation, Writing – review & editing. **Richard Jiang:** Conceptualization, Investigation, Methodology, Validation, Writing – review & editing. **Hassan Khosravi:** Conceptualization, Investigation, Methodology, Validation, Writing – review & editing. **Freddy Lecue:** Conceptualization, Investigation, Methodology, Validation, Writing – review & editing. **Gianclaudio Malgieri:** Conceptualization, Investigation, Methodology, Validation, Writing – review & editing. **Andrés Páez:** Conceptualization, Investigation, Methodology, Validation, Writing – review & editing. **Wojciech Samek:** Conceptualization, Investigation, Methodology, Validation, Writing – review & editing. **Johannes Schneider:** Conceptualization, Investigation, Methodology, Validation, Writing – review & editing. **Timo Speith:** Conceptualization, Investigation, Methodology, Validation, Writing – review & editing. **Simone Stumpf:** Conceptualization, Investigation, Methodology, Validation, Writing – review & editing.

## 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.

## Acknowledgements

T. Speith acknowledges funding support provided by the Volkswagen Foundation grants AZ 9B830, AZ 98509, and AZ 98514 "Explainable Intelligent Systems" (EIS; https://explainable-intelligent.systems) and by the DFG grant 389792660 as part of TRR 248 (https://perspicuous-computing.science).

J. Del Ser acknowledges funding support from the 'Centro para el Desarrollo Tecnologico Industrial'(CDTI), by the European Union (AI4ES, grant no. CER-20211030), and by the Basque Government (MATHMODE, ref. IT1456-22).

F. Herrera acknowledges funding support by the Spanish grant PID2020-119478GB-I00 funded by MCIN/AEI/10.13039/501100011033.

R. Confalonieri acknowledges funding support from the 'Neurosymbolic XAI' project (BIRD231830).

J. Choi is supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. 2022-0-00984).

R. Guidotti acknowledges funding support provided by the funding schemes ERC-2018-ADG G.A. 834756 "XAI: Science and technology for the eXplanation of AI decision making" (https://xai-project.eu/), "INFRAIA-01-2018-2019 – Integrating Activities for Advanced Communities", G.A. 871042, "SoBigData++: European Integrated Infrastructure for Social Mining and Big Data Analytics" (http://www.sobigdata.eu), by the European Commission under the NextGeneration EU programme National Recovery and Resilience Plan (Piano Nazionale di Ripresa e Resilienza, PNRR) – Project: "SoBigData.it – Strengthening the Italian RI for Social Mining and Big Data Analytics" – Prot. IR0000013 – Avviso n. 3264 del 28/12/2021, and M4C2 - Investimento 1.3, Partenariato Esteso PE00000013 - "FAIR - Future Artificial Intelligence Research" - Spoke 1 "Human-centered AI", and by the Italian Project Fondo Italiano per la Scienza FIS00001966 MIMOSA.

A. Holzinger acknowledges funding support by the XAI Project of the Austrian Science Fund FWF P-32554.

F. Cabitza acknowledges funding support provided by the Italian project PRIN 2022 PNRR 1409/22, InXAID - Interaction with eXplainable Artificial Intelligence in (medical) Decision making. CUP: H53D23008090001 funded by MUR.

W. Samek acknowledges funding support from the German Research Foundation (DFG) as research unit DeSBi (KI-FOR 5363).

## References

[1] W. Swartout, C. Paris, J. Moore, Explanations in knowledge systems: Design for explainable expert systems, IEEE Expert 6 (3) (1991) 58–64.

[2] C.L. Paris, Generation and explanation: Building an explanation facility for the explainable expert systems framework, in: Natural Language Generation in Artificial Intelligence and Computational Linguistics, Springer, 1991, pp. 49–82.

[3] 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, URL https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/widm.1391.

[4] T. Speith, A review of taxonomies of explainable artificial intelligence (XAI) methods, in: C. Isbell, S. Lazar, A. Oh, A. Xiang (Eds.), Proceedings of the 5th ACM Conference on Fairness, Accountability, and Transparency, Association for Computing Machinery, New York, NY, USA, 2022, pp. 2239–2250, http://dx.doi.org/10.1145/3531146.3534639.

[5] G. Schwalbe, B. Finzel, A comprehensive taxonomy for explainable artificial intelligence: a systematic survey of surveys on methods and concepts, Data Min. Knowl. Discov. (2023) 1–59.

[6] M. Langer, D. Oster, T. Speith, H. Hermanns, L. Kästner, E. Schmidt, A. Sesing, K. Baum, What do we want from Explainable Artificial Intelligence (XAI)?–A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research, Artificial Intelligence 296 (2021) 103473.

[7] M. Langer, K. Baum, K. Hartmann, S. Hessel, T. Speith, J. Wahl, Explainability auditing for intelligent systems: A rationale for multi-disciplinary perspectives, in: T. Yue, M. Mirakhorli (Eds.), 29th IEEE International Requirements Engineering Conference Workshops, in: REW 2021, IEEE, Piscataway, NJ, USA, 2021, pp. 164–168, http://dx.doi.org/10.1109/REW53955.2021.00030.

[8] S. Ali, T. Abuhmed, S. El-Sappagh, K. Muhammad, J.M. Alonso-Moral, R. Confalonieri, R. Guidotti, J. Del Ser, N. Díaz-Rodríguez, F. Herrera, Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence, Inf. Fusion (2023) 101805.

[9] L. Cao, Ai in finance: challenges, techniques, and opportunities, ACM Comput. Surv. 55 (3) (2022) 1–38.

[10] R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, N. Elhadad, Intelligible models for HealthCare: Predicting pneumonia risk and hospital 30-day readmission, in: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '15, Association for Computing Machinery, New York, NY, USA, 2015, pp. 1721–1730, http://dx.doi.org/10.1145/2783258.2788613.

[11] AI High-Level Expert Group, Ethics guidelines for trustworthy AI, 2019, B-1049 Brussels. URL https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai.

[12] T. Freiesleben, G. König, Dear XAI community, we need to talk!, in: L. Longo (Ed.), Explainable Artificial Intelligence, Springer Nature Switzerland, Cham, 2023, pp. 48–65.

[13] L. Kästner, M. Langer, V. Lazar, A. Schomäcker, T. Speith, S. Sterz, On the relation of trust and explainability: Why to engineer for trustworthiness, in: T. Yue, M. Mirakhorli (Eds.), 29th IEEE International Requirements Engineering Conference Workshops, in: REW 2021, IEEE, Piscataway, NJ, USA, 2021, pp. 169–175, http://dx.doi.org/10.1109/REW53955.2021.00031.

[14] A. Papenmeier, G. Englebienne, C. Seifert, How model accuracy and explanation fidelity influence user trust, 2019, arXiv preprint arXiv:1907.12652.

[15] X. Huang, J. Marques-Silva, From robustness to explainability and back again, 2023, arXiv preprint arXiv:2306.03048.

[16] J. Marques-Silva, X. Huang, Explainability is NOT a game, 2023, arXiv preprint arXiv:2307.07514.

[17] T. Miller, P. Howe, L. Sonenberg, Explainable AI: Beware of inmates running the asylum. or: How I learnt to stop worrying and love the social and behavioural sciences, in: D.W. Aha, T. Darrell, M. Pazzani, D. Reid, C. Sammut, P. Stone (Eds.), Proceedings of the IJCAI 2017 Workshop on Explainable Artificial Intelligence, IJCAI, Santa Clara County, CA, USA, 2017, pp. 36–42, arXiv:1712.00547.

[18] 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, URL https://www.sciencedirect.com/science/article/pii/S1566253519308103.

[19] K. Haresamudram, S. Larsson, F. Heintz, Three levels of AI transparency, Computer 56 (2) (2023) 93–100, http://dx.doi.org/10.1109/MC.2022.3213181.

[20] J. Zerilli, Explaining machine learning decisions, Philos. Sci. 89 (1) (2022) 1–19.

[21] L. Chazette, K. Schneider, Explainability as a non-functional requirement: challenges and recommendations, Requir. Eng. 25 (4) (2020) 493–514, http://dx.doi.org/10.1007/s00766-020-00333-1.

[22] M.A. Köhl, K. Baum, D. Bohlender, M. Langer, D. Oster, T. Speith, Explainability as a non-functional requirement, in: D.E. Damian, A. Perini, S. Lee (Eds.), IEEE 27th International Requirements Engineering Conference, in: RE 2019, IEEE, Piscataway, NJ, USA, 2019, pp. 363–368, http://dx.doi.org/10.1109/RE.2019.00046.

[23] 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.

[24] A. Páez, The pragmatic turn in explainable artificial intelligence (XAI), Minds Mach. 29 (3) (2019) 441–459, http://dx.doi.org/10.1007/s11023-019-09502-w.

[25] S. Bruckert, B. Finzel, U. Schmid, The next generation of medical decision support: A roadmap toward transparent expert companions, Frontiers Artificial Intelligence 3 (2020) 507973.

