Explainable Artificial Intelligence (XAI) 2.0: A Manifesto of Open Challenges and Interdisciplinary Research Directions

23 authors, including:

Luca Longo
TU Dublin
147 PUBLICATIONS   **2,538** CITATIONS

SEE PROFILE

Mario Brcic
University of Zagreb
50 PUBLICATIONS   **1,299** CITATIONS

SEE PROFILE

Federico Cabitza
Università degli Studi di Milano-Bicocca
321 PUBLICATIONS   **5,426** CITATIONS

SEE PROFILE

Roberto Confalonieri
University of Padova
63 PUBLICATIONS   **1,111** CITATIONS

SEE PROFILE

Explainable Artificial Intelligence (XAI) 2.0: A Manifesto of Open Challenges and Interdisciplinary Research Directions

Luca Longo
The Artificial Intelligence and Cognitive Load Research Lab, Technological University Dublin, Ireland
Corresponding author: `luca.longo@tudublin.ie`

Mario Brcic
University of Zagreb, Faculty of Electrical Engineering and Computing, Croatia

Federico Cabitza
University of Milano-Bicocca, Milan, Italy, IRCCS Ospedale Galeazzi Sant'Ambrogio, Milan, Italy

Jaesik Choi
Kim Jaechul Graduate School of AI, Korea Advanced Institute of Science & Technology, Republic of Korea
INEEJI Corporation, Republic of Korea

Roberto Confalonieri
Department of Mathematics, University of Padua, Italy

Javier Del Ser
TECNALIA, Basque Research & Technology Alliance (BRTA), Derio, Spain,
University of the Basque Country (UPV/EHU), Bilbao, Spain

Riccardo Guidotti
University of Pisa, Pisa, Italy

Yoichi Hayashi
Department of Computer Science, Meiji University, Tokyo, Japan

Francisco Herrera
Department of Computer Science and Artificial Intelligence,
DaSCI Andalusian Institute in Data Science & Computational Intelligence, University of Granada, Granada, Spain

Andreas Holzinger
Human-Centered AI Lab, University of Natural Resources and Life Sciences Vienna, Austria

Richard Jiang
School of Computing and Communications, Lancaster University, UK

Hassan Khosravi
The University of Queensland, Brisbane, Australia

Freddy Lecue
National Institute for Research in Digital Science and Technology (INRIA), Sophia Antipolis, France

Gianclaudio Malgieri
eLaw Center for Law and Digital Technologies, Leiden University, Netherlands

Andrés Páez
Department of Philosophy, Universidad de los Andes, Bogotá, Colombia
Center for Research & Formation in Artificial Intelligence, Universidad de los Andes, Bogotá, Colombia

Wojciech Samek
Technical University of Berlin, Berlin, Germany,
Fraunhofer Heinrich Hertz Institute, Berlin, Germany,
Berlin Institute for the Foundations of Learning and Data (BIFOLD), Berlin, Germany

Johannes Schneider
Department of Information Systems and Computer Science, University of Liechtenstein, Liechtenstein

Timo Speith
Department of Philosophy, University of Bayreuth, Bayreuth, Germany
Center for Perspicuous Computing, Saarland University, Saarbrücken, Germany

Simone Stumpf
School of Computing Science, University of Glasgow, United Kingdom

ABSTRACT

As systems based on opaque Artificial Intelligence (AI) continue to flourish in diverse real-world applications, understanding these black box models has become paramount. In response, Explainable AI (XAI) has emerged as a field of research with practical and ethical benefits across various domains. This paper not only highlights the advancements in XAI and its application in real-world scenarios but also 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. Our goal is to put forward a comprehensive proposal for advancing XAI. To achieve this goal, we present a manifesto of 27 open problems categorized into nine categories. These 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.

Keywords Explainable Artificial Intelligence; XAI; Interpretability; Manifesto; Open Challenges; Interdisciplinarity; Ethical AI; Large Language Models; Trustworthy AI; Responsible AI; Generative AI; Multi-Faceted Explanations; Concept-Based Explanations; Causality; Actionable XAI; Falsifiability.

1 Introduction

The field of Explainable AI (XAI) has undergone significant growth and development over the past few years. It has evolved from being a niche research topic within the larger field of Artificial Intelligence (AI) [1, 2, 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, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17]. 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 [17, 18, 19, 20]. The growth of XAI can be attributed to the increasing success of AI. In recent years, Machine Learning (ML), specifically Deep Learning (DL), 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. ML and 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 [21] and healthcare [22] 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 the behaviour and outputs of AI systems. In other words, AI's need for transparency and interpretability has made XAI an area of study with practical and ethical value in various fields [4, 18]. Despite much progress in XAI, many open problems require further exploration. For instance, XAI alone is not enough for trustworthiness, and there is a lack of consensus concerning the priorities and directions needed to advance this research field. Often these open problems are viewed through siloed perspectives [23, 4]. A broader, multidisciplinary approach that draws on the expertise of researchers across different fields could bring about advances towards XAI 2.0. Our work addresses this gap by bringing together a wide range of experts from diverse fields to collaborate on identifying and tackling open problems in XAI. The focus is on synchronizing the research agendas of scholars working in the field, identifying directions that could catalyze XAI in real-world applications.

The goal is to form a proposal open to discussion, a sincere attempt at fostering a debate around XAI and what research should be addressed 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 Explainable AI (XAI). To achieve the goal described above, this article has come about through a very specific synthesis. To get different perspectives on XAI various experts from different disciplines, including philosophy, psychology, HCI, and, of course, computer science were brought together. This significant effort has resulted in a total of 27 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. Through the illustration of some of the many possible use-cases of AI, 2 presents a variety of advances of XAI techniques and methods, along with their application in real-world settings. This is meant to highlight the benefits of XAI to people, businesses, institutions, and society. Subsequently, the article's core follows in 3 by describing 27 problems in XAI, the challenges in solving them, and our suggestions for possible solutions. Finally, 4 summarizes our manifesto concisely, offering a roadmap for future research.

2 Advances and Applications of xAI Research

In this section, we showcase that research in XAI is alive and useful. In particular, 2.1 focuses on synthesizing the main breakthrough in XAI, demonstrating its enormous potential. Similarly, 2.2 demonstrates the large and increasing number of applications of XAI methods, techniques, and tools and their utility in real-world scenarios.

2.1 XAI Trends, Advances, and Breakthroughs

The prime goal of explanations is to make a model understandable or comprehensible to its stakeholders [24, 17, 25, 26]. 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 [4, 17, 27, 9, 10]. Synthesizing explanations for AI systems has been shown to have the potential to solve several technical and societal problems. Explanations can facilitate the understanding of how learning from data has occurred, for instance, via feature attribution methods. Furthermore, explanations can reveal information about 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 generalisation when predicting unseen data [28]. 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 [29]. 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 [30]. Explanations can also drive pruning and model compression strategies, linking irrelevant concepts to specific neurons that can hence be removed from a neural network [31].

2.1.1 Attribution Methods

A lot of work exists on explaining the decisions of a classifier with *attribution methods* [32]. For instance, model agnostic attribution methods such as Local Interpretable Model-Agnostic Explanations (LIME) [33], Shapley Additive Explanation (SHAP) [34], and many others can contribute to the interpretation of DL models by computing the importance of input features [35, 36]. 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, explanations using attributions 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, which is often present in images of this category. Such misbehaving classifiers have been termed 'Clever Hans' predictors [37] or 'Short-Cuts' [38]. However, identifying such misbehaviour or understanding the meaning of an attribution-based explanation can be significantly more difficult in other domains [39]. For instance, an attribution map computed on a multivariate time series signal or a complex biological sequence can be significantly more difficult to understand for the human receiver; that means the 'interpretation gap' is much larger 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 important cue for the prediction or the fact that the person smiles [40].

2.1.2 Interpretable Models

Explainability in contexts like finance often has a special 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) [41]. The benefit of these techniques is, among others, that supposedly they lead to inherently interpretable 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 [42] (cf., [43]). Accordingly, models that are interpretable by design, such as DTs [44], are preferred for many applications [45]. 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 [46]. Although these methods can improve the overall performance of XAI systems by synthesizing effective explanations, they are still largely ignored when prioritizing interpretability. One reason might be that the coverage and specificity of the generated trees or rules are low. In methods based on rule extraction, an opaque 'black box' model is typically trained first and then used to construct a transparent

'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 [47].

2.1.3 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 [48, 49]. One such approach has integrated attention-based explanations into a neural architecture to achieve an efficient computation of tabular data and to increase its interpretability [50]. 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 [51], LRP adaptation [52], or attention memory [53]. Such techniques are promising in enhancing explanations for complex models but the properties of explanation need to be further investigated, especially concerning stability, robustness, and fidelity [54, 39, 55]. 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 [56, 57, 58]. In particular, computational argumentation can be useful to explain all the steps towards a rational decision, as well as enabling reasoning under uncertainty to find solutions with conflictual pieces of information [59, 60, 61]. In this context, rules are seen as arguments, and their interaction is seen as a conflict that can be resolved with argumentation semantics [62]. Typically, computational argumentation implements non-monotonic reasoning, a type of reasoning where conclusions can be retracted in the light of new reasons [63, 64, 65]. This formalism is appealing within XAI because it mirrors one-way human reasoning works [58].

2.2 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.

2.2.1 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 have a significant impact on 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 [66]. In these high-stakes scenarios, there is evidence that AI-based systems can have superior diagnostic capabilities than human experts [67]. 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 [68].

2.2.2 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 optimisation purposes. In these contexts, applying AI often requires transparency and explainability 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. It becomes increasingly difficult for banks, when adopting AI algorithms, to provide explanations which are stable and trustworthy [69, 70] In other words, it becomes increasingly hard to justify the inferences of AI models, both with simpler transparent models [71, 72, 73], and even more with complex models [50, 48, 49]. This lack of transparency 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 [74], which could be subjected to extreme market conditions and unexpected events. If an AI system makes inferences that are difficult to validate, it could lead to disastrous outcomes.

2.2.3 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 important 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 [75]. 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 [76, 77]. 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 in ascertaining the causal factors that underlie them. Even the smallest perturbations in the input data can dramatically affect a model's output, leading to completely different inferences and thus undermining the trustworthiness of such models [78, 79]. Additionally, in these naturalistic environments, a challenge for forest engineering is the development of methods for uncertainty quantification and propagation. In fact, 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, extremely important to analyse the robustness of AI methods—for instance, through explainability—and enhance it for the produced models and their inferences [80, 81].

2.2.4 Education

AI in Education (AIED) focuses on developing AI-powered educational technologies to aid students, instructors, and educational institutions [82, 83, 84] in their teaching and learning activities. On the one hand, for students, AIED has focused on developing models [85], 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 [86]. These are, for example, focused on improving their meta-cognitive processes of self-monitoring, reflection, and planning [87]. On the other hand, for instructors, AIED tools can act as smart teaching assistants [88], help them orchestrate the classroom [89], grade assessments [90], and answer student queries [91], minimizing students dropout [92]. The most recent example of an application of AIED includes the use of some Large Language Model (LLM) capable of generating new textual content based on human input prompts. This can be used for writing essays, producing software code, or generating educational content such as multiple-choice questions or worked 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 certain outputs, and what impact they may have. In particular, AIED tools such as educational recommender systems are increasingly used to automate and personalize learning activities [93]. These tools pose various concerns about their use in high-stakes decisions, including fairness, accountability, transparency, and ethics [94, 95, 96]. The impact of these technologies on students' agency and self-regulated learning is a growing concern, as the lack of transparency 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 [97].