[26] V. Arya, R.K.E. Bellamy, P.-Y. Chen, A. Dhurandhar, M. Hind, S.C. Hoffman, S. Houde, Q.V. Liao, R. Luss, A. Mojsilović, S. Mourad, P. Pedemonte, R. Raghavendra, J. Richards, P. Sattigeri, K. Shanmugam, M. Singh, K.R. Varshney, D. Wei, Y. Zhang, One explanation does not fit all: A toolkit and taxonomy of ai explainability techniques, 2021, arXiv:1909.03012.

[27] G. Vilone, L. Longo, Notions of explainability and evaluation approaches for explainable artificial intelligence, Inf. Fusion 76 (2021) 89–106.

[28] K. Sokol, P. Flach, Explainability fact sheets: A framework for systematic assessment of explainable approaches, in: M. Hildebrandt, C. Castillo, L.E. Celis, S. Ruggieri, L. Taylor, G. Zanfir-Fortuna (Eds.), Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, in: FAT* 2020, Association for Computing Machinery, New York, NY, USA, 2020, pp. 56–67, http://dx.doi.org/10.1145/3351095.3372870.

[29] L. Chazette, W. Brunotte, T. Speith, Exploring explainability: A definition, a model, and a knowledge catalogue, in: J. Cleland-Huang, A. Moreira, K. Schneider, M. Vierhauser (Eds.), IEEE 29th International Requirements Engineering Conference, in: RE 2021, IEEE, Piscataway, NJ, USA, 2021, pp. 197–208, http://dx.doi.org/10.1109/RE51729.2021.00025.

[30] L. Weber, S. Lapuschkin, A. Binder, W. Samek, Beyond explaining: Opportunities and challenges of XAI-based model improvement, Inf. Fusion 92 (2023) 154–176.

[31] F. Bodria, F. Giannotti, R. Guidotti, F. Naretto, D. Pedreschi, S. Rinzivillo, Benchmarking and survey of explanation methods for black box models, Data Min. Knowl. Discov. (2023) 1–60.

[32] R. Guidotti, Counterfactual explanations and how to find them: literature review and benchmarking, Data Min. Knowl. Discov. (2022) 1–55.

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

[34] Y. Mei, Q. Chen, A. Lensen, B. Xue, M. Zhang, Explainable artificial intelligence by genetic programming: A survey, IEEE Trans. Evol. Comput. 27 (3) (2023) 621–641, http://dx.doi.org/10.1109/TEVC.2022.3225509.

[35] D. Minh, H.X. Wang, Y.F. Li, T.N. Nguyen, Explainable artificial intelligence: a comprehensive review, Artif. Intell. Rev. (2022) 1–66.

[36] A. Theissler, F. Spinnato, U. Schlegel, R. Guidotti, Explainable AI for time series classification: A review, taxonomy and research directions, IEEE Access 10 (2022) 100700–100724, http://dx.doi.org/10.1109/ACCESS.2022.3207765.

[37] G. Yang, Q. Ye, J. Xia, Unbox the black-box for the medical explainable AI via multi-modal and multi-centre data fusion: A mini-review, two showcases and beyond, Inf. Fusion 77 (2022) 29–52.

[38] J.E. Zini, M. Awad, On the explainability of natural language processing deep models, ACM Comput. Surv. 55 (5) (2022) 1–31.

[39] 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.

[40] A. Heuillet, F. Couthouis, N. Díaz-Rodríguez, Explainability in deep reinforcement learning, Knowl.-Based Syst. 214 (2021) 106685.

[41] 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) 11, http://dx.doi.org/10.1016/j.jbi.2020.103655.

[42] S. Mohseni, N. Zarei, E.D. Ragan, A multidisciplinary survey and framework for design and evaluation of explainable AI systems, ACM Trans. Interact. Intell. Syst. (TiiS) 11 (3–4) (2021) 1–45.

[43] 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.

[44] 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.

[45] G. Vilone, L. Longo, Classification of explainable artificial intelligence methods through their output formats, Mach. Learn. Knowl. Extr. 3 (3) (2021) 615–661, http://dx.doi.org/10.3390/make3030032, URL https://www.mdpi.com/2504-4990/3/3/32.

[46] 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.

[47] E. Tjoa, C. Guan, A survey on explainable artificial intelligence (xai): Toward medical xai, IEEE Trans. Neural Netw. Learn. Syst. 32 (11) (2020) 4793–4813.

[48] D.V. Carvalho, E.M. Pereira, J.S. Cardoso, Machine learning interpretability: A survey on methods and metrics, Electronics 8 (8) (2019) 832.

[49] A. Adadi, M. Berrada, Peeking inside the black-box: a survey on explainable artificial intelligence (XAI), IEEE Access 6 (2018) 52138–52160.

[50] 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) 1–42, http://dx.doi.org/10.1145/3236009.

[51] A. Rawal, J. McCoy, D.B. Rawat, B.M. Sadler, R.S. Amant, Recent advances in trustworthy explainable artificial intelligence: Status, challenges, and perspectives, IEEE Trans. Artif. Intell. 3 (6) (2021) 852–866.

[52] F. Hinder, B. Hammer, Counterfactual explanations of concept drift, 2020, arXiv preprint arXiv:2006.12822.

[53] A. Khan, M.t. Thij, A. Wilbik, Vertical federated learning: A structured literature review, 2022, arXiv preprint arXiv:2212.00622.

[54] S.-K. Yeom, P. Seegerer, S. Lapuschkin, A. Binder, S. Wiedemann, K.-R. Müller, W. Samek, Pruning by explaining: A novel criterion for deep neural network pruning, Pattern Recognit. 115 (2021) 107899.

[55] M.T. Ribeiro, S. Singh, C. Guestrin, "Why should I trust you?": Explaining the predictions of any classifier, in: C. Aggarwal, B. Krishnapuram, R. Rastogi, D. Shen, M. Shah, A. Smola (Eds.), Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, in: KDD 2016, Association for Computing Machinery, New York, NY, USA, 2016, pp. 1135–1144, http://dx.doi.org/10.1145/2939672.2939778.

[56] S.M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, in: I. Guyon, U.V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), in: Advances in Neural Information Processing Systems, vol. 30, Curran Associates, Inc., 2017, pp. 4768–4777, URL https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf.

[57] D. Garreau, U. Luxburg, Explaining the explainer: A first theoretical analysis of LIME, in: International Conference on Artificial Intelligence and Statistics, PMLR, 2020, pp. 1287–1296.

[58] 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.

[59] S. Lapuschkin, S. Wäldchen, A. Binder, G. Montavon, W. Samek, K.-R. Müller, Unmasking Clever Hans predictors and assessing what machines really learn, Nature Commun. 10 (1) (2019) 1096.

[60] R. Geirhos, J.-H. Jacobsen, C. Michaelis, R. Zemel, W. Brendel, M. Bethge, F.A. Wichmann, Shortcut learning in deep neural networks, Nat. Mach. Intell. 2 (11) (2020) 665–673, http://dx.doi.org/10.1038/s42256-020-00257-z.

[61] T. Speith, How to evaluate explainability – a case for three criteria, in: E. Knauss, G. Mussbacher, C. Arora, M. Bano, J.-G. Schneider (Eds.), Proceedings of the 30th IEEE International Requirements Engineering Conference Workshops, in: REW 2022, IEEE, Piscataway, NJ, USA, 2022, pp. 92–97, http://dx.doi.org/10.1109/REW56159.2022.00024.

[62] S. Lapuschkin, A. Binder, K.-R. Müller, W. Samek, Understanding and comparing deep neural networks for age and gender classification, in: Proceedings of the IEEE International Conference on Computer Vision Workshops, 2017, pp. 1629–1638.

[63] L. Grinsztajn, E. Oyallon, G. Varoquaux, Why do tree-based models still outperform deep learning on typical tabular data? in: S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, A. Oh (Eds.), in: Advances in Neural Information Processing Systems, vol. 35, Curran Associates, Inc., 2022, pp. 507–520, URL https://proceedings.neurips.cc/paper_files/paper/2022/file/0378c7692da36807bdec87ab043cdadc-Paper-Datasets_and_Benchmarks.pdf.

[64] 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.

[65] B. Crook, M. Schlüter, T. Speith, Revisiting the performance-explainability trade-off in explainable artificial intelligence (XAI), in: F. Dalpiaz, J. Horkoff, K. Schneider (Eds.), Proceedings of the 31st IEEE International Requirements Engineering Conference Workshops, IEEE, Piscataway, NJ, USA, 2023, pp. 316–324.

[66] L. Rokach, Decision forest: Twenty years of research, Inf. Fusion 27 (2016) 111–125.

[67] J. Hatwell, M.M. Gaber, R.M.A. Azad, CHIRPS: Explaining random forest classification, Artif. Intell. Rev. 53 (2020) 5747–5788.

[68] J. Fürnkranz, T. Kliegr, H. Paulheim, On cognitive preferences and the plausibility of rule-based models, Mach. Learn. 109 (4) (2020) 853–898.

[69] D.C. Krakauer, Unifying complexity science and machine learning, Front. Complex Syst. 1 (2023) http://dx.doi.org/10.3389/fcpxs.2023.1235202, URL https://www.frontiersin.org/articles/10.3389/fcpxs.2023.1235202.

[70] A. Fernandez, F. Herrera, O. Cordon, M.J. del Jesus, F. Marcelloni, Evolutionary fuzzy systems for explainable artificial intelligence: Why, when, what for, and where to? IEEE Comput. Intell. Mag. 14 (1) (2019) 69–81.

[71] X. Huang, J. Marques-Silva, From decision trees to explained decision sets, in: 26th European Conference on Artificial Intelligence, ECAI 2023, Vol. 372, IOS Press, 2023, pp. 1100–1108.

[72] X. Huang, A. Khetan, M. Cvitkovic, Z. Karnin, Tabtransformer: Tabular data modeling using contextual embeddings, 2020, arXiv preprint arXiv:2012.06678.

[73] Y. Gorishniy, I. Rubachev, V. Khrulkov, A. Babenko, Revisiting deep learning models for tabular data, Adv. Neural Inf. Process. Syst. 34 (2021) 18932–18943.

[74] S.Ö. Arik, T. Pfister, Tabnet: Attentive interpretable tabular learning, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 35, 2021, pp. 6679–6687, (8).

[75] S. Abnar, W. Zuidema, Quantifying attention flow in transformers, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 4190–4197, http://dx.doi.org/10.18653/v1/2020.acl-main.385, URL https://aclanthology.org/2020.acl-main.385.

[76] A. Ali, T. Schnake, O. Eberle, G. Montavon, K.-R. Müller, L. Wolf, XAI for transformers: Better explanations through conservative propagation, in: International Conference on Machine Learning, PMLR, 2022, pp. 435–451.

[77] M. Deb, B. Deiseroth, S. Weinbach, P. Schramowski, K. Kersting, AtMan: Understanding transformer predictions through memory efficient attention manipulation, 2023, arXiv preprint arXiv:2301.08110.

[78] S.M.V.H. Reduan Achtibat, M. Dreyer, A. Jain, T. Wiegand, S. Lapuschkin, W. Samek, Attnlrp: attention-aware layer-wise relevance propagation for transformers, arXiv:2402.05602 (2024) https://arxiv.org/abs/2402.05602.

[79] F. Lécué, On the role of knowledge graphs in explainable AI, Semant. Web 11 (1) (2020) 41–51, http://dx.doi.org/10.3233/SW-190374.

[80] T. Speith, M. Langer, A new perspective on evaluation methods for explainable artificial intelligence (XAI), in: F. Dalpiaz, J. Horkoff, K. Schneider (Eds.), Proceedings of the 31st IEEE International Requirements Engineering Conference Workshops, IEEE, Piscataway, NJ, USA, 2023, pp. 325–331.

[81] K. Čyras, A. Rago, E. Albini, P. Baroni, F. Toni, Argumentative XAI: A survey, in: Z.-H. Zhou (Ed.), Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21, International Joint Conferences on Artificial Intelligence Organization, 2021, pp. 4392–4399, http://dx.doi.org/10.24963/ijcai.2021/600, Survey Track.

[82] K. Baum, H. Hermanns, T. Speith, From machine ethics to machine explainability and back, in: M. Charles, D.I. Diochnos, J. Dix, F. Hoffman, G.R. Simari (Eds.), International Symposium on Artificial Intelligence and Mathematics, International Symposium on Artificial Intelligence and Mathematics, Fort Lauderdale, FL, USA, 2018, pp. 1–8.

[83] K. Baum, H. Hermanns, T. Speith, Towards a framework combining machine ethics and machine explainability, in: B. Finkbeiner, S. Kleinberg (Eds.), Proceedings of the 3rd Workshop on Formal Reasoning about Causation, Responsibility, and Explanations in Science and Technology, Electronic Proceedings in Theoretical Computer Science, Sydney, NSW, AU, 2018, pp. 34–49, http://dx.doi.org/10.4204/EPTCS.286.4.

[84] A. Vassiliades, N. Bassiliades, T. Patkos, Argumentation and explainable artificial intelligence: a survey, Knowl. Eng. Rev. 36 (2021) e5.

[85] L. Longo, Argumentation for knowledge representation, conflict resolution, defeasible inference and its integration with machine learning, in: Machine Learning for Health Informatics: State-of-the-Art and Future Challenges, Springer, 2016, pp. 183–208.

[86] Z. Zeng, C. Miao, C. Leung, J.J. Chin, Building more explainable artificial intelligence with argumentation, in: S.A. McIlraith, K.Q. Weinberger (Eds.), Proceedings of the 32nd AAAI Conference on Artificial Intelligence, the 30th Innovative Applications of Artificial Intelligence Conference, and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence, AAAI Press, Palo Alto, CA, USA, 2018, pp. 8044–8046, http://dx.doi.org/10.1609/aaai.v32i1.11353.

[87] P. Baroni, M. Caminada, M. Giacomin, An introduction to argumentation semantics, Knowl. Eng. Rev. 26 (4) (2011) 365–410.

[88] L. Rizzo, L. Longo, Inferential models of mental workload with defeasible argumentation and non-monotonic fuzzy reasoning: a comparative study, in: Proceedings of the 2nd Workshop on Advances in Argumentation in Artificial Intelligence, Co-Located with XVII International Conference of the Italian Association for Artificial Intelligence, AI$^3$@AI*IA 2018, 20-23 November 2018, Trento, Italy, 2018, pp. 11–26.

[89] L. Rizzo, L. Majnaric, L. Longo, A comparative study of defeasible argumentation and non-monotonic fuzzy reasoning for elderly survival prediction using biomarkers, in: AI* IA 2018–Advances in Artificial Intelligence: XVIIth International Conference of the Italian Association for Artificial Intelligence, Trento, Italy, November 20–23, 2018, Proceedings 17, Springer, 2018, pp. 197–209.

[90] L. Longo, L. Rizzo, P. Dondio, Examining the modelling capabilities of defeasible argumentation and non-monotonic fuzzy reasoning, Knowl.-Based Syst. 211 (2021) 106514.

[91] S. S Band, A. Yarahmadi, C.-C. Hsu, M. Biyari, M. Sookhak, R. Ameri, I. Dehzangi, A.T. Chronopoulos, H.-W. Liang, Application of explainable artificial intelligence in medical health: A systematic review of interpretability methods, Inform. Med. Unlocked 40 (2023) 101286, http://dx.doi.org/10.1016/j.imu.2023.101286, URL https://www.sciencedirect.com/science/article/pii/S2352914823001302.

[92] P. Tschandl, N. Codella, B.N. Akay, G. Argenziano, R.P. Braun, H. Cabo, D. Gutman, A. Halpern, B. Helba, R. Hofmann-Wellenhof, et al., Comparison of the accuracy of human readers versus machine-learning algorithms for pigmented skin lesion classification: an open, web-based, international, diagnostic study, Lancet Oncol. 20 (7) (2019) 938–947.

[93] J. Amann, A. Blasimme, E. Vayena, D. Frey, V.I. Madai, Explainability for artificial intelligence in healthcare: a multidisciplinary perspective, BMC Med. Inform. Decis. Mak. 20 (1) (2020) 1–9.

[94] Coalition for Health AI (CHAI), Blueprint for trustworthy AI implementation guidance and assurance for healthcare, 2023, URL https://www.coalitionforhealthai.org/papers/Blueprint%20for%20Trustworthy%20AI.pdf.

[95] T. Han, S. Srinivas, H. Lakkaraju, Which explanation should I choose? A function approximation perspective to characterizing post hoc explanations, Adv. Neural Inf. Process. Syst. 35 (2020) URL https://par.nsf.gov/biblio/10396110.

[96] C. Agarwal, S. Krishna, E. Saxena, M. Pawelczyk, N. Johnson, I. Puri, M. Zitnik, H. Lakkaraju, OpenXAI: Towards a transparent evaluation of model explanations, in: S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, A. Oh (Eds.), in: Advances in Neural Information Processing Systems, vol. 35, Curran Associates, Inc., 2022, pp. 15784–15799.

[97] N. Bussmann, P. Giudici, D. Marinelli, J. Papenbrock, Explainable machine learning in credit risk management, Comput. Econ. 57 (2021) 203–216.

[98] S. Sachan, J.-B. Yang, D.-L. Xu, D.E. Benavides, Y. Li, An explainable AI decision-support-system to automate loan underwriting, Expert Syst. Appl. 144 (2020) 113100.

[99] C. Rudin, J. Radin, Why are we using black box models in AI when we don't need to? A lesson from an explainable AI competition, Harv. Data Sci. Rev. 1 (2) (2019) 10–1162.