3 Challenges and Research Directions

Despite the many advances, breakthroughs, and potential applications of XAI methods, more research is clearly 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, exactly, XAI is related to trustworthiness. Many surveys tackling some of these aspects of XAI exist and keep appearing in conference proceedings and journals [20]. However, they are rather 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 27 problems, which we have grouped into nine high-level categories—our manifesto—as depicted in 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 concept-based 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.
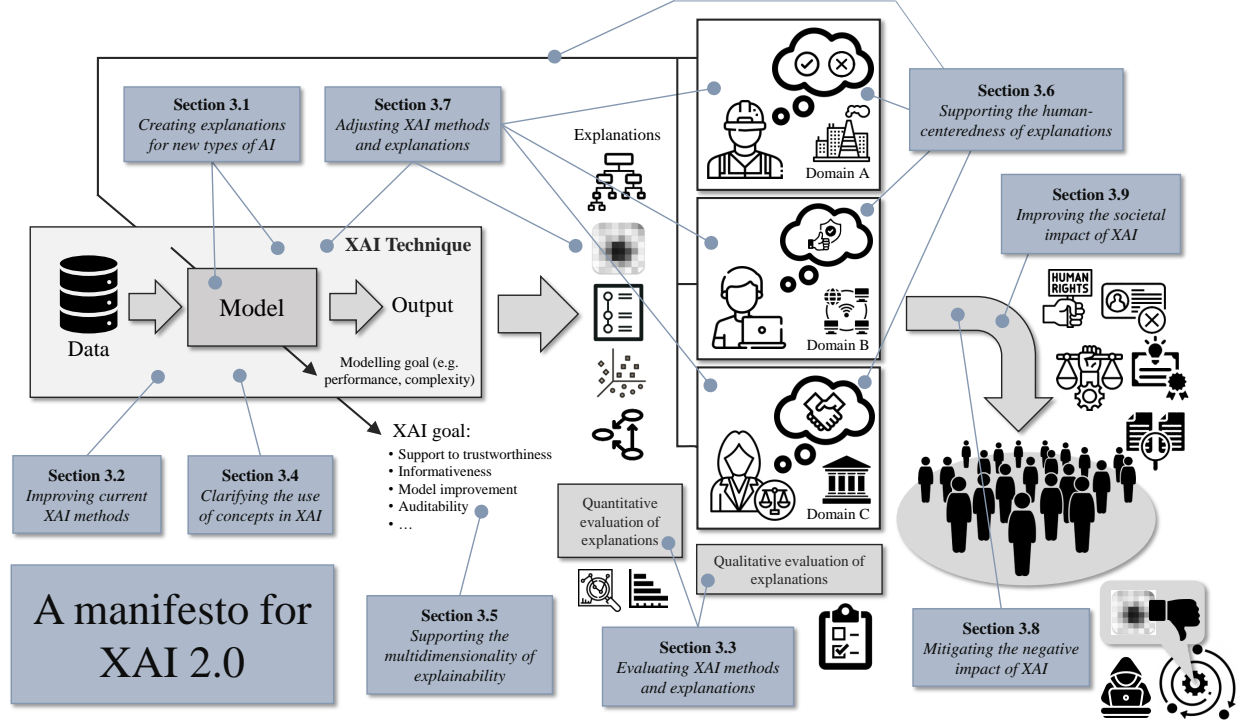
Figure 1: A manifesto for eXplainable Artificial Intelligence (XAI): High-level challenges

3.1.1 Creating Explanations for Generative and Large Language Models

Generative AI models, such as those employed for diffusion denoising [98, 99] or the family of GPT models for large-scale language generation [100], are disrupting many sectors. These models deliver exceptional performance due to their immense scale. With billions, and in some cases, nearly trillions of parameters, however, their sheer size poses a significant challenge to existing XAI methods [14]. 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, which is thought to arise from a superposition of multiple independent features [101]. XAI methods have so far been mostly limited to classification and regression problems. Accordingly, completely 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 explanations for them is very challenging. Another challenge, particularly for LLMs, concerns scaling laws. Neural scaling laws are functional relationships that relate variables associated to a neural network, such as the number of layers in its architecture and its achieved accuracy after training. Such functional relation of two x and y variable is of the general form $y = a \cdot x^{\alpha}$, where a and α are constants of the scaling law. 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. It is an open issue whether scaling laws can be used to infer the quality of the artefacts or concepts learned by LLMs. Even if this were possible, such laws might reveal plateaus of XAI scaling relative to general ability scaling or indicate a gap between the two.

Solution Ideas Mechanistic interpretability [102, 103] is a promising approach to gain deeper insights into the functioning and scaling laws of generative models, such as for grokking mechanics [104] and the ability to solve problems recursively [105]. 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 [106] and sparse autoencoders have been successfully used for the mono-semanticity of Deep Neural Network (DNN) models [101]. There are also challenges in mechanistic interpretability, such as disentangling multiple algorithm implementations and finding unknown algorithms [107]. Furthermore, there are preliminary results from vision models that scaling does not help the mechanistic interpretability

of models [108], calling for designing models for mechanistic interpretability. A potential complement to mechanistic interpretability may be information geometry [109], which can help analyze high-dimensional spaces involved in the processing of LLMs. Furthermore, constraints may have to be imposed on the training and functioning of LLMs to ensure safety as well as explainability [110, 108]. 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 good direction is in [111] which introduces a training procedure that encourages modularity and interpretability by discouraging non-local connections between neurons through local L1 regularization with swaps of neuron locations. Finally, it remains to be seen if these methods can be scaled to relevant models and problem sizes and complexities.

3.1.2   Creating Explanations for Concept-Based Learning Algorithms

Concept-based learning algorithms are another class of new forms of AI for which good XAI methods do not yet exist. Several such algorithms have been proposed over the years to directly learn features that describe 'prototypical concepts' or 'prototypes' present in each input to the model, including ProtoPNet [112], ProtoTree [113], ProtoPShare [114], Concept Bottleneck Models [115], Concept Activation Vectors [116], Concept Embedding Models [117] or Concept Atlases [118]. Neuro-symbolic learning, namely, the symbiosis between connectionist and concept-based symbolic learning, has recently also gained momentum [119, 120, 121]. 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 a priori known concepts and the interrelationships among them [15]. 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 large inductive bias in their explanation-producing process, not properly generalising when explaining new inputs that are distributionally novel with respect to the training data. Furthermore, the continuous proposal of new datasets for concept learning, including Clevr/Clevrer [122, 123], Kandinsky Patterns [124] or Closure [125], sheds evidence on the need for eliciting local explanations that can be formulated in terms of concepts and their spatial distribution.

Solution Ideas   One interesting avenue of research explores the potential for genetically evolvable connections between identifiable concepts in input data using object detection models and evolutionary programming solvers [16]. This hybridization could offer the advantage of employing symbolic classifiers that are interpretable, algorithmically transparent, and well-suited for handling datasets that encapsulate discriminative, concept-wise compositional information. Additionally, there is a growing demand to expand hybrid approaches that unite Knowledge Graphs 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. [54].

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 [126], one of the most prominent classes of XAI methods used for computer vision tasks. Such methods are often based on perturbations [127, 128] or gradients [129, 130]. Despite the great success of these methods to, for instance, detect biases and flaws in the learned prediction strategies (so-called 'Clever Hans Effect' [37], see 2.1), attribution-based explanation methods also have limitations. For instance, saliency masks on the level of pixels are often unsuited for laypersons [39]. A major technical limitation of attribution methods is their sensitivity to 1) internal hyper-parameter tuning and customisation (such as baselines), 2) the manual settings of interpretable interfaces, and 3) the assumption surrounding 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 such as LRP have to use different methods depending on the layer of the DNN. Additionally, some methods have issues with computational efficiency, requiring many passes for calculating attributions.

Solution Ideas   We can tackle the inherent issues in attribution methods by combining them with other approaches to XAI to get a portfolio approach that hedges the weak characteristics of each individual approach. Mechanistic interpretability is an orthogonal approach with different characteristics that could play along well with these approaches. Methods in the portfolio could negotiate like in a market to tune themselves and converge to a majority view, or even better, a list of hypotheses with their plausibilities based on the votes of each portfolio participant.

3.2.2 Removing Artifacts in Synthesis-Based Explanations

Generating explanations through synthesis is a promising direction to advance the field of XAI. While a user is unlikely to directly understand the layer activations of specific classes, it may be different for 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 actually be present to activate the neuron strongly. Two existing methods for synthesis in the literature are a decoder for layer activation [131], and a GAN for single neurons [132]. 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 [131] and the chapter concept vectors in [133]. Most methods explaining concepts rely on a given dataset of human-defined concepts [133], 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.

Solution Ideas   To minimize artefacts, state-of-the-art models and recent popular techniques in DL [134], especially diffusion models, could be leveraged [98]. However, even state-of-the-art generative models do not ensure the absence of artefacts. Thus, to verify whether there are any distortions due to the synthesis, one idea is to compute a reconstruction of the original input serving as a reference [135]. 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 fairly 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.

3.2.3 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 [136] highlights the challenge of creating robust 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 [137, 138], for instance, to also 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 [74, 139, 140]. A satisfying solution, however, does not yet exist. The difficulty lies especially in the fact that for a robust explanation, the model itself must be robust.

Solution Ideas   As a first step towards robust explanations, evaluations on standard 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. For example, a proposal exists to blend uncertainty quantification and XAI [141]. Other research has placed emphasis on the robustness of the AI model itself, for instance, in the form of explanations that inform about model inversion or extraction attacks [142, 143]. 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 [144].

3.3 Evaluating XAI Methods and Explanations

Evaluation is an important aspect of the development and deployment of XAI systems. Evaluating XAI methods, however, is a complex task, and no gold standard exists on what makes for a good explanation [39].

3.3.1 Facilitating Human Evaluation of Explanations

One problem concerning the evaluation of XAI methods is that they often lack user studies. Current evaluation approaches typically only analyze certain properties of the XAI methods themselves without accounting for the interaction with the final user [145, 146, 147, 148, 55]. 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 [149]. Individual differences in understanding, prior knowledge, and the cognitive load required to comprehend explanations add further challenges to evaluating XAI methods. In general, it is difficult

to compare different forms and types of explanations to determine which of them are the most effective. Additionally, users are typically 'passive recipients' of explanations, and the actual usage or exploitation of such explanations is barely tested. For certain properties, there are no approaches at all that test for them [55]. While some studies evaluated the impact of synthesized explanations of AI systems on humans when compared to the scenario where no explanations were provided [150, 151, 152], there is clearly 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 fields like 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 [23, 153]. To streamline the evaluation process, it is imperative to establish standardized frameworks encompassing every stage of it, from formulating hypotheses to data collection, analysis, and utilising 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 not only benefits from theoretical foundations but is also shaped by empirical user-centric research, ultimately enhancing its practical utility.