[100] S. Mishra, S. Dutta, J. Long, D. Magazzeni, A survey on the robustness of feature importance and counterfactual explanations, 2021, arXiv preprint arXiv:2111.00358.

[101] S. Sharma, S. Dhal, T. Rout, B.S. Acharya, Drones and machine learning for estimating forest carbon storage, Carbon Res. 1 (1) (2022) 21.

[102] T.B. Möllmann, B. Möhring, A practical way to integrate risk in forest management decisions, Ann. For. Sci. 74 (2017) 1–12.

[103] C. Gollob, T. Ritter, A. Nothdurft, Forest inventory with long range and high-speed personal laser scanning (PLS) and simultaneous localization and mapping (SLAM) technology, Remote Sens. 12 (9) (2020) 1509.

[104] A. Holzinger, A. Saranti, A. Angerschmid, C.O. Retzlaff, A. Gronauer, V. Pejakovic, F. Medel-Jimenez, T. Krexner, C. Gollob, K. Stampfer, Digital transformation in smart farm and forest operations needs human-centered AI: challenges and future directions, Sensors 22 (8) (2022) 3043.

[105] A. Holzinger, K. Stampfer, A. Nothdurft, C. Gollob, P. Kieseberg, Challenges in Artificial Intelligence for Smart Forestry, Vol. 130, Eur. Res. Consort. Informatics Math.(ERCIM) News, 2022, pp. 40–41.

[106] A. Holzinger, The next frontier: AI we can really trust, in: Machine Learning and Principles and Practice of Knowledge Discovery in Databases: International Workshops of ECML PKDD 2021, Virtual Event, September 13-17, 2021, Proceedings, Part I, Springer, 2022, pp. 427–440.

[107] A. Holzinger, M. Dehmer, F. Emmert-Streib, R. Cucchiara, I. Augenstein, J. Del Ser, W. Samek, I. Jurisica, N. Díaz-Rodríguez, Information fusion as an integrative cross-cutting enabler to achieve robust, explainable, and trustworthy medical artificial intelligence, Inf. Fusion 79 (2022) 263–278.

[108] R. Luckin, W. Holmes, M. Griffiths, L.B. Forcier, Intelligence Unleashed: An Argument for AI in Education, Tech. rep., The Open University, 2016.

[109] O. Zawacki-Richter, V.I. Marín, M. Bond, F. Gouverneur, Systematic review of research on artificial intelligence applications in higher education–where are the educators? Int. J. Educ. Technol. High. Educ. 16 (1) (2019) 1–27.

[110] L. Longo, Empowering qualitative research methods in education with artificial intelligence, in: A.P. Costa, L.P. Reis, A. Moreira (Eds.), Computer Supported Qualitative Research, Springer International Publishing, Cham, 2020, pp. 1–21.

[111] M.C. Desmarais, R.S.d. Baker, A review of recent advances in learner and skill modeling in intelligent learning environments, User Model. User-Adapt. Interact. 22 (2012) 9–38.

[112] K. VanLehn, The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems, Educ. Psychol. 46 (4) (2011) 197–221.

[113] S. Bull, There are open learner models about!, IEEE Trans. Learn. Technol. 13 (2) (2020) 425–448.

[114] B.B. (du), Artificial intelligence as an effective classroom assistant, IEEE Intell. Syst. 31 (6) (2016) 76–81.

[115] K. Holstein, B.M. McLaren, V. Aleven, Co-designing a real-time classroom orchestration tool to support teacher-AI complementarity, Grantee Submiss. (2019).

[116] A. Singh, S. Karayev, K. Gutowski, P. Abbeel, Gradescope: a fast, flexible, and fair system for scalable assessment of handwritten work, in: Proceedings of the Fourth (2017) Acm Conference on Learning@ Scale, 2017, pp. 81–88.

[117] G. Hiremath, A. Hajare, P. Bhosale, R. Nanaware, K. Wagh, Chatbot for education system, Int. J. Adv. Res. Ideas Innov. Technol. 4 (3) (2018) 37–43.

[118] M. Liz-Domínguez, M. Caeiro-Rodríguez, M. Llamas-Nistal, F.A. Mikic-Fonte, Systematic literature review of predictive analysis tools in higher education, Appl. Sci. 9 (24) (2019) 5569.

[119] H. Khosravi, K. Kitto, W. Joseph, RiPPLE: A crowdsourced adaptive platform for recommendation of learning activities, J. Learn. Anal. 6 (3) (2019) 91–105.

[120] W. Holmes, K. Porayska-Pomsta, K. Holstein, E. Sutherland, T. Baker, S.B. Shum, O.C. Santos, M.T. Rodrigo, M. Cukurova, I.I. Bittencourt, et al., Ethics of AI in education: Towards a community-wide framework, Int. J. Artif. Intell. Educ. (2021) 1–23.

[121] R.S. Baker, A. Hawn, Algorithmic bias in education, Int. J. Artif. Intell. Educ. (2021) 1–41.

[122] R.F. Kizilcec, H. Lee, Algorithmic fairness in education, in: The Ethics of Artificial Intelligence in Education, Routledge, 2022, pp. 174–202.

[123] S. Abdi, H. Khosravi, S. Sadiq, D. Gasevic, Complementing educational recommender systems with open learner models, in: Proceedings of the Tenth International Conference on Learning Analytics & Knowledge, 2020, pp. 360–365.

[124] F.-A. Croitoru, V. Hondru, R.T. Ionescu, M. Shah, Diffusion models in vision: A survey, IEEE Trans. Pattern Anal. Mach. Intell. 45 (9) (2023) 10850–10869, http://dx.doi.org/10.1109/TPAMI.2023.3261988.

[125] L. Yang, Z. Zhang, Y. Song, S. Hong, R. Xu, Y. Zhao, W. Zhang, B. Cui, M.-H. Yang, Diffusion models: A comprehensive survey of methods and applications, ACM Comput. Surv. (2023) http://dx.doi.org/10.1145/3626235, Just Accepted.

[126] M.O. Topal, A. Bas, I. van Heerden, Exploring transformers in natural language generation: GPT, BERT, and XLNet, 2021, ArXiv abs/2102.08036. URL https://api.semanticscholar.org/CorpusID:231933669.

[127] T. Bricken, A. Templeton, J. Batson, B. Chen, A. Jermyn, Towards Monosemanticity: Decomposing Language Models With Dictionary Learning, Tech. rep., Anthropic, 2023, URL https://transformer-circuits.pub/2023/monosemantic-features/index.html.

[128] N. Cammarata, S. Carter, G. Goh, C. Olah, M. Petrov, L. Schubert, C. Voss, B. Egan, S.K. Lim, Thread: Circuits, Distill (2020) http://dx.doi.org/10.23915/distill.00024.

[129] N. Elhage, N. Nanda, C. Olsson, T. Henighan, A Mathematical Framework for Transformer Circuits, Tech. rep., Anthropic, 2021, URL https://transformer-circuits.pub/2021/framework/index.html.

[130] C. Olah, N. Cammarata, L. Schubert, G. Goh, M. Petrov, S. Carter, Zoom in: An introduction to circuits, Distill (2020) http://dx.doi.org/10.23915/distill.00024.001, https://distill.pub/2020/circuits/zoom-in.

[131] N. Nanda, L. Chan, T. Lieberum, J. Smith, J. Steinhardt, Progress measures for grokking via mechanistic interpretability, 2023, URL https://arxiv.org/abs/2301.05217v2.

[132] S.D. Zhang, C. Tigges, S. Biderman, M. Raginsky, T. Ringer, Can transformers learn to solve problems recursively? 2023, http://dx.doi.org/10.48550/arXiv.2305.14699, arXiv:2305.14699 [cs]. URL http://arxiv.org/abs/2305.14699.

[133] S. Black, L. Sharkey, L. Grinsztajn, E. Winsor, D. Braun, J. Merizian, K. Parker, C.R. Guevara, B. Millidge, G. Alfour, C. Leahy, Interpreting neural networks through the polytope lens, 2022, URL https://arxiv.org/abs/2211.12312v1.

[134] Z. Zhong, Z. Liu, M. Tegmark, J. Andreas, The clock and the pizza: Two stories in mechanistic explanation of neural networks, 2023, URL https://arxiv.org/abs/2306.17844v1.

[135] R.S. Zimmermann, T. Klein, W. Brendel, Scale alone does not improve mechanistic interpretability in vision models, 2023, URL https://arxiv.org/abs/2307.05471v1.

[136] S.-i. Amari, Information Geometry and Its Applications, Springer, 2016, Google-Books-ID: UkSFCwAAQBAJ.

[137] M. Brcic, R.V. Yampolskiy, Impossibility results in AI: A survey, ACM Comput. Surv. 56 (1) (2023) 8:1–8:24, http://dx.doi.org/10.1145/3603371, URL https://dl.acm.org/doi/10.1145/3603371.

[138] Z. Liu, E. Gan, M. Tegmark, Seeing is believing: Brain-inspired modular training for mechanistic interpretability, 2023, http://dx.doi.org/10.48550/arXiv.2305.08746, arXiv:2305.08746 [cond-mat, q-bio]. URL http://arxiv.org/abs/2305.08746.

[139] N. Rodríguez-Barroso, D. Jiménez-López, M.V. Luzón, F. Herrera, E. Martínez-Cámara, Survey on federated learning threats: Concepts, taxonomy on attacks and defences, experimental study and challenges, Inf. Fusion 90 (2023) 148–173.

[140] J.L.C. Bárcena, P. Ducange, F. Marcelloni, G. Nardini, A. Noferi, A. Renda, F. Ruffini, A. Schiavo, G. Stea, A. Virdis, Enabling federated learning of explainable AI models within beyond-5G/6G networks, Comput. Commun. 210 (2023) 356–375.

[141] W. Du, M.J. Atallah, Secure multi-party computation problems and their applications: a review and open problems, in: Proceedings of the 2001 Workshop on New Security Paradigms, 2001, pp. 13–22.

[142] K. Simonyan, A. Vedaldi, A. Zisserman, Deep inside convolutional networks: Visualising image classification models and saliency maps, in: Y. Bengio, Y. LeCun (Eds.), 2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Workshop Track Proceedings, 2014, pp. 1–8, URL http://arxiv.org/abs/1312.6034.

[143] M.D. Zeiler, R. Fergus, Visualizing and understanding convolutional networks, in: European Conference on Computer Vision, Springer, 2014, pp. 818–833.

[144] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: Visual explanations from deep networks via gradient-based localization, in: IEEE International Conference on Computer Vision, ICCV, 2017, pp. 618–626.

[145] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, W. Samek, On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation, PLoS One 10 (7) (2015).

[146] B. Kim, M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, F. Viegas, et al., Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV), in: International Conference on Machine Learning, PMLR, 2018, pp. 2668–2677.

[147] C. Chen, O. Li, D. Tao, A. Barnett, C. Rudin, J.K. Su, This looks like that: deep learning for interpretable image recognition, Adv. Neural Inf. Process. Syst. 32 (2019).

[148] M. Nauta, R. Van Bree, C. Seifert, Neural prototype trees for interpretable fine-grained image recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 14933–14943.

[149] D. Rymarczyk, Ł. Struski, J. Tabor, B. Zieliński, Protopshare: Prototypical parts sharing for similarity discovery in interpretable image classification, in: Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, 2021, pp. 1420–1430.

[150] P.W. Koh, T. Nguyen, Y.S. Tang, S. Mussmann, E. Pierson, B. Kim, P. Liang, Concept bottleneck models, in: International Conference on Machine Learning, PMLR, 2020, pp. 5338–5348.

[151] M.E. Zarlenga, P. Barbiero, G. Ciravegna, G. Marra, F. Giannini, M. Diligenti, Z. Shams, F. Precioso, S. Melacci, A. Weller, et al., Concept embedding models, 2022, arXiv preprint arXiv:2209.09056.

[152] R. Achtibat, M. Dreyer, I. Eisenbraun, S. Bosse, T. Wiegand, W. Samek, S. Lapuschkin, From attribution maps to human-understandable explanations through Concept Relevance Propagation, Nat. Mach. Intell. 5 (9) (2023) 1006–1019, http://dx.doi.org/10.1038/s42256-023-00711-8, Number: 9, Publisher: Nature Publishing Group. URL https://www.nature.com/articles/s42256-023-00711-8.

[153] J. Mao, C. Gan, P. Kohli, J.B. Tenenbaum, J. Wu, The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision, in: 7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019, OpenReview.net, 2019, URL https://openreview.net/forum?id=rJgMlhRctm.

[154] M.K. Sarker, L. Zhou, A. Eberhart, P. Hitzler, Neuro-symbolic artificial intelligence, AI Commun. 34 (3) (2021) 197–209.

[155] K. Hamilton, A. Nayak, B. Božić, L. Longo, Is neuro-symbolic AI meeting its promises in natural language processing? A structured review, Semant. Web 15 (Preprint) (2022) 1–42.

[156] I. Tiddi, S. Schlobach, Knowledge graphs as tools for explainable machine learning: A survey, Artificial Intelligence 302 (2022) 103627.

[157] T. Räuker, A. Ho, S. Casper, D. Hadfield-Menell, Toward transparent ai: A survey on interpreting the inner structures of deep neural networks, in: 2023 IEEE Conference on Secure and Trustworthy Machine Learning, SaTML, IEEE, 2023, pp. 464–483.

[158] J. Johnson, B. Hariharan, L. Van Der Maaten, L. Fei-Fei, C. Lawrence Zitnick, R. Girshick, Clevr: A diagnostic dataset for compositional language and elementary visual reasoning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2901–2910.

[159] K. Yi, C. Gan, Y. Li, P. Kohli, J. Wu, A. Torralba, J.B. Tenenbaum, CLEVRER: collision events for video representation and reasoning, in: ICLR, 2020.

[160] H. Müller, A. Holzinger, Kandinsky patterns, Artificial Intelligence 300 (2021) 103546.

[161] H. de Vries, D. Bahdanau, S. Murty, A.C. Courville, P. Beaudoin, CLOSURE: assessing systematic generalization of CLEVR models, in: Visually Grounded Interaction and Language (ViGIL), NeurIPS 2019 Workshop, Vancouver, Canada, December 13, 2019, 2019, URL https://vigilworkshop.github.io/static/papers/28.pdf.

[162] J. Schneider, G. Apruzzese, Concept-based adversarial attacks: Tricking humans and classifiers alike, in: 2022 IEEE Security and Privacy Workshops, SPW, IEEE, 2022, pp. 66–72.

[163] A. Nguyen, A. Dosovitskiy, J. Yosinski, T. Brox, J. Clune, Synthesizing the preferred inputs for neurons in neural networks via deep generator networks, in: Advances in Neural Information Processing Systems, 2016, pp. 3387–3395.

[164] J. Schneider, M. Vlachos, A survey of deep learning: From activations to transformers, 2023, arXiv preprint arXiv:2302.00722.

[165] J. Schneider, M. Vlachos, Explaining classifiers by constructing familiar concepts, Mach. Learn. (2022) 1–34.

[166] C.-K. Yeh, C.-Y. Hsieh, A. Suggala, D.I. Inouye, P.K. Ravikumar, On the (in) fidelity and sensitivity of explanations, Adv. Neural Inf. Process. Syst. 32 (2019).

[167] Y. Gao, S. Gu, J. Jiang, S.R. Hong, D. Yu, L. Zhao, Going beyond XAI: A systematic survey for explanation-guided learning, 2022, arXiv preprint arXiv:2212.03954.

[168] A. Ferrario, M. Loi, The robustness of counterfactual explanations over time, IEEE Access 10 (2022) 82736–82750.

[169] L. Qiu, Y. Yang, C.C. Cao, Y. Zheng, H. Ngai, J. Hsiao, L. Chen, Generating perturbation-based explanations with robustness to out-of-distribution data, in: Proceedings of the ACM Web Conference 2022, 2022, pp. 3594–3605.

[170] D. Seuß, Bridging the gap between explainable AI and uncertainty quantification to enhance trustability, 2021, arXiv preprint arXiv:2105.11828.

[171] A. Kuppa, N.-A. Le-Khac, Black box attacks on explainable artificial intelligence (XAI) methods in cyber security, in: 2020 International Joint Conference on Neural Networks, IJCNN, IEEE, 2020, pp. 1–8.

[172] A.C. Oksuz, A. Halimi, E. Ayday, AUTOLYCUS: Exploiting explainable AI (XAI) for model extraction attacks against decision tree models, 2023, arXiv preprint arXiv:2302.02162.

[173] F. Pahde, M. Dreyer, W. Samek, S. Lapuschkin, Reveal to revise: An explainable AI life cycle for iterative bias correction of deep models, in: H. Greenspan, A. Madabhushi, P. Mousavi, S. Salcudean, J. Duncan, T. Syeda-Mahmood, R. Taylor (Eds.), Medical Image Computing and Computer Assisted Intervention – MICCAI 2023, Springer Nature Switzerland, Cham, 2023, pp. 596–606.

[174] M. Krishnan, Against interpretability: A critical examination of the interpretability problem in machine learning, Philos. Technol. 33 (3) (2020) 487–502, http://dx.doi.org/10.1007/s13347-019-00372-9.

[175] 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.

[176] U. Ehsan, M.O. Riedl, Social construction of XAI: Do we need one definition to rule them all? in: M. Muller, P. Angelov, H. Daume III, S. Guha, Q.V. Liao, N. Oliver, D. Piorkowski (Eds.), Proceedings of the NeurIPS 2022 Workshop on Human-Centered AI, 2022, arXiv:2211.06499.