3.3.2   Creating an Evaluation Framework for XAI Methods

There are several works that address the evaluation of XAI methods. Hoffman et al. [154], for instance, integrate extensive literature and various psychometric assessments to introduce key concepts for measuring the quality of an XAI system. Similarly, Vilone and Longo [12] aggregate evaluation approaches for XAI methods from several scientific studies via a hierarchical system. Furthermore, Van der Lee et al. [155] 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, there is currently a lack of standardized methods and metrics for evaluating XAI systems. In other words, despite the broad interest in the design of XAI methods [156, 145, 8, 10, 157], 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 good 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 [158]. Furthermore, the recently developed XAI evaluation framework Quantus [159] implements over 30 evaluation metrics from six categories. Frameworks such as Quantus allow for the evaluation and comparison of explanations in a standardized and reproducible manner. Furthermore, publicly available XAI evaluation datasets with ground truth information, such as CLEVR-XAI [160], allow for objective evaluations. In the future, artefacts like these need to be extended to more application areas, especially outside of computer vision.

3.3.3   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 generalise well [161]. 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 [150, 151, 152, 162, 163, 164, 165].

Solution Ideas   A potential solution involves augmenting human studies with synthetic data and virtual participants. By creating synthetic datasets that span a wide range of demographic characteristics, behaviours, and preferences, researchers can address the issue of limited sample representativeness. 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. To enhance the reproducibility and rigour of XAI evaluations, standardized methodologies and statistical analyses must be employed. 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.

3.4 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.

3.4.1 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 effective and useful XAI systems. Some researchers have attempted to define an explainable Artificial System such as one that produces details or reasons to make its functioning clear or easy to understand [166]. Other researchers use terms like explainability and interpretability synonymously [153, 167], while others draw major distinctions between them [42]. These differences pose problems for applied research and interdisciplinary collaboration. Discussions about clarifying terms in the field of XAI 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, [168, 169, 170]). On the other hand, some attempt to provide explicit definitions, intending to differentiate between the various terms employed (for example, [171, 172, 148, 173, 174]).

Solution Ideas As the lack of a clear and consistent definition of terms related to explainability can hinder progress in developing effective and useful 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 completely 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, therefore, to identify current usages of them and to create an overview and comparison of them. For instance, some work identifies relevant notions [148], but limited work exists in comparing them. As a next step, the merit of the various proposed definitions must be determined. For this purpose, quality criteria should be established (see, for instance, [171]), which can be consulted to evaluate each proposed definition.

3.4.2 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 [175]: human agency and oversight, technical robustness and safety, privacy awareness and data governance, transparency and explainability, diversity, non-discrimination and fairness, societal and environmental well-being, and accountability. While XAI has the potential to help with most of these [176], it is taken to help with one of them primarily: transparency and explainability. 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 main goal of XAI [166], but XAI is also claimed to be a part of trustworthiness [177]. XAI is purported to change the belief in the trustworthiness of a system [178], while it should also support the trustworthy integration of systems [179]. These are just a few examples, and in other articles, it is also possible to find completely different relationships [176, 180, 148]. One reason for this divergence in descriptions is that there is no uniform way of using terms like trustworthiness (and other terms in XAI, see 3.4.1). 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.

Solution Ideas The relationship between XAI and trustworthiness is widely discussed. We must distinguish between trustworthiness as a property of an AI system, trustworthy AI as a technology enabler to accomplish a responsible and safe AI, and the technical requirements required for an AI system to be trustworthy. XAI is identified as one of the seven trustworthy AI requirements [175, 181]. On the other hand, XAI must contribute towards achieving trustworthiness. Currently, it is necessary to connect XAI with the fundamental properties of AI trustworthiness for AI risk management and the AI lifecycle, measuring their presence and impact. In this regard, we must highlight the report recently published by the UC Berkeley Center for Long-Term Cybersecurity (CLTC) [1]. This report aims to help

---

[1] `https://cltc.berkeley.edu/publication/a-taxonomy-of-trustworthiness-for-artificial-intelligence/` (Last access: 2023/10/28)

organizations develop and deploy more trustworthy AI technologies, including 150 properties related to one of the seven "characteristics of trustworthiness" 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 important aspect is AI governance [182] and the need of governance measures linked to the importance of managing AI risks, another scenario where XAI becomes of utmost importance. These are new fundamental scenarios posing essential challenges for the design, development, and safe deployment of responsible AI systems [181]. In the current debate, XAI is identified as a vital technology to decrease the uncertainty and worry about AI systems in the society.

3.4.3 Finding a Useful Account of Understanding

Another challenge to bring about conceptual clarity is finding a useful 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 [183, 184]. The other end of the spectrum is occupied by philosophers who allow other paths to understanding, even if they offer distorted or false accounts of their targets [185, 186]. Finally, intermediate views exist, allowing that some, but not all of the pieces of information used to provide understanding can be false [187, 188, 189]. There is no consensus regarding which of these views is more adequate in the context of AI explanations. For example, while [190] sides with the traditional view, [24] adopts a more pragmatic stance. Another obstacle arises from the fact that the understanding provided by XAI methods that account for singular predictions need not be the same type of understanding provided by proxy or surrogate models that offer 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' [191], while proxy or surrogate models provide 'objectual explanations' [192] of their targets. The relation between the two types of understanding requires clarification, not only from a philosophical perspective but also from the point of view of psychology. In addition, there is a third type of understanding that depends entirely on the functional correlations between inputs and outputs [193]. 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 [24, 194, 195, 196, 190, 197, 198, 199, 26, 25] 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 "the inmates running the asylum" [23, p. 36]. The proposed and endorsed alternative is to incorporate results from psychology and philosophy to XAI [17, 200, 153, 201]. Existing theories of how people formulate questions and how they select and evaluate answers should inform the discussion [153].

3.5 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.

3.5.1 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 not accurate, not lawful, or not fair. 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 transparency and 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.

---

[2] lhttps://www.nist.gov/itl/ai-risk-management-framework (Last access: 2023/10/28)

11

Solution Ideas   One approach to such multi-faceted explanations could involve developing trustworthiness metrics that encapsulate dimensions like safety, fairness, and accountability. 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 aims to ensure that trustworthiness considerations are an integral part of the XAI process, leading to multi-faceted explanations. A complementary way to tackle the multidimensionality of explanation concerns its operationalisation which should be performed as it happens with other psychological constructs such as 'intelligence' or 'cognitive load' [165]. 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 main 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 the use of quantitative methods for greater reproducibility, replicability and falsifiability.

3.5.2   Enabling Interdisciplinary Work in XAI

XAI is an interdisciplinary research field [17, 18]. For example, through the collaboration of philosophers and computer scientists, XAI is envisioned to ensure the ethical use of AI [18]. 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 own 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 certain terms [172]. This can lead to confusion and difficulty adapting to different usage in XAI. Eventually, for terms for which there is no common usage, different disciplines may establish their own 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.

3.6   Supporting the Human-Centeredness of Explanations

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

3.6.1   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 [153]. Confalonieri et al. discussed further properties for explanations, including integrating symbolic knowledge and statistical approaches to explainability [3, 202]. 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 [203, 129, 130], 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 [23], and the presentation of explanations in the form of pixel attributions may not be comprehensible to laypersons [39]. 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 [204, 205], 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 the face of a person 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 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 (that means, explaining individual predictions) with global XAI methods (that means, explaining the whole model) might lead to semantically richer and more human-understandable explanations. This '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 individual input (local) [118]. 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 the way in which 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 [206, 60, 207, 208, 209, 57, 58, 59]. Another possible solution to create human-understandable explanations is to map explanations to a more interpretable domain. For instance, one approach to providing more interpretable explanations on time series data has been recently explored in [210]. 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, by using autoencoders [211], for this and other domains.

3.6.2   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 are hard to grasp for humans. On the other hand, humans use concepts that are coarse-grained representations of reality [212, 213]. 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 [35, 36]. 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 [214]). However, explanations based on human-understandable concepts are still in early development. In particular, concept-based explanations are mostly elaborated only for classification or regression models, leaving aside other problems and models in which concept-based explanations could be useful. 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 [215]. 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 [216]. Accordingly, using a method that is generalisable and applicable across diverse areas and different 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 that takes into account a broad range of sub-problems. It begins with finding reliable ways for the extraction and identification of 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 real-world, comprehensible terms. Next, the concepts must be personalized so that they are tailored to the individual consuming them. Allowing users to define their own concepts would be one way

to ensure personalization. Interdisciplinary collaboration and continuous feedback loops could refine these concepts, making them more meaningful and interpretable. A supplementary avenue could be to organize concepts within a hierarchical structure. This structure could be useful for delivering explanations that can be tailored to 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, which could be tackled by introducing quality metrics for concept-based explanations. Finally, the application of XAI methods based on concept synthesis in different domains and applications is another sub-problem. This might be solved by personalization, as described above.

3.6.3   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 are becoming smarter, their decision-making is becoming more intricate. AI systems might start using concepts that are impossible to convey to humans [217, 218]. This means that the concepts humans use to understand the world might no longer suffice to describe reality in a meaningful and useful way [219]. 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 [220, 110].

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 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 information from ontologies. 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 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 [47]. This comprehensive approach aims to improve the alignment of AI explainability with the dynamic nature of real-world scenarios.

3.6.4   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 [27, 10, 221]. 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 [222]. Explanations which are purely based on correlations can hinder decision-making when a model's outputs contain important 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, especially in scenarios where decisions may affect people. To address this issue, counterfactual generation methods for ML methods have garnered attention [147]. Contrary to most XAI approaches, counterfactuals attempt to answer why a black box model leads to a certain prediction by helping users understand what would need to change at its output to achieve a desired result [223]. In this answer, several desired properties should be met, namely, proximity, plausibility, sparsity, diversity, and feasibility [147]. 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 [224].

---

[3]The EU AI Act: `https://artificialintelligenceact.eu/the-act/`

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 of the desired properties for explanations such as proximity, plausibility, sparsity, diversity, and feasibility. This has been approached recently in [225], where counterfactuals are produced by means of 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 [226, 227], 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 automatically construct causal graphs for models that do not necessarily operate on concepts or entities but rather 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.

3.7 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 have to 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

An explanation can be required by many different kinds of stakeholders during the development, evaluation, and use of an AI system [17]. Each stakeholder brings their own attitudes, preferences, aptitudes, abilities, and previous experiences that influence the kind of explanation they require. Designing and tailoring explanations that are appropriate for each of these stakeholder types, both in terms of *content* as well as *format and presentation*, is an ongoing challenge. For example, in the business context, the same objective facts must be explained and tailored to the stakeholders' respective interests and objectives. A business person is usually mostly interested in the bottom line impact of an AI system, a technical person is interested in the process and the validity of implementation, 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 [228] and, more specifically, creating personalized explanations [229] 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 [230, 231].

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 greatly 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 on whose peculiarities and context they are built. In this line of thought, research is starting to emerge focused on distinguishing between high-stakes and low-stakes domains [232, 233, 17]. 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 [218, 234, 17]. 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 [19, 235, 236, 17]. 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 stakeholder that has 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 for transparency 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 need to be avoided or mitigated.

3.8.1 Mitigating Failed Support by XAI

In some domains, especially in the medical domain [237], the ineffective support by XAI can sometimes be harmful. This has been associated with the so-called 'white-box paradox' [238, 239], 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 [238]. 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 [240].

Solution Ideas A first option would be to detect failure situations and label them 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, as for instance, in radiological settings (see [237, 239]). 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 originally conceived as *generative* and *persuasive* technologies [241], not oracular ones. This oracular nature can be characterized as an *alethic* nature, which assumes that machines can, and should, always state the truth [242]. 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 makes the final decision, offering a more balanced perspective than the oracular approach. [243, 242, 244]

3.8.2 Devising Criteria for the Falsifiability of Explanations

Explanations are often requested to clarify issues such as accountability [217, 245, 201, 246]. 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 interpretability. Therefore, the question of establishing what are ground truths for explainability in benchmarks and how they were produced are open questions. As a more ambitious goal, we may ask about the discriminability between very 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 [247]. 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 [248]. 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 important aspect of human coordination with machines [249, 220, 236, 250]. 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 providing a clear explanation of the logic and reasoning behind these outcomes. The effectiveness and adequacy of explainability as a tool for AI safety may be limited in certain scenarios [251]. For instance, AI systems in the hands of malicious human actors, can pose significant challenges to explainability [220] through manipulation and adversarial attacks. As an example, employers may systematically discriminate job applicants by 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 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 [131]. 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 [252]. Similarly, another application context includes forensic analysis, which aims to understand the concepts learned by a classifier [253]. Concept-based explanations could also be helpful for reflective learning from data [254], 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 important aspect of AI safety. Many of the challenges, highlighted in the literature [255, 110] 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 [256]. 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 work for us and tell us 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 [110].

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 3.6.3). The effectiveness and adequacy of explainability as a tool for AI safety may be limited in certain scenarios [251]. For this reason, explainability should be only one part of every safety toolkit as it has its strengths and limitations that need to be complemented in a portfolio of approaches. Work on building such a portfolio is welcome, as there is a growing need. 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 powerful super-intelligent systems without strong 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 with respect to 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 [257, 258]. Likewise, plagiarism detection is becoming central in LLMs that excel across different domains, for example, with ChatGPT [259]). The massive usage of these models to produce apparently original textual content has disrupted the idea of plagiarism, as such content has been shown to evade mainstream tools for plagiarism detection easily. 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 generalisation 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* [260], in the produced content [261]. 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 [262] has been on the rise [263], it is unclear how to efficiently ensure that data owned by a certain 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 transparency 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. In general, humans should be granted the chance 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 guaranteeing more transparency 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 [264, 265, 218, 266]. 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 certain 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* [267] 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 [268, 269]. 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 [270, 271]. 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 [272]. Current XAI methods answer this issue, but they can address only a small part of the problem [273].

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 [272, 274]. There are different modalities of participative approach to AI design, but an essential consideration is the participative impact assessment [275]. Vulnerable impacted stakeholders should be included, through an open and circular approach, in the key value-sensitive decisions in the AI design [276]. Following the example of environmental impact assessment or workers' participation in business decision-making [277], 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 succinctly 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 concept-based learning algorithms.

2. **Improving (and Augmenting) Current XAI Methods**: To augment and improve attribution methods, remove artefacts in synthesis-based explanations, and create robust explanations.

3. **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.

4. **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.

5. **Supporting the Multi-Dimensionality of Explainability:** To create multi-faceted explanations and enable interdisciplinary work in XAI.

6. **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.

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 certainly inspire more scholars to join us in advancing eXplainable Artificial Intelligence as a field. This manifesto is a genuine attempt at it, an exciting opportunity for shaping the future of AI-based systems for the benefit of human society.

References

[1] William Swartout, Cecile Paris, and Johanna Moore. Explanations in knowledge systems: Design for explainable expert systems. *IEEE Expert*, 6(3):58–64, 1991.

[2] Cécile 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*, pages 49–82. Springer, 1991.

[3] Roberto Confalonieri, Ludovik Coba, Benedikt Wagner, and Tarek R. Besold. A historical perspective of explainable artificial intelligence. *WIREs Data Mining and Knowledge Discovery*, 11(1):e1391, 2021.

[4] Timo Speith. A review of taxonomies of explainable artificial intelligence (XAI) methods. In Charles Isbell, Seth Lazar, Alice Oh, and Alice Xiang, editors, *Proceedings of the 5th ACM Conference on Fairness, Accountability, and Transparency*, pages 2239–2250, New York, NY, USA, 2022. Association for Computing Machinery.

[5] Nava Tintarev and Judith Masthoff. A Survey of Explanations in Recommender Systems. In *Proceedings of the 2007 IEEE 23rd International Conference on Data Engineering Workshop*, pages 801–810. IEEE Computer Society, 2007.

[6] Supriyo Chakraborty, Richard Tomsett, Ramya Raghavendra, Daniel Harborne, Moustafa Alzantot, Federico Cerutti, Mani B. Srivastava, Alun D. Preece, Simon Julier, Raghuveer M. Rao, Troy D. Kelley, Dave Braines, Murat Sensoy, Christopher J. Willis, and Prudhvi Gurram. Interpretability of deep learning models: A survey of results. In *2017 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computed, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation, SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI 2017, San Francisco, CA, USA, August 4-8, 2017*, pages 1–6. IEEE, 2017.

[7] Or Biran and Courtenay Cotton. Explanation and justification in machine learning: A survey. In David W Aha, Trevor Darrell, Michael Pazzani, Darryn Reid, Claude Sammut, and Peter Stone, editors, *Proceedings of the IJCAI 2017 Workshop on Explainable Artificial Intelligence*, pages 8–13, 2017.

[8] Amina Adadi and Mohammed Berrada. Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access*, 6:52138–52160, 2018.

[9] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A survey of methods for explaining black box models. *ACM Comp. Surv.*, 51(5):1–42, 2018.

[10] Diogo V. Carvalho, Eduardo M. Pereira, and Jaime S. Cardoso. Machine learning interpretability: A survey on methods and metrics. *Electronics*, 8(8), 2019.

[11] Aniek F. Markus, Jan A. Kors, and Peter 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. *Journal of Biomedical Informatics*, 113:11, 2021.

[12] Giulia Vilone and Luca Longo. Notions of explainability and evaluation approaches for explainable artificial intelligence. *Information Fusion*, 76:89–106, 2021.

[13] Jianlong Zhou, Amir H. Gandomi, Fang Chen, and Andreas Holzinger. Evaluating the quality of machine learning explanations: A survey on methods and metrics. *Electronics*, 10(5), 2021.

[14] Julia El Zini and Mariette Awad. On the explainability of natural language processing deep models. *ACM Computing Surveys*, 55(5):1–31, 2022.

[15] Ilaria Tiddi and Stefan Schlobach. Knowledge graphs as tools for explainable machine learning: A survey. *Artificial Intelligence*, 302:103627, 2022.

[16] Yi Mei, Qi Chen, Andrew Lensen, Bing Xue, and Mengjie Zhang. Explainable artificial intelligence by genetic programming: A survey. *IEEE Transactions on Evolutionary Computation*, 2022.

[17] Markus Langer, Daniel Oster, Timo Speith, Holger Hermanns, Lena Kästner, Eva Schmidt, Andreas Sesing, and Kevin 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.

[18] Markus Langer, Kevin Baum, Kathrin Hartmann, Stefan Hessel, Timo Speith, and Jonas Wahl. Explainability auditing for intelligent systems: A rationale for multi-disciplinary perspectives. In Tao Yue and Mehdi Mirakhorli, editors, *29th IEEE International Requirements Engineering Conference Workshops*, REW 2021, pages 164–168, Piscataway, NJ, USA, 2021. IEEE.

[19] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

[20] Sajid Ali, Tamer Abuhmed, Shaker El-Sappagh, Khan Muhammad, Jose M Alonso-Moral, Roberto Confalonieri, Riccardo Guidotti, Javier Del Ser, Natalia Díaz-Rodríguez, and Francisco Herrera. Explainable artificial intelligence (xai): What we know and what is left to attain trustworthy artificial intelligence. *Information Fusion*, page 101805, 2023.

[21] Longbing Cao. Ai in finance: Challenges, techniques, and opportunities. *ACM Comput. Surv.*, 55(3), feb 2022.

[22] Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie 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, pages 1721–1730, New York, NY, USA, August 2015. Association for Computing Machinery.

[23] Tim Miller, Piers Howe, and Liz Sonenberg. Explainable AI: Beware of inmates running the asylum. or: How I learnt to stop worrying and love the social and behavioural sciences. In David W. Aha, Trevor Darrell, Michael Pazzani, Darryn Reid, Claude Sammut, and Peter Stone, editors, *Proceedings of the IJCAI 2017 Workshop on Explainable Artificial Intelligence*, pages 36–42, Santa Clara County, CA, USA, 2017. IJCAI.

[24] Andrés Páez. The pragmatic turn in explainable artificial intelligence (XAI). *Minds and Machines*, 29(3):441–459, 2019.

[25] Maximilian A. Köhl, Kevin Baum, Dimitri Bohlender, Markus Langer, Daniel Oster, and Timo Speith. Explainability as a non-functional requirement. In Daniela E. Damian, Anna Perini, and Seok-Won Lee, editors, *IEEE 27th International Requirements Engineering Conference*, RE 2019, pages 363–368, Piscataway, NJ, USA, 2019. IEEE.

[26] Larissa Chazette, Wasja Brunotte, and Timo Speith. Exploring explainability: A definition, a model, and a knowledge catalogue. In Jane Cleland-Huang, Ana Moreira, Kurt Schneider, and Michael Vierhauser, editors, *IEEE 29th International Requirements Engineering Conference*, RE 2021, pages 197–208, Piscataway, NJ, USA, 2021. IEEE.

[27] Amina Adadi and Mohammed Berrada. Peeking inside the black-box: A survey on explainable artificial intelligence (xai). *IEEE Access*, 6, 2018.

[28] Atul Rawal, James McCoy, Danda B Rawat, Brian M Sadler, and Robert St Amant. Recent advances in trustworthy explainable artificial intelligence: Status, challenges, and perspectives. *IEEE Transactions on Artificial Intelligence*, 3(6):852–866, 2021.

[29] Fabian Hinder and Barbara Hammer. Counterfactual explanations of concept drift. *arXiv preprint arXiv:2006.12822*, 2020.

[30] Afsana Khan, Marijn ten Thij, and Anna Wilbik. Vertical federated learning: A structured literature review. *arXiv preprint arXiv:2212.00622*, 2022.

[31] Seul-Ki Yeom, Philipp Seegerer, Sebastian Lapuschkin, Alexander Binder, Simon Wiedemann, Klaus-Robert Müller, and Wojciech Samek. Pruning by explaining: A novel criterion for deep neural network pruning. *Pattern Recognition*, 115:107899, 2021.

[32] Wojciech Samek, Grégoire Montavon, Sebastian Lapuschkin, Christopher J Anders, and Klaus-Robert Müller. Explaining deep neural networks and beyond: A review of methods and applications. *Proceedings of the IEEE*, 109(3):247–278, 2021.

[33] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 1135–1144, New York, NY, USA, 2016. ACM.

[34] Scott M Lundberg and Su-In Lee. A Unified Approach to Interpreting Model Predictions. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30, page 4768–4777. Curran Associates, Inc., 2017.

[35] Damien Garreau and Ulrike Luxburg. Explaining the explainer: A first theoretical analysis of lime. In *International Conference on Artificial Intelligence and Statistics*, pages 1287–1296. PMLR, 2020.

[36] Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju. Fooling lime and shap: Adversarial attacks on post hoc explanation methods. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 180–186, 2020.

[37] Sebastian Lapuschkin, Stephan Wäldchen, Alexander Binder, Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. Unmasking clever hans predictors and assessing what machines really learn. *Nature Communications*, 10(1):1096, 2019.