[177] M.-A. Clinciu, H. Hastie, A survey of explainable AI terminology, in: J.M. Alonso, A. Catala (Eds.), Proceedings of the 1st Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence, in: NL4XAI 2019, Association for Computational Linguistics, Stroudsburg, PA, USA, 2019, pp. 8–13, http://dx.doi.org/10.18653/v1/W19-8403.

[178] M. Graziani, L. Dutkiewicz, D. Calvaresi, J.P. Amorim, K. Yordanova, M. Vered, R. Nair, P.H. Abreu, T. Blanke, V. Pulignano, J.O. Prior, L. Lauwaert, W. Reijers, A. Depeursinge, V. Andrearczyk, H. Müller, A global taxonomy of interpretable AI: Unifying the terminology for the technical and social sciences, Artif. Intell. Rev. (2022) 1–32, http://dx.doi.org/10.1007/s10462-022-10256-8.

[179] N. Díaz-Rodríguez, J. Del Ser, M. Coeckelbergh, M. López de Prado, E. Herrera-Viedma, F. Herrera, Connecting the dots in trustworthy Artificial Intelligence: From AI principles, ethics, and key requirements to responsible AI systems and regulation, Inf. Fusion 99 (2023) 101896, http://dx.doi.org/10.1016/j.inffus.2023.101896.

[180] S. Robbins, A misdirected principle with a catch: Explicability for AI, Minds Mach. 29 (4) (2019) 495–514, http://dx.doi.org/10.1007/s11023-019-09509-3.

[181] R.F. Kizilcec, How much information? Effects of transparency on trust in an algorithmic interface, in: J. Kaye, A. Druin, C. Lampe, D. Morris, J.P. Hourcade (Eds.), Proceedings of the 34th Conference on Human Factors in Computing Systems, in: CHI 2016, Association for Computing Machinery, New York, NY, USA, 2016, pp. 2390–2395, http://dx.doi.org/10.1145/2858036.2858402.

[182] B. Ghosh, D. Malioutov, K.S. Meel, Interpretable classification rules in relaxed logical form, in: T. Miller, R. Weber, D. Magazzeni (Eds.), Proceedings of the IJCAI 2019 Workshop on Explainable Artificial Intelligence, in: IJCAI XAI 2019, 2019, pp. 14–20.

[183] J. Newman, A Taxonomy of Trustworthiness for Artificial Intelligence, CLTC White Paper Series, North Charleston, SC, USA, 2023, URL https://cltc.berkeley.edu/publication/a-taxonomy-of-trustworthiness-for-artificial-intelligence/.

[184] N. Palladino, A 'biased' emerging governance regime for artificial intelligence? How AI ethics get skewed moving from principles to practices, Telecommun. Policy 47 (5) (2023) 102479, http://dx.doi.org/10.1016/j.telpol.2022.102479.

[185] K. Khalifa, Inaugurating understanding or repackaging explanation? Philos. Sci. 79 (1) (2012) 15–37.

[186] M. Strevens, No understanding without explanation, Stud. Hist. Philos. Sci. A 44 (3) (2013) 510–515.

[187] P. Lipton, Understanding without explanation, in: Scientific Understanding: Philosophical Perspectives, 2009, pp. 43–63.

[188] C.Z. Elgin, True Enough, MIT Press, 2017.

[189] J. Kvanvig, Responses to critics, in: Epistemic Value, Oxford University Press, Oxford, 2009, pp. 339–351.

[190] M. Mizrahi, Idealizations and scientific understanding, Philos. Stud. 160 (2012) 237–252.

[191] J.A. Carter, E.C. Gordon, Objectual understanding, factivity and belief, in: Epistemic Reasons, Norms and Goals, Vol. 423, De Gruyter, Berlin, 2016.

[192] A. Erasmus, T.D. Brunet, E. Fisher, What is interpretability? Philos. Technol. 34 (4) (2021) 833–862.

[193] D. Pritchard, Knowing the Answer, Understanding and Epistemic Value, Citeseer, 2008.

[194] L. Zagzebski, On Epistemology, Wadsworth, 2009.

[195] T. Lombrozo, D. Wilkenfeld, T. Lombrozo, D. Wilkenfeld, Mechanistic versus functional understanding, in: Varieties of Understanding: New Perspectives from Philosophy, Psychology, and Theology, Oxford University Press, New York, NY, 2019, pp. 209–229.

[196] E. Sullivan, Understanding from machine learning models, British J. Philos. Sci. (2022).

[197] K.A. Creel, Transparency in complex computational systems, Philos. Sci. 87 (4) (2020) 568–589.

[198] J.M. Durán, Dissecting scientific explanation in AI (sXAI): A case for medicine and healthcare, Artificial Intelligence 297 (2021) 103498.

[199] C. Zednik, Solving the black box problem: A normative framework for explainable artificial intelligence, Philos. Technol. 34 (2) (2021) 265–288.

[200] W. Fleisher, Understanding, idealization, and explainable AI, Episteme 19 (4) (2022) 534–560.

[201] P. Pirozelli, Sources of understanding in supervised machine learning models, Philos. Technol. 35 (2) (2022) 23.

[202] M.M. De Graaf, B.F. Malle, How people explain action (and autonomous intelligent systems should too), in: 2017 AAAI Fall Symposium Series, 2017, pp. 19–26.

[203] B. Mittelstadt, C. Russell, S. Wachter, Explaining explanations in AI, in: Proceedings of the Conference on Fairness, Accountability, and Transparency, 2019, pp. 279–288.

[204] R. Guidotti, Evaluating local explanation methods on ground truth, Artificial Intelligence 291 (2021) 103428.

[205] I. Sevillano-García, J. Luengo, F. Herrera, REVEL framework to measure local linear explanations for black-box models: Deep learning image classification case study, Int. J. Intell. Syst. 2023 (2023) 1–34.

[206] M.T. Keane, E.M. Kenny, E. Delaney, B. Smyth, If only we had better counterfactual explanations: Five key deficits to rectify in the evaluation of counterfactual xai techniques, in: Proceedings of the 30th International Joint Conference on Artificial Intelligence, IJCAI-21, 2021, pp. 4466–4474.

[207] J. Dodge, Q.V. Liao, Y. Zhang, R.K.E. Bellamy, C. Dugan, Explaining models: an empirical study of how explanations impact fairness judgment, in: IUI, ACM, 2019, pp. 275–285.

[208] A. Lucic, H. Haned, M. de Rijke, Why does my model fail?: contrastive local explanations for retail forecasting, in: FAT*, ACM, 2020, pp. 90–98.

[209] C. Metta, R. Guidotti, Y. Yin, P. Gallinari, S. Rinzivillo, Exemplars and counterexemplars explanations for skin lesion classifiers, in: HHAI, in: Frontiers in Artificial Intelligence and Applications, vol. 354, IOS Press, 2022, pp. 258–260.

[210] R.R. Hoffman, S.T. Mueller, G. Klein, J. Litman, Metrics for explainable AI: Challenges and prospects, 2018, ArXiv abs/1812.04608.

[211] C. van der Lee, A. Gatt, E. van Miltenburg, E. Krahmer, Human evaluation of automatically generated text: Current trends and best practice guidelines, Comput. Speech Lang. 67 (2021) 101151, http://dx.doi.org/10.1016/j.csl.2020.101151, URL https://www.sciencedirect.com/science/article/pii/S088523082030084X.

[212] M. Nauta, J. Trienes, S. Pathak, E. Nguyen, M. Peters, Y. Schmitt, J. Schlötterer, M. van Keulen, C. Seifert, From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable AI, ACM Comput. Surv. (2023) http://dx.doi.org/10.1145/3583558.

[213] R. Confalonieri, J.M. Alonso-Moral, An operational framework for guiding human evaluation in Explainable and Trustworthy AI, IEEE Intell. Syst. (2023) 1–13, http://dx.doi.org/10.1109/MIS.2023.3334639.

[214] A. Hedström, L. Weber, D. Bareeva, F. Motzkus, W. Samek, S. Lapuschkin, M.M.C. Höhne, Quantus: An explainable AI toolkit for responsible evaluation of neural network explanation, J. Mach. Learn. Res. 24 (34) (2023) 1–11, URL http://jmlr.org/papers/v24/22-0142.html.

[215] L. Arras, A. Osman, W. Samek, CLEVR-XAI: A benchmark dataset for the ground truth evaluation of neural network explanations, Inf. Fusion 81 (2022) 14–40, http://dx.doi.org/10.1016/j.inffus.2021.11.008.

[216] F. Pahde, M. Dreyer, W. Samek, S. Lapuschkin, Reveal to revise: An explainable AI life cycle for iterative bias correction of deep models, in: H. Greenspan, A. Madabhushi, P. Mousavi, S. Salcudean, J. Duncan, T. Syeda-Mahmood, R. Taylor (Eds.), Medical Image Computing and Computer Assisted Intervention – MICCAI 2023, Springer Nature Switzerland, Cham, 2023, pp. 596–606.