[38] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, Nov 2020.

[39] Timo Speith. How to evaluate explainability – a case for three criteria. In Eric Knauss, Gunter Mussbacher, Chetan Arora, Muneera Bano, and Jean-Guy Schneider, editors, *Proceedings of the 30th IEEE International Requirements Engineering Conference Workshops*, REW 2022, pages 92–97, Piscataway, NJ, USA, 2022. IEEE.

[40] Sebastian Lapuschkin, Alexander Binder, Klaus-Robert Müller, and Wojciech Samek. Understanding and comparing deep neural networks for age and gender classification. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 1629–1638, 2017.

[41] Léo Grinsztajn, Edouard Oyallon, and Gaël Varoquaux. Why do tree-based models still outperform deep learning on typical tabular data? In *NeurIPS*, 2022.

[42] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, 1(5):206–215, 2019.

[43] Barnaby Crook, Maximilian Schlüter, and Timo Speith. Revisiting the performance-explainability trade-off in explainable artificial intelligence (XAI). In Fabiano Dalpiaz, Jennifer Horkoff, and Kurt Schneider, editors, *Proceedings of the 31st IEEE International Requirements Engineering Conference Workshops*, pages 316–324, Piscataway, NJ, USA, 2023. IEEE.

[44] Lior Rokach. Decision forest: Twenty years of research. *Information Fusion*, 27:111–125, 2016.

[45] Julian Hatwell, Mohamed Medhat Gaber, and R Muhammad Atif Azad. Chirps: Explaining random forest classification. *Artificial Intelligence Review*, 53:5747–5788, 2020.

[46] Johannes Fürnkranz, Tomáš Kliegr, and Heiko Paulheim. On cognitive preferences and the plausibility of rule-based models. *Machine Learning*, 109(4):853–898, 2020.

[47] David C. Krakauer. Unifying complexity science and machine learning. *Frontiers in Complex Systems*, 1, 2023.

[48] Xin Huang, Ashish Khetan, Milan Cvitkovic, and Zohar Karnin. Tabtransformer: Tabular data modeling using contextual embeddings. *arXiv preprint arXiv:2012.06678*, 2020.

[49] Yury Gorishniy, Ivan Rubachev, Valentin Khrulkov, and Artem Babenko. Revisiting deep learning models for tabular data. *Advances in Neural Information Processing Systems*, 34:18932–18943, 2021.

[50] Sercan Ö Arik and Tomas Pfister. Tabnet: Attentive interpretable tabular learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(8):6679–6687, 2021.

[51] Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4190–4197, Online, July 2020. Association for Computational Linguistics.

[52] Ameen Ali, Thomas Schnake, Oliver Eberle, Grégoire Montavon, Klaus-Robert Müller, and Lior Wolf. Xai for transformers: Better explanations through conservative propagation. In *International Conference on Machine Learning*, pages 435–451. PMLR, 2022.

[53] Mayukh Deb, Björn Deiseroth, Samuel Weinbach, Patrick Schramowski, and Kristian Kersting. Atman: Understanding transformer predictions through memory efficient attention manipulation. *arXiv preprint arXiv:2301.08110*, 2023.

[54] Freddy Lécué. On the role of knowledge graphs in explainable AI. *Semantic Web*, 11(1):41–51, 2020.

[55] Timo Speith and Markus Langer. A new perspective on evaluation methods for explainable artificial intelligence (xai). In Fabiano Dalpiaz, Jennifer Horkoff, and Kurt Schneider, editors, *Proceedings of the 31st IEEE International Requirements Engineering Conference Workshops*, pages 325–331, Piscataway, NJ, USA, 2023. IEEE.

[56] Kristijonas Čyras, Antonio Rago, Emanuele Albini, Pietro Baroni, and Francesca Toni. Argumentative xai: A survey. In Zhi-Hua Zhou, editor, *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 4392–4399. International Joint Conferences on Artificial Intelligence Organization, 8 2021. Survey Track.

[57] Kevin Baum, Holger Hermanns, and Timo Speith. From machine ethics to machine explainability and back. In Martin Charles, Dimitrios I. Diochnos, Jürgen Dix, Frederick Hoffman, and Guillermo R. Simari, editors, *International Symposium on Artificial Intelligence and Mathematics*, pages 1–8, Fort Lauderdale, FL, USA, 2018. International Symposium on Artificial Intelligence and Mathematics.

[58] Kevin Baum, Holger Hermanns, and Timo Speith. Towards a framework combining machine ethics and machine explainability. In Bernd Finkbeiner and Samantha Kleinberg, editors, *Proceedings of the 3rd Workshop on Formal Reasoning about Causation, Responsibility, and Explanations in Science and Technology*, pages 34–49, Sydney, NSW, AU, 2018. Electronic Proceedings in Theoretical Computer Science.

[59] Alexandros Vassiliades, Nick Bassiliades, and Theodore Patkos. Argumentation and explainable artificial intelligence: a survey. *The Knowledge Engineering Review*, 36:e5, 2021.

[60] Luca Longo. Argumentation for knowledge representation, conflict resolution, defeasible inference and its integration with machine learning. *Machine Learning for Health Informatics: State-of-the-Art and Future Challenges*, pages 183–208, 2016.

[61] Zhiwei Zeng, Chunyan Miao, Cyril Leung, and Jing Jih Chin. Building more explainable artificial intelligence with argumentation. In Sheila A. McIlraith and Kilian Q. Weinberger, editors, *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*, pages 8044–8046, Palo Alto, CA, USA, 2018. AAAI Press.

[62] Pietro Baroni, Martin Caminada, and Massimiliano Giacomin. An introduction to argumentation semantics. *The knowledge engineering review*, 26(4):365–410, 2011.

[63] Lucas Rizzo and Luca 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, AI³@AI*IA 2018, 20-23 November 2018, Trento, Italy*, pages 11–26, 2018.

[64] Lucas Rizzo, Ljiljana Majnaric, and Luca 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*, pages 197–209. Springer, 2018.

[65] Luca Longo, Lucas Rizzo, and Pierpaolo Dondio. Examining the modelling capabilities of defeasible argumentation and non-monotonic fuzzy reasoning. *Knowledge-Based Systems*, 211:106514, 2021.

[66] Shahab S Band, Atefeh Yarahmadi, Chung-Chian Hsu, Meghdad Biyari, Mehdi Sookhak, Rasoul Ameri, Iman Dehzangi, Anthony Theodore Chronopoulos, and Huey-Wen Liang. Application of explainable artificial intelligence in medical health: A systematic review of interpretability methods. *Informatics in Medicine Unlocked*, 40:101286, 2023.

[67] Philipp Tschandl, Noel Codella, Bengü Nisa Akay, Giuseppe Argenziano, Ralph P Braun, Horacio Cabo, David Gutman, Allan Halpern, Brian Helba, Rainer 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. *The lancet oncology*, 20(7):938–947, 2019.

[68] Julia Amann, Alessandro Blasimme, Effy Vayena, Dietmar Frey, and Vince I Madai. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. *BMC medical informatics and decision making*, 20(1):1–9, 2020.

[69] Tessa Han, Suraj Srinivas, and Himabindu Lakkaraju. Which explanation should i choose? a function approximation perspective to characterizing post hoc explanations. *Advances in neural information processing systems*, 35, 2020.

[70] Chirag Agarwal, Satyapriya Krishna, Eshika Saxena, Martin Pawelczyk, Nari Johnson, Isha Puri, Marinka Zitnik, and Himabindu Lakkaraju. Openxai: Towards a transparent evaluation of model explanations. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 15784–15799. Curran Associates, Inc., 2022.

[71] Niklas Bussmann, Paolo Giudici, Dimitri Marinelli, and Jochen Papenbrock. Explainable machine learning in credit risk management. *Computational Economics*, 57:203–216, 2021.

[72] Swati Sachan, Jian-Bo Yang, Dong-Ling Xu, David Eraso Benavides, and Yang Li. An explainable ai decision-support-system to automate loan underwriting. *Expert Systems with Applications*, 144:113100, 2020.

[73] Cynthia Rudin and Joanna Radin. Why are we using black box models in ai when we don't need to? a lesson from an explainable ai competition. *Harvard Data Science Review*, 1(2):10–1162, 2019.

[74] Saumitra Mishra, Sanghamitra Dutta, Jason Long, and Daniele Magazzeni. A survey on the robustness of feature importance and counterfactual explanations. *arXiv preprint arXiv:2111.00358*, 2021.

[75] Sadikshya Sharma, Sambandh Dhal, Tapas Rout, and Bharat Sharma Acharya. Drones and machine learning for estimating forest carbon storage. *Carbon Research*, 1(1):21, 2022.

[76] Torsten B Möllmann and Bernhard Möhring. A practical way to integrate risk in forest management decisions. *Annals of Forest Science*, 74:1–12, 2017.

[77] Christoph Gollob, Tim Ritter, and Arne Nothdurft. Forest inventory with long range and high-speed personal laser scanning (pls) and simultaneous localization and mapping (slam) technology. *Remote Sensing*, 12(9):1509, 2020.

[78] Andreas Holzinger, Anna Saranti, Alessa Angerschmid, Carl Orge Retzlaff, Andreas Gronauer, Vladimir Pejakovic, Francisco Medel-Jimenez, Theresa Krexner, Christoph Gollob, and Karl Stampfer. Digital transformation in smart farm and forest operations needs human-centered ai: challenges and future directions. *Sensors*, 22(8):3043, 2022.

[79] A Holzinger, K Stampfer, A Nothdurft, C Gollob, and P Kieseberg. Challenges in artificial intelligence for smart forestry. *Eur. Res. Consort. Informatics Math.(ERCIM) News*, 130:40–41, 2022.

[80] Andreas 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*, pages 427–440. Springer, 2022.

[81] Andreas Holzinger, Matthias Dehmer, Frank Emmert-Streib, Rita Cucchiara, Isabelle Augenstein, Javier Del Ser, Wojciech Samek, Igor Jurisica, and Natalia Díaz-Rodríguez. Information fusion as an integrative cross-cutting enabler to achieve robust, explainable, and trustworthy medical artificial intelligence. *Information Fusion*, 79:263–278, 2022.

[82] Rose Luckin, Wayne Holmes, Mark Griffiths, and Laurie B Forcier. Intelligence unleashed: An argument for ai in education. Technical report, The Open University, 2016.

[83] Olaf Zawacki-Richter, Victoria I Marín, Melissa Bond, and Franziska Gouverneur. Systematic review of research on artificial intelligence applications in higher education–where are the educators? *International Journal of Educational Technology in Higher Education*, 16(1):1–27, 2019.

[84] Luca Longo. Empowering qualitative research methods in education with artificial intelligence. In António Pedro Costa, Luís Paulo Reis, and António Moreira, editors, *Computer Supported Qualitative Research*, pages 1–21, Cham, 2020. Springer International Publishing.

[85] Michel C Desmarais and Ryan SJ d Baker. A review of recent advances in learner and skill modeling in intelligent learning environments. *User Modeling and User-Adapted Interaction*, 22:9–38, 2012.

[86] Kurt VanLehn. The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational psychologist*, 46(4):197–221, 2011.

[87] Susan Bull. There are open learner models about! *IEEE Transactions on Learning Technologies*, 13(2):425–448, 2020.

[88] Boulay B. (du). Artificial intelligence as an effective classroom assistant. *IEEE Intelligent Systems*, 31(6):76–81, 2016.

[89] Kenneth Holstein, Bruce M McLaren, and Vincent Aleven. Co-designing a real-time classroom orchestration tool to support teacher-ai complementarity. *Grantee Submission*, 2019.

[90] Arjun Singh, Sergey Karayev, Kevin Gutowski, and Pieter 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*, pages 81–88, 2017.

[91] Guruswami Hiremath, Aishwarya Hajare, Priyanka Bhosale, Rasika Nanaware, and KS Wagh. Chatbot for education system. *International Journal of Advance Research, Ideas and Innovations in Technology*, 4(3):37–43, 2018.

[92] Martín Liz-Domínguez, Manuel Caeiro-Rodríguez, Martín Llamas-Nistal, and Fernando A Mikic-Fonte. Systematic literature review of predictive analysis tools in higher education. *Applied Sciences*, 9(24):5569, 2019.

[93] Hassan Khosravi, Kirsty Kitto, and Williams Joseph. Ripple: A crowdsourced adaptive platform for recommendation of learning activities. *Journal of Learning Analytics*, 6(3):91–105, 2019.

[94] Wayne Holmes, Kaska Porayska-Pomsta, Ken Holstein, Emma Sutherland, Toby Baker, Simon Buckingham Shum, Olga C Santos, Mercedes T Rodrigo, Mutlu Cukurova, Ig Ibert Bittencourt, et al. Ethics of ai in education: Towards a community-wide framework. *International Journal of Artificial Intelligence in Education*, pages 1–23, 2021.

[95] Ryan S Baker and Aaron Hawn. Algorithmic bias in education. *International Journal of Artificial Intelligence in Education*, pages 1–41, 2021.

[96] René F Kizilcec and Hansol Lee. Algorithmic fairness in education. In *The ethics of artificial intelligence in education*, pages 174–202. Routledge, 2022.

[97] Solmaz Abdi, Hassan Khosravi, Shazia Sadiq, and Dragan Gasevic. Complementing educational recommender systems with open learner models. In *Proceedings of the tenth international conference on learning analytics & knowledge*, pages 360–365, 2020.

[98] Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, and Mubarak Shah. Diffusion models in vision: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(9):10850–10869, sep 2023.

[99] Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. Diffusion models: A comprehensive survey of methods and applications. *ACM Comput. Surv.*, sep 2023. Just Accepted.

[100] M. Onat Topal, Anil Bas, and Imke van Heerden. Exploring transformers in natural language generation: Gpt, bert, and xlnet. *ArXiv*, abs/2102.08036, 2021.

[101] Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, and Adam Jermyn. Towards Monosemanticity: Decomposing Language Models With Dictionary Learning. Technical report, Anthropic, October 2023.

[102] Nick Cammarata, Shan Carter, Gabriel Goh, Chris Olah, Michael Petrov, Ludwig Schubert, Chelsea Voss, Ben Egan, and Swee Kiat Lim. Thread: Circuits. *Distill*, 2020.

[103] Nelson Elhage, Neel Nanda, Catherine Olsson, and Tom Henighan. A Mathematical Framework for Transformer Circuits. Technical report, Anthropic, December 2021.

[104] Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. Progress measures for grokking via mechanistic interpretability, January 2023.

[105] Shizhuo Dylan Zhang, Curt Tigges, Stella Biderman, Maxim Raginsky, and Talia Ringer. Can Transformers Learn to Solve Problems Recursively?, June 2023. arXiv:2305.14699 [cs].

[106] Sid Black, Lee Sharkey, Leo Grinsztajn, Eric Winsor, Dan Braun, Jacob Merizian, Kip Parker, Carlos Ramón Guevara, Beren Millidge, Gabriel Alfour, and Connor Leahy. Interpreting neural networks through the polytope lens, November 2022.

[107] Ziqian Zhong, Ziming Liu, Max Tegmark, and Jacob Andreas. The Clock and the Pizza: Two Stories in Mechanistic Explanation of Neural Networks, June 2023.

[108] Roland S. Zimmermann, Thomas Klein, and Wieland Brendel. Scale Alone Does not Improve Mechanistic Interpretability in Vision Models, July 2023.

[109] Shun-ichi Amari. *Information Geometry and Its Applications*. Springer, February 2016. Google-Books-ID: UkSFCwAAQBAJ.

[110] Mario Brcic and Roman V. Yampolskiy. Impossibility Results in AI: A Survey. *ACM Computing Surveys*, 56(1):8:1–8:24, August 2023.

[111] Ziming Liu, Eric Gan, and Max Tegmark. Seeing is Believing: Brain-Inspired Modular Training for Mechanistic Interpretability, June 2023. arXiv:2305.08746 [cond-mat, q-bio].

[112] Chaofan Chen, Oscar Li, Daniel Tao, Alina Barnett, Cynthia Rudin, and Jonathan K Su. This looks like that: deep learning for interpretable image recognition. *Advances in neural information processing systems*, 32, 2019.

[113] Meike Nauta, Ron Van Bree, and Christin Seifert. Neural prototype trees for interpretable fine-grained image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14933–14943, 2021.

[114] Dawid Rymarczyk, Łukasz Struski, Jacek Tabor, and Bartosz 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*, pages 1420–1430, 2021.

[115] Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *International Conference on Machine Learning*, pages 5338–5348. PMLR, 2020.

[116] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In *International conference on machine learning*, pages 2668–2677. PMLR, 2018.

[117] Mateo Espinosa Zarlenga, Pietro Barbiero, Gabriele Ciravegna, Giuseppe Marra, Francesco Giannini, Michelangelo Diligenti, Zohreh Shams, Frederic Precioso, Stefano Melacci, Adrian Weller, et al. Concept embedding models. *arXiv preprint arXiv:2209.09056*, 2022.

[118] Reduan Achtibat, Maximilian Dreyer, Ilona Eisenbraun, Sebastian Bosse, Thomas Wiegand, Wojciech Samek, and Sebastian Lapuschkin. From attribution maps to human-understandable explanations through Concept Relevance Propagation. *Nature Machine Intelligence*, 5(9):1006–1019, September 2023. Number: 9, Publisher: Nature Publishing Group.

[119] Jiayuan Mao, Chuang Gan, Pushmeet Kohli, Joshua B. Tenenbaum, and Jiajun 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.

[120] Md Kamruzzaman Sarker, Lu Zhou, Aaron Eberhart, and Pascal Hitzler. Neuro-symbolic artificial intelligence. *AI Communications*, 34(3):197–209, 2021.

[121] Kyle Hamilton, Aparna Nayak, Bojan Božić, and Luca Longo. Is neuro-symbolic ai meeting its promises in natural language processing? a structured review. *Semantic Web*, 15(Preprint):1–42, 2022.

[122] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2901–2910, 2017.

[123] Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B. Tenenbaum. CLEVRER: collision events for video representation and reasoning. In *ICLR*, 2020.

[124] Heimo Müller and Andreas Holzinger. Kandinsky patterns. *Artificial intelligence*, 300:103546, 2021.

[125] Harm de Vries, Dzmitry Bahdanau, Shikhar Murty, Aaron C. Courville, and Philippe Beaudoin. CLOSURE: assessing systematic generalization of CLEVR models. In *Visually Grounded Interaction and Language (ViGIL), NeurIPS 2019 Workshop, Vancouver, Canada, December 13, 2019*, 2019.

[126] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In Yoshua Bengio and Yann LeCun, editors, *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Workshop Track Proceedings*, 2014.

[127] Marco Ribeiro, Sameer Singh, and Carlos Guestrin. "why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 97–101, San Diego, California, June 2016. Association for Computational Linguistics.

[128] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014.

[129] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, 2017.

[130] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7), 2015.

[131] Johannes Schneider and Giovanni Apruzzese. Concept-based adversarial attacks: Tricking humans and classifiers alike. In *2022 IEEE Security and Privacy Workshops (SPW)*, pages 66–72. IEEE, 2022.

[132] Anh Nguyen, Alexey Dosovitskiy, Jason Yosinski, Thomas Brox, and Jeff Clune. Synthesizing the preferred inputs for neurons in neural networks via deep generator networks. In *Advances in neural information processing systems*, pages 3387–3395, 2016.

[133] Tilman Räuker, Anson Ho, Stephen Casper, and Dylan 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)*, pages 464–483. IEEE, 2023.

[134] Johannes Schneider and Michalis Vlachos. A Survey of Deep Learning: From Activations to Transformers. *arXiv preprint arXiv:2302.00722*, 2023.

[135] Johannes Schneider and Michalis Vlachos. Explaining classifiers by constructing familiar concepts. *Machine Learning*, pages 1–34, 2022.

[136] Chih-Kuan Yeh, Cheng-Yu Hsieh, Arun Suggala, David I Inouye, and Pradeep K Ravikumar. On the (in) fidelity and sensitivity of explanations. *Advances in Neural Information Processing Systems*, 32, 2019.

[137] Leander Weber, Sebastian Lapuschkin, Alexander Binder, and Wojciech Samek. Beyond explaining: Opportunities and challenges of xai-based model improvement. *Information Fusion*, 92:154–176, 2023.

[138] Yuyang Gao, Siyi Gu, Junji Jiang, Sungsoo Ray Hong, Dazhou Yu, and Liang Zhao. Going beyond xai: A systematic survey for explanation-guided learning. *arXiv preprint arXiv:2212.03954*, 2022.

[139] Andrea Ferrario and Michele Loi. The robustness of counterfactual explanations over time. *IEEE Access*, 10:82736–82750, 2022.

[140] Luyu Qiu, Yi Yang, Caleb Chen Cao, Yueyuan Zheng, Hilary Ngai, Janet Hsiao, and Lei Chen. Generating perturbation-based explanations with robustness to out-of-distribution data. In *Proceedings of the ACM Web Conference 2022*, pages 3594–3605, 2022.

[141] Dominik Seuß. Bridging the gap between explainable ai and uncertainty quantification to enhance trustability. *arXiv preprint arXiv:2105.11828*, 2021.

[142] Aditya Kuppa and Nhien-An Le-Khac. Black box attacks on explainable artificial intelligence (xai) methods in cyber security. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2020.

[143] Abdullah Caglar Oksuz, Anisa Halimi, and Erman Ayday. Autolycus: Exploiting explainable ai (xai) for model extraction attacks against decision tree models. *arXiv preprint arXiv:2302.02162*, 2023.

[144] Frederik Pahde, Maximilian Dreyer, Wojciech Samek, and Sebastian Lapuschkin. Reveal to revise: An explainable ai life cycle for iterative bias correction of deep models. In Hayit Greenspan, Anant Madabhushi, Parvin Mousavi, Septimiu Salcudean, James Duncan, Tanveer Syeda-Mahmood, and Russell Taylor, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, pages 596–606, Cham, 2023. Springer Nature Switzerland.

[145] Francesco Bodria, Fosca Giannotti, Riccardo Guidotti, Francesca Naretto, Dino Pedreschi, and Salvatore Rinzivillo. Benchmarking and survey of explanation methods for black box models. *Data Mining and Knowledge Discovery*, 37:1719–1778, 2023.

[146] Riccardo Guidotti. Evaluating local explanation methods on ground truth. *Artif. Intell.*, 291:103428, 2021.