[217] L. Longo, Formalising human mental workload as non-monotonic concept for adaptive and personalised web-design, in: User Modeling, Adaptation, and Personalization: 20th International Conference, UMAP 2012, Montreal, Canada, July 16-20, 2012. Proceedings 20, Springer, 2012, pp. 369–373.

[218] L. Longo, Designing medical interactive systems via assessment of human mental workload, in: 2015 IEEE 28th International Symposium on Computer-Based Medical Systems, IEEE, 2015, pp. 364–365.

[219] G. Hancock, L. Longo, M. Young, P. Hancock, Mental workload, in: Handbook of Human Factors and Ergonomics, Wiley Online Library, 2021, pp. 203–226.

[220] L. Longo, C.D. Wickens, P.A. Hancock, G.M. Hancock, Human mental workload: A survey and a novel inclusive definition, Front. Psychol. 13 (2022) http://dx.doi.org/10.3389/fpsyg.2022.883321, URL https://www.frontiersin.org/article/10.3389/fpsyg.2022.883321.

[221] R. Confalonieri, T. Weyde, T.R. Besold, F.M. del Prado Martín, Using ontologies to enhance human understandability of global post-hoc explanations of Black-box models, Artificial Intelligence 296 (2021) http://dx.doi.org/10.1016/j.artint.2021.103471.

[222] I.E. Nielsen, D. Dera, G. Rasool, R.P. Ramachandran, N.C. Bouaynaya, Robust explainability: A tutorial on gradient-based attribution methods for deep neural networks, IEEE Signal Process. Mag. 39 (4) (2022) 73–84.

[223] J. Yuan, T. Chen, B. Li, X. Xue, Compositional scene representation learning via reconstruction: A survey, IEEE Trans. Pattern Anal. Mach. Intell. 45 (10) (2023) 11540–11560, http://dx.doi.org/10.1109/TPAMI.2023.3286184.

[224] T. Klinger, D. Adjodah, V. Marois, J. Joseph, M. Riemer, A. Pentland, M. Campbell, A study of compositional generalization in neural models, 2020, arXiv preprint arXiv:2006.09437.

[225] L. Rizzo, L. Longo, An empirical evaluation of the inferential capacity of defeasible argumentation, non-monotonic fuzzy reasoning and expert systems, Expert Syst. Appl. 147 (2020) 113220.

[226] L. Rizzo, L. Longo, A qualitative investigation of the explainability of defeasible argumentation and non-monotonic fuzzy reasoning, in: Proceedings for the 26th AIAI Irish Conference on Artificial Intelligence and Cognitive Science Trinity College Dublin, Dublin, Ireland, December 6-7th, 2018, 2018, pp. 138–149.

[227] G. Vilone, L. Longo, A novel human-centred evaluation approach and an argument-based method for explainable artificial intelligence, in: Artificial Intelligence Applications and Innovations: 18th IFIP WG 12.5 International Conference, AIAI 2022, Hersonissos, Crete, Greece, June 17–20, 2022, Proceedings, Part I, Springer, 2022, pp. 447–460.

[228] G. Vilone, L. Longo, An XAI method for the automatic formation of an abstract argumentation framework from a neural network and its objective evaluation, in: 1st International Workshop on Argumentation for EXplainable AI Co-Located with 9th International Conference on Computational Models of Argument (COMMA 2022), in: CEUR Workshop Proceedings, vol. 3209, CEUR-WS.org, 2022, URL http://ceur-ws.org/Vol-3209/2119.pdf.

[229] J. Vielhaben, S. Lapuschkin, G. Montavon, W. Samek, Explainable ai for time series via virtual inspection layers, Pattern Recognition 150 (2024) 110309, http://dx.doi.org/10.1016/j.patcog.2024.110309, https://www.sciencedirect.com/science/article/pii/S0031320324000608.

[230] T. Ahmed, L. Longo, Interpreting disentangled representations of person-specific convolutional variational autoencoders of spatially preserving EEG topographic maps via clustering and visual plausibility, Information 14 (9) (2023) http://dx.doi.org/10.3390/info14090489, URL https://www.mdpi.com/2078-2489/14/9/489.

[231] W.V. Quine, On what there is, in: W.V. Quine (Ed.), From a Logical Point of View, Harvard University Press, Cambridge, Mass., 1953, pp. 1–19.

[232] D. Krakauer, The computational systems of the world, BioScience 64 (4) (2014) 351–354, http://dx.doi.org/10.1093/biosci/biu024.

[233] S. Badreddine, A.d. Garcez, L. Serafini, M. Spranger, Logic tensor networks, Artificial Intelligence 303 (2022) 103649.

[234] C. Meske, E. Bunde, J. Schneider, M. Gersch, Explainable artificial intelligence: Objectives, stakeholders and future research opportunities, Inf. Syst. Manage. (2020).

[235] A. Weller, Transparency: Motivations and challenges, in: W. Samek, G. Montavon, A. Vedaldi, L.K. Hansen, K.-R. Müller (Eds.), Explainable AI: Interpreting, Explaining and Visualizing Deep Learning, in: Lecture Notes in Computer Science, Springer International Publishing, Cham, 2019, pp. 23–40.

[236] R. Hamon, H. Junklewitz, G. Malgieri, P.D. Hert, L. Beslay, I. Sanchez, Impossible explanations? Beyond explainable AI in the GDPR from a COVID-19 use case scenario, in: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 2021, pp. 549–559.

[237] J.C. Flack, Multiple time-scales and the developmental dynamics of social systems, Philos. Trans. R. Soc. B 367 (1597) (2012) 1802–1810, http://dx.doi.org/10.1098/rstb.2011.0214, Publisher: Royal Society. URL https://royalsocietypublishing.org/doi/10.1098/rstb.2011.0214.

[238] M. Juric, A. Sandic, M. Brcic, AI safety: state of the field through quantitative lens, in: 2020 43rd International Convention on Information, Communication and Electronic Technology (MIPRO), 2020, pp. 1254–1259, http://dx.doi.org/10.23919/MIPRO48935.2020.9245153, ISSN: 2623-8764.

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

[240] S. Beckers, Causal explanations and XAI, in: Conference on Causal Learning and Reasoning, PMLR, 2022, pp. 90–109.

[241] Y.-L. Chou, C. Moreira, P. Bruza, C. Ouyang, J. Jorge, Counterfactuals and causability in explainable artificial intelligence: Theory, algorithms, and applications, Inf. Fusion 81 (2022) 59–83.

[242] M. Cinquini, R. Guidotti, CALIME: Causality-aware local interpretable model-agnostic explanations, 2022, arXiv preprint arXiv:2212.05256.

[243] J. Del Ser, A. Barredo-Arrieta, N. Díaz-Rodríguez, F. Herrera, A. Saranti, A. Holzinger, On generating trustworthy counterfactual explanations, Inform. Sci. 655 (2024) 119898.

[244] P. Sanchez, S.A. Tsaftaris, Diffusion causal models for counterfactual estimation, in: Conference on Causal Learning and Reasoning, CLeaR, 2022.

[245] M. Augustin, V. Boreiko, F. Croce, M. Hein, Diffusion visual counterfactual explanations, in: NeurIPS, 2022.

[246] J. Schneider, M. Vlachos, Personalization of deep learning, in: Data Science–Analytics and Applications: Proceedings of the 3rd International Data Science Conference–iDSC2020, Springer, 2021, pp. 89–96.

[247] J. Schneider, J.P. Handali, Personalized explanation for machine learning: a conceptualization, in: European Conference on Information Systems, ECIS, 2019.

[248] B. Zhu, M. Jordan, J. Jiao, Principled reinforcement learning with human feedback from pairwise or K-wise comparisons, in: A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, J. Scarlett (Eds.), Proceedings of the 40th International Conference on Machine Learning, in: Proceedings of Machine Learning Research, vol. 202, PMLR, 2023, pp. 43037–43067.

[249] T. Bewley, F. Lecue, Interpretable preference-based reinforcement learning with tree-structured reward functions, in: Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems, AAMAS '22, International Foundation for Autonomous Agents and Multiagent Systems, 2022, pp. 118–126.

[250] A. Bunt, M. Lount, C. Lauzon, Are explanations always important?: a study of deployed, low-cost intelligent interactive systems, in: Proceedings of the 2012 ACM International Conference on Intelligent User Interfaces, IUI '12, ACM, New York, NY, USA, 2012, pp. 169–178, http://dx.doi.org/10.1145/2166966.2166996, URL http://doi.acm.org/10.1145/2166966.2166996.

[251] 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, http://dx.doi.org/10.1038/s42256-019-0048-x, Number: 5 Publisher: Nature Publishing Group. URL https://www.nature.com/articles/s42256-019-0048-x.