[147] Riccardo Guidotti. Counterfactual explanations and how to find them: literature review and benchmarking. *Data Mining and Knowledge Discovery*, pages 1–55, 2022.

[148] Giulia Vilone and Luca Longo. Notions of explainability and evaluation approaches for explainable artificial intelligence. *Information Fusion*, 76:89–106, 2021.

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

[150] Jonathan Dodge, Q. Vera Liao, Yunfeng Zhang, Rachel K. E. Bellamy, and Casey Dugan. Explaining models: an empirical study of how explanations impact fairness judgment. In *IUI*, pages 275–285. ACM, 2019.

[151] Ana Lucic, Hinda Haned, and Maarten de Rijke. Why does my model fail?: contrastive local explanations for retail forecasting. In *FAT**, pages 90–98. ACM, 2020.

[152] Carlo Metta, Riccardo Guidotti, Yuan Yin, Patrick Gallinari, and Salvatore Rinzivillo. Exemplars and counterexemplars explanations for skin lesion classifiers. In *HHAI*, volume 354 of *Frontiers in Artificial Intelligence and Applications*, pages 258–260. IOS Press, 2022.

[153] Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38, 2019.

[154] Robert R. Hoffman, Shane T. Mueller, Gary Klein, and Jordan Litman. Metrics for explainable AI: Challenges and prospects. *ArXiv*, abs/1812.04608, 2018.

[155] Chris van der Lee, Albert Gatt, Emiel van Miltenburg, and Emiel Krahmer. Human evaluation of automatically generated text: Current trends and best practice guidelines. *Computer Speech & Language*, 67:101151, 2021.

[156] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A survey of methods for explaining black box models. *ACM Comput. Surv.*, 51(5):93:1–93:42, 2019.

[157] Alejandro Barredo Arrieta, Natalia Díaz Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf. Fusion*, 58:82–115, 2020.

[158] Meike Nauta, Jan Trienes, Shreyasi Pathak, Elisa Nguyen, Michelle Peters, Yasmin Schmitt, Jörg Schlötterer, Maurice van Keulen, and Christin Seifert. From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai. *ACM Comput. Surv.*, feb 2023.

[159] Anna Hedström, Leander Weber, Dilyara Bareeva, Franz Motzkus, Wojciech Samek, Sebastian Lapuschkin, and Marina M. C. Höhne. Quantus: An explainable ai toolkit for responsible evaluation of neural network explanation. *Journal of Machine Learning Research*, 24(34):1–11, 2023.

[160] Leila Arras, Ahmed Osman, and Wojciech Samek. Clevr-xai: A benchmark dataset for the ground truth evaluation of neural network explanations. *Information Fusion*, 81:14–40, 2022.

[161] Frederik Pahde, Maximilian Dreyer, Wojciech Samek, and Sebastian Lapuschkin. Reveal to revise: An explainable ai life cycle for iterative bias correction of deep models. In Hayit Greenspan, Anant Madabhushi, Parvin Mousavi, Septimiu Salcudean, James Duncan, Tanveer Syeda-Mahmood, and Russell Taylor, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, pages 596–606, Cham, 2023. Springer Nature Switzerland.

[162] Luca 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*, pages 369–373. Springer, 2012.

[163] Luca Longo. Designing medical interactive systems via assessment of human mental workload. In *2015 IEEE 28th International Symposium on Computer-Based Medical Systems*, pages 364–365. IEEE, 2015.

[164] GM Hancock, L Longo, MS Young, and PA Hancock. Mental workload. *Handbook of human factors and ergonomics*, pages 203–226, 2021.

[165] Luca Longo, Christoper D. Wickens, Peter A. Hancock, and Gabriela M. Hancock. Human mental workload: A survey and a novel inclusive definition. *Frontiers in Psychology*, 13, 2022.

[166] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-Lopez, Daniel Molina, Benjamins Richard, Raja Chatila, and Francisco Herrera. Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58:82–115, 2020.

[167] Vijay Arya, Rachel K. E. Bellamy, Pin-Yu Chen, Amit Dhurandhar, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Q. Vera Liao, Ronny Luss, Aleksandra Mojsilović, Sami Mourad, Pablo Pedemonte, Ramya Raghavendra, John Richards, Prasanna Sattigeri, Karthikeyan Shanmugam, Moninder Singh, Kush R. Varshney, Dennis Wei, and Yunfeng Zhang. One explanation does not fit all: A toolkit and taxonomy of ai explainability techniques, 2021.

[168] Maya Krishnan. Against interpretability: A critical examination of the interpretability problem in machine learning. *Philosophy & Technology*, 33(3):487–502, 2020.

[169] Zachary C Lipton. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3):31–57, 2018.

[170] Upol Ehsan and Mark O Riedl. Social construction of xai: Do we need one definition to rule them all? In Michael Muller, Plamen Angelov, Hal Daume III, Shion Guha, Q. Vera Liao, Nuria Oliver, and David Piorkowski, editors, *Proceedings of the NeurIPS 2022 Workshop on Human-Centered AI*, 2022.

[171] Miruna-Adriana Clinciu and Helen Hastie. A survey of explainable ai terminology. In Jose M. Alonso and Alejandro Catala, editors, *Proceedings of the 1st Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence*, NL4XAI 2019, pages 8–13, Stroudsburg, PA, USA, 2019. Association for Computational Linguistics.

[172] Mara Graziani, Lidia Dutkiewicz, Davide Calvaresi, José Pereira Amorim, Katerina Yordanova, Mor Vered, Rahul Nair, Pedro Henriques Abreu, Tobias Blanke, Valeria Pulignano, John O. Prior, Lode Lauwaert, Wessel Reijers, Adrien Depeursinge, Vincent Andrearczyk, and Henning Müller. A global taxonomy of interpretable ai: Unifying the terminology for the technical and social sciences. *Artificial Intelligence Review*, pages 1–32, 2022.

[173] Tathagata Chakraborti, Anagha Kulkarni, Sarath Sreedharan, David E. Smith, and Subbarao Kambhampati. Explicability? legibility? predictability? transparency? privacy? security? the emerging landscape of interpretable agent behavior. In J. Benton, Nir Lipovetzky, Eva Onaindia, David E. Smith, and Siddharth Srivastava, editors, *Proceedings of the 29th International Conference on Automated Planning and Scheduling*, ICAPS 2019, pages 86–96, Palo Alto, CA, USA, 2019. AAAI Press.

[174] Sarah Sterz, Kevin Baum, Anne Lauber-Rönsberg, and Holger Hermanns. Towards perspicuity requirements. In Tao Yue and Mehdi Mirakhorli, editors, *29th IEEE International Requirements Engineering Conference Workshops*, REW 2021, pages 159–163, Piscataway, NJ, USA, 2021. IEEE.

[175] AI High-Level Expert Group. Ethics guidelines for trustworthy AI. *B-1049 Brussels*, 2019.

[176] Lena Kästner, Markus Langer, Veronika Lazar, Astrid Schomäcker, Timo Speith, and Sarah Sterz. On the relation of trust and explainability: Why to engineer for trustworthiness. In Tao Yue and Mehdi Mirakhorli, editors, *29th IEEE International Requirements Engineering Conference Workshops*, REW 2021, pages 169–175, Piscataway, NJ, USA, 2021. IEEE.

[177] Scott Robbins. A misdirected principle with a catch: Explicability for ai. *Minds and Machines*, 29(4):495–514, 2019.

[178] René F. Kizilcec. How much information? effects of transparency on trust in an algorithmic interface. In Jofish Kaye, Allison Druin, Cliff Lampe, Dan Morris, and Juan P. Hourcade, editors, *Proceedings of the 34th Conference on Human Factors in Computing Systems*, CHI 2016, pages 2390–2395, New York, NY, USA, 2016. Association for Computing Machinery.

[179] Bishwamittra Ghosh, Dmitry Malioutov, and Kuldeep S. Meel. Interpretable classification rules in relaxed logical form. In Tim Miller, Rosina Weber, and Daniele Magazzeni, editors, *Proceedings of the IJCAI 2019 Workshop on Explainable Artificial Intelligence*, IJCAI XAI 2019, pages 14–20, 2019.

[180] Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why should I trust you?": Explaining the predictions of any classifier. In Charu Aggarwal, Balaji Krishnapuram, Rajeev Rastogi, Dou Shen, Mohak Shah, and Alex Smola, editors, *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD 2016, pages 1135–1144, New York, NY, USA, 2016. Association for Computing Machinery.

[181] Natalia Díaz-Rodríguez, Javier Del Ser, Mark Coeckelbergh, Marcos López de Prado, Enrique Herrera-Viedma, and Francisco Herrera. Connecting the dots in trustworthy Artificial Intelligence: From AI principles, ethics, and key requirements to responsible AI systems and regulation. *Inf. Fusion*, 99:101896, 2023.

[182] Nicola Palladino. A 'biased' emerging governance regime for artificial intelligence? how ai ethics get skewed moving from principles to practices. *Telecommunications Policy*, 47(5):102479, 2023.

[183] Kareem Khalifa. Inaugurating understanding or repackaging explanation? *Philosophy of Science*, 79(1):15–37, 2012.

[184] Michael Strevens. No understanding without explanation. *Studies in history and philosophy of science Part A*, 44(3):510–515, 2013.

[185] Peter Lipton. Understanding without explanation. *Scientific understanding: Philosophical perspectives*, pages 43–63, 2009.

[186] Catherine Z Elgin. *True enough*. MIT press, 2017.

[187] Jonathan Kvanvig. Responses to critics. *Epistemic value*, pages 339–351, 2009.

[188] Moti Mizrahi. Idealizations and scientific understanding. *Philosophical Studies*, 160:237–252, 2012.

[189] J Adam Carter and Emma C Gordon. Objectual understanding, factivity and belief. *Epistemic reasons, norms and goals*, 423, 2016.

[190] Adrian Erasmus, Tyler DP Brunet, and Eyal Fisher. What is interpretability? *Philosophy & Technology*, 34(4):833–862, 2021.

[191] Duncan Pritchard. *Knowing the answer, understanding and epistemic value*. Citeseer, 2008.

[192] Linda Zagzebski. *On Epistemology*. Wadsworth, 2009.

[193] Tania Lombrozo, Daniel Wilkenfeld, T Lombrozo, and D Wilkenfeld. Mechanistic versus functional understanding. *Varieties of understanding: New perspectives from philosophy, psychology, and theology*, pages 209–229, 2019.

[194] Emily Sullivan. Understanding from machine learning models. *The British Journal for the Philosophy of Science*, 2022.

[195] Kathleen A Creel. Transparency in complex computational systems. *Philosophy of Science*, 87(4):568–589, 2020.

[196] Juan M Durán. Dissecting scientific explanation in ai (sxai): A case for medicine and healthcare. *Artificial Intelligence*, 297:103498, 2021.

[197] Carlos Zednik. Solving the black box problem: A normative framework for explainable artificial intelligence. *Philosophy & technology*, 34(2):265–288, 2021.

[198] Will Fleisher. Understanding, idealization, and explainable ai. *Episteme*, 19(4):534–560, 2022.

[199] Paulo Pirozelli. Sources of understanding in supervised machine learning models. *Philosophy & Technology*, 35(2):23, 2022.

[200] Maartje MA De Graaf and Bertram F Malle. How people explain action (and autonomous intelligent systems should too). In *2017 AAAI Fall Symposium Series*, pages 19–26, 2017.

[201] Brent Mittelstadt, Chris Russell, and Sandra Wachter. Explaining explanations in ai. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 279–288, 2019.

[202] Roberto Confalonieri, Tillman Weyde, Tarek R. Besold, and Fermín Moscoso del Prado Martín. Using ontologies to enhance human understandability of global post-hoc explanations of black-box models. *Artificial Intelligence*, 296, 2021.

[203] Ian E Nielsen, Dimah Dera, Ghulam Rasool, Ravi P Ramachandran, and Nidhal Carla Bouaynaya. Robust explainability: A tutorial on gradient-based attribution methods for deep neural networks. *IEEE Signal Processing Magazine*, 39(4):73–84, 2022.

[204] Jinyang Yuan, Tonglin Chen, Bin Li, and Xiangyang Xue. Compositional scene representation learning via reconstruction: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(10):11540–11560, 2023.

[205] Tim Klinger, Dhaval Adjodah, Vincent Marois, Josh Joseph, Matthew Riemer, Alex'Sandy' Pentland, and Murray Campbell. A study of compositional generalization in neural models. *arXiv preprint arXiv:2006.09437*, 2020.

[206] Lucas Rizzo and Luca Longo. An empirical evaluation of the inferential capacity of defeasible argumentation, non-monotonic fuzzy reasoning and expert systems. *Expert Systems with Applications*, 147:113220, 2020.

[207] Lucas Rizzo and Luca 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.*, pages 138–149, 2018.

[208] Giulia Vilone and Luca 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*, pages 447–460. Springer, 2022.

[209] Giulia Vilone and Luca 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)*, volume 3209 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2022.

[210] Johanna Vielhaben, Sebastian Lapuschkin, Grégoire Montavon, and Wojciech Samek. Explainable ai for time series via virtual inspection layers. *arXiv preprint arXiv:2303.06365*, 2023.

[211] Taufique Ahmed and Luca 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.

[212] W. V. Quine. On what there is. In W. V. Quine, editor, *From a Logical Point of View*, pages 1–19. Cambridge, Mass.: Harvard University Press, 1953.

[213] David Krakauer. The Computational Systems of the World. *BioScience*, 64(4):351–354, April 2014.

[214] Samy Badreddine, Artur d'Avila Garcez, Luciano Serafini, and Michael Spranger. Logic tensor networks. *Artificial Intelligence*, 303:103649, 2022.

[215] Alexandre Heuillet, Fabien Couthouis, and Natalia Díaz-Rodríguez. Explainability in deep reinforcement learning. *Knowledge-Based Systems*, 214:106685, 2021.

[216] C Meske, E Bunde, J Schneider, and M Gersch. Explainable artificial intelligence: Objectives, stakeholders and future research opportunities. *Information Systems Management*, 2020.

[217] Adrian Weller. Transparency: Motivations and Challenges. In Wojciech Samek, Grégoire Montavon, Andrea Vedaldi, Lars Kai Hansen, and Klaus-Robert Müller, editors, *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, Lecture Notes in Computer Science, pages 23–40. Springer International Publishing, Cham, 2019.

[218] Ronan Hamon, Henrik Junklewitz, Gianclaudio Malgieri, Paul De Hert, Laurent Beslay, and Ignacio 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*, pages 549–559, 2021.

[219] Jessica C. Flack. Multiple time-scales and the developmental dynamics of social systems. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367(1597):1802–1810, July 2012. Publisher: Royal Society.

[220] Mislav Juric, Agneza Sandic, and Mario Brcic. AI safety: state of the field through quantitative lens. In *2020 43rd International Convention on Information, Communication and Electronic Technology (MIPRO)*, pages 1254–1259, September 2020. ISSN: 2623-8764.

[221] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning, 2017.

[222] Sander Beckers. Causal explanations and xai. In *Conference on Causal Learning and Reasoning*, pages 90–109. PMLR, 2022.

[223] Yu-Liang Chou, Catarina Moreira, Peter Bruza, Chun Ouyang, and Joaquim Jorge. Counterfactuals and causability in explainable artificial intelligence: Theory, algorithms, and applications. *Information Fusion*, 81:59–83, 2022.

[224] Martina Cinquini and Riccardo Guidotti. Calime: Causality-aware local interpretable model-agnostic explanations. *arXiv preprint arXiv:2212.05256*, 2022.

[225] Javier Del Ser, Alejandro Barredo-Arrieta, Natalia Díaz-Rodríguez, Francisco Herrera, and Andreas Holzinger. Exploring the trade-off between plausibility, change intensity and adversarial power in counterfactual explanations using multi-objective optimization. *arXiv preprint arXiv:2205.10232*, 2022.

[226] Pedro Sanchez and Sotirios A Tsaftaris. Diffusion causal models for counterfactual estimation. *Conference on Causal Learning and Reasoning (CLeaR)*, 2022.

[227] Maximilian Augustin, Valentyn Boreiko, Francesco Croce, and Matthias Hein. Diffusion visual counterfactual explanations. In *NeurIPS*, 2022.

[228] Johannes Schneider and Michalis Vlachos. Personalization of deep learning. In *Data Science–Analytics and Applications: Proceedings of the 3rd International Data Science Conference–iDSC2020*, pages 89–96. Springer, 2021.

[229] Johannes Schneider and Joshua Peter Handali. Personalized explanation for machine learning: a conceptualization. In *European Conference on Information Systems (ECIS)*, 2019.

[230] Banghua Zhu, Michael Jordan, and Jiantao Jiao. Principled reinforcement learning with human feedback from pairwise or k-wise comparisons. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 43037–43067. PMLR, 23–29 Jul 2023.

[231] Tom Bewley and Freddy 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, page 118–126. International Foundation for Autonomous Agents and Multiagent Systems, 2022.

[232] Andrea Bunt, Matthew Lount, and Catherine 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, pages 169–178, New York, NY, USA, 2012. ACM.

[233] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, May 2019. Number: 5 Publisher: Nature Publishing Group.

[234] Ronan Hamon, Henrik Junklewitz, Ignacio Sanchez, Gianclaudio Malgieri, and Paul De Hert. Bridging the gap between ai and explainability in the gdpr: Towards trustworthiness-by-design in automated decision-makin. *IEEE Computational Intelligence Magazine*, 17(1):72–85, 2022.

[235] David Gunning, Mark Stefik, Jaesik Choi, Timothy Miller, Simone Stumpf, and Guang-Zhong Yang. XAI—Explainable artificial intelligence. *Science Robotics*, 4(37), December 2019.

[236] Agneza Krajna, Mario Brcic, Tomislav Lipic, and Juraj Doncevic. Explainability in reinforcement learning: perspective and position, March 2022.

[237] Marzyeh Ghassemi, Luke Oakden-Rayner, and Andrew L Beam. The false hope of current approaches to explainable artificial intelligence in health care. *The Lancet Digital Health*, 3(11):e745–e750, 2021.

[238] Federico Cabitza, Andrea Campagner, Lorenzo Famiglini, Enrico Gallazzi, and Giovanni Andrea 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*, pages 31–50. Springer, 2022.

[239] Federico Cabitza, Andrea Campagner, Luca Ronzio, Matteo Cameli, Giulia Elena Mandoli, Maria Concetta Pastore, Luca Sconfienza, Duarte Folgado, Marília Barandas, and Hugo Gamboa. Rams, hounds and white boxes: Investigating human-ai collaboration protocols in medical diagnosis. *Artificial Intelligence in Medicine*, page 102506, 2023.

[240] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel 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*, pages 1–16, 2021.

[241] Simone Natale. *Deceitful media: Artificial intelligence and social life after the Turing test*. Oxford University Press, USA, 2021.

[242] Federico Cabitza, Andrea Campagner, and Edoardo 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*, pages 36–49. Springer, 2021.

[243] Federico Cabitza, Andrea Campagner, and Carla Simone. The need to move away from agential-ai: Empirical investigations, useful concepts and open issues. *International Journal of Human-Computer Studies*, 155:102696, 2021.

[244] Tim 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*, pages 333–342, 2023.

[245] Ashraf Abdul, Jo Vermeulen, Danding Wang, Brian Y. Lim, and Mohan 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, pages 582:1–582:18, New York, NY, USA, 2018. ACM.

[246] Kevin Baum, Susanne Mantel, Eva Schmidt, and Timo Speith. From responsibility to reason-giving explainable artificial intelligence. *Philosophy & Technology*, 35(1):1–30, 2022.

[247] Stephen Thornton. Karl Popper. In Edward N. Zalta and Uri Nodelman, editors, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Winter 2023 edition, 2023.

[248] Matthew L. Leavitt and Ari Morcos. Towards falsifiable interpretability research. *NeurIPS 2020 Workshop: ML Retrospectives, Surveys and Meta-Analyses (ML-RSA)*, 2020.

[249] Filip Karlo Dosilovic, Mario Brcic, and Nikica Hlupic. Explainable artificial intelligence: A survey. In *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, pages 0210–0215, May 2018.

[250] Agneza Krajna, Mario Brcic, Mihael Kovac, and Ana Sarcevic. Explainable Artificial Intelligence: An Updated Perspective. In *Proceedings of 45th Jubilee International Convention on Information, Communication and Electronic Technology (MIPRO) 2022*, pages 859–864, Opatija, Croatia, May 2022.

[251] W. James Murdoch, Chandan Singh, Karl Kumbier, Reza Abbasi-Asl, and Bin Yu. Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences*, 116(44):22071–22080, October 2019. Publisher: Proceedings of the National Academy of Sciences.

[252] Johannes Schneider, Christian Meske, and Michalis Vlachos. Deceptive AI explanations: Creation and detection. In *International Conference on Agents and Artificial Intelligence(ICAART)*, pages 44–55, 2022.

[253] Johannes Schneider and Frank Breitinger. AI Forensics: Did the artificial intelligence system do it? why? *arXiv preprint arXiv:2005.13635*, 2020.

[254] Johannes Schneider and Michalis Vlachos. Reflective-net: Learning from explanations. *Data Mining and Knowledge Discovery*, 2023.

[255] Roman V. Yampolskiy. Unexplainability and Incomprehensibility of AI. *Journal of Artificial Intelligence and Consciousness*, 07(02):277–291, September 2020. Publisher: World Scientific Publishing Co.

[256] Roman V. Yampolskiy. What are the ultimate limits to computational techniques: verifier theory and unverifiability. *Physica Scripta*, 92(9):093001, July 2017. Publisher: IOP Publishing.

[257] Victor Boutin, Thomas Fel, Lakshya Singhal, Rishav Mukherji, Akash Nagaraj, Julien Colin, and Thomas Serre. Diffusion models as artists: Are we closing the gap between humans and machines? In *International Conference on Machine Learning*, 2023.

[258] H Holden Thorp. Chatgpt is fun, but not an author, 2023.

[259] Eva AM van Dis, Johan Bollen, Willem Zuidema, Robert van Rooij, and Claudi L Bockting. Chatgpt: five priorities for research. *Nature*, 614(7947):224–226, 2023.

[260] Franziska Boenisch. A systematic review on model watermarking for neural networks. *Frontiers in big Data*, 4:729663, 2021.

[261] John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. A watermark for large language models. In *Proceedings of the 40th International Conference on Machine Learning*, pages 202:17061–17084, 2023.

[262] Lucas Bourtoule, Varun