[252] R. Hamon, H. Junklewitz, I. Sanchez, G. Malgieri, P. De Hert, Bridging the gap between AI and explainability in the GDPR: Towards trustworthiness-by-design in automated decision-makin, IEEE Comput. Intell. Mag. 17 (1) (2022) 72–85.

[253] D. Gunning, M. Stefik, J. Choi, T. Miller, S. Stumpf, G.-Z. Yang, XAI—Explainable artificial intelligence, Science Robotics 4 (37) (2019) http://dx.doi.org/10.1126/scirobotics.aay7120, URL https://robotics.sciencemag.org/content/4/37/eaay7120.

[254] A. Krajna, M. Brcic, T. Lipic, J. Doncevic, Explainability in reinforcement learning: perspective and position, 2022, http://dx.doi.org/10.48550/arXiv.2203.11547.

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

[256] F. Cabitza, A. Campagner, L. Famiglini, E. Gallazzi, G.A. La Maida, Color shadows (part I): Exploratory usability evaluation of activation maps in radiological machine learning, in: Machine Learning and Knowledge Extraction: 6th IFIP TC 5, TC 12, WG 8.4, WG 8.9, WG 12.9 International Cross-Domain Conference, CD-MAKE 2022, Vienna, Austria, August 23–26, 2022, Proceedings, Springer, 2022, pp. 31–50.

[257] F. Cabitza, A. Campagner, L. Ronzio, M. Cameli, G.E. Mandoli, M.C. Pastore, L. Sconfienza, D. Folgado, M. Barandas, H. Gamboa, Rams, hounds and white boxes: Investigating human-AI collaboration protocols in medical diagnosis, Artif. Intell. Med. (2023) 102506.

[258] G. Bansal, T. Wu, J. Zhou, R. Fok, B. Nushi, E. Kamar, M.T. Ribeiro, D. Weld, Does the whole exceed its parts? the effect of ai explanations on complementary team performance, in: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, 2021, pp. 1–16.

[259] S. Natale, Deceitful Media: Artificial Intelligence and Social Life After the Turing Test, Oxford University Press, USA, 2021.

[260] F. Cabitza, A. Campagner, E. Datteri, To err is (only) human. Reflections on how to move from accuracy to trust for medical AI, in: Exploring Innovation in a Digital World: Cultural and Organizational Challenges, Springer, 2021, pp. 36–49.

[261] F. Cabitza, A. Campagner, C. Simone, The need to move away from agential-AI: Empirical investigations, useful concepts and open issues, Int. J. Hum.-Comput. Stud. 155 (2021) 102696.

[262] T. Miller, Explainable AI is dead, long live explainable AI! hypothesis-driven decision support using evaluative AI, in: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, 2023, pp. 333–342.

[263] 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, CHI '18, ACM, New York, NY, USA, 2018, pp. 582:1–582:18, http://dx.doi.org/10.1145/3173574.3174156, URL http://doi.acm.org/10.1145/3173574.3174156.

[264] K. Baum, S. Mantel, E. Schmidt, T. Speith, From responsibility to reason-giving explainable artificial intelligence, Philos. Technol. 35 (1) (2022) 1–30, http://dx.doi.org/10.1007/s13347-022-00510-w.

[265] S. Thornton, Karl Popper, in: E.N. Zalta, U. Nodelman (Eds.), The Stanford Encyclopedia of Philosophy, Winter 2023 ed., Metaphysics Research Lab, Stanford University, 2023.

[266] M.L. Leavitt, A. Morcos, Towards falsifiable interpretability research, in: NeurIPS 2020 Workshop: ML Retrospectives, Surveys and Meta-Analyses, ML-RSA, 2020.

[267] F.K. Dosilovic, M. Brcic, N. Hlupic, 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.

[268] A. Krajna, M. Brcic, M. Kovac, A. Sarcevic, Explainable artificial intelligence: An updated perspective, in: Proceedings of 45th Jubilee International Convention on Information, Communication and Electronic Technology (MIPRO) 2022, Opatija, Croatia, 2022, pp. 859–864.

[269] 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, Publisher: Proceedings of the National Academy of Sciences. URL https://www.pnas.org/doi/10.1073/pnas.1900654116.

[270] J. Schneider, C. Meske, M. Vlachos, Deceptive AI explanations: Creation and detection, in: International Conference on Agents and Artificial Intelligence, ICAART, 2022, pp. 44–55.

[271] J. Schneider, F. Breitinger, AI Forensics: Did the artificial intelligence system do it? why? 2020, arXiv preprint arXiv:2005.13635.

[272] J. Schneider, M. Vlachos, Reflective-net: Learning from explanations, Data Min. Knowl. Discov. (2023).

[273] R.V. Yampolskiy, Unexplainability and incomprehensibility of AI, J. Artif. Intell. Conscious. 07 (02) (2020) 277–291, http://dx.doi.org/10.1142/S2705078520500150, Publisher: World Scientific Publishing Co. URL https://www.worldscientific.com/doi/10.1142/S2705078520500150.

[274] R.V. Yampolskiy, What are the ultimate limits to computational techniques: verifier theory and unverifiability, Phys. Scr. 92 (9) (2017) 093001, http://dx.doi.org/10.1088/1402-4896/aa7ca8, Publisher: IOP Publishing.

[275] V. Boutin, T. Fel, L. Singhal, R. Mukherji, A. Nagaraj, J. Colin, T. Serre, Diffusion models as artists: Are we closing the gap between humans and machines? in: International Conference on Machine Learning, 2023, URL https://api.semanticscholar.org/CorpusID:256358696.

[276] H.H. Thorp, ChatGPT is fun, but not an author, Science 379 (6630) (2023) 313.

[277] E.A. van Dis, J. Bollen, W. Zuidema, R. van Rooij, C.L. Bockting, ChatGPT: five priorities for research, Nature 614 (7947) (2023) 224–226.

[278] F. Boenisch, A systematic review on model watermarking for neural networks, Front. Big Data 4 (2021) 729663.

[279] J. Kirchenbauer, J. Geiping, Y. Wen, J. Katz, I. Miers, T. Goldstein, A watermark for large language models, in: Proceedings of the 40th International Conference on Machine Learning, 2023, pp. 202:17061–17084.

[280] L. Bourtoule, V. Chandrasekaran, C.A. Choquette-Choo, H. Jia, A. Travers, B. Zhang, D. Lie, N. Papernot, Machine unlearning, in: 2021 IEEE Symposium on Security and Privacy, SP, IEEE, 2021, pp. 141–159.

[281] T.T. Nguyen, T.T. Huynh, P.L. Nguyen, A.W.-C. Liew, H. Yin, Q.V.H. Nguyen, A survey of machine unlearning, 2022, arXiv preprint arXiv:2209.02299.

[282] J.E. Cohen, Between Truth and Power, Oxford University Press, 2019.

[283] F. Cabitza, A. Campagner, G. Malgieri, C. Natali, D. Schneeberger, K. Stoeger, A. Holzinger, Quod erat demonstrandum?-Towards a typology of the concept of explanation for the design of explainable AI, Expert Syst. Appl. 213 (2023) 118888.

[284] G. Malgieri, "Just" algorithms: justification (beyond explanation) of automated decisions under the general data protection regulation, Law Bus. 1 (1) (2021) 16–28.

[285] E. Bayamlioglu, Contesting automated decisions, Eur. Data Prot. L. Rev. 4 (2018) 433.

[286] C. Henin, D. Le Métayer, Beyond explainability: justifiability and contestability of algorithmic decision systems, AI Soc. (2021) 1–14.

[287] C. Henin, D. Le Métayer, A framework to contest and justify algorithmic decisions, AI Ethics 1 (4) (2021) 463–476.

[288] L.M. Austin, Enough about me: why privacy is about power, not consent (or harm), in: A. Sarat (Ed.), A World Without Privacy: What Law Can and Should Do?, 2014, pp. 131–189.

[289] L. Wilsdon, Carissa véliz, privacy is power: Why and how you should take back control of your data, 2022.

[290] S. Costanza-Chock, Design Justice: Community-Led Practices to Build the Worlds We Need, The MIT Press, 2020.

[291] M.E. Kaminski, G. Malgieri, Algorithmic impact assessments under the GDPR: producing multi-layered explanations, Int. Data Priv. Law (2020) 19–28.

[292] J. Gregory, Scandinavian approaches to participatory design, Int. J. Eng. Educ. 19 (1) (2003) 62–74.

[293] A. Mantelero, Beyond Data: Human Rights, Ethical and Social Impact Assessment in AI, Springer Nature, 2022.

[294] G. Malgieri, In/acceptable marketing and consumers' privacy expectations: Four tests from EU data protection law, J. Consum. Mark. 40 (2) (2023) 209–223.

[295] K. Bodker, F. Kensing, J. Simonsen, Participatory IT Design: Designing for Business and Workplace Realities, MIT Press, 2009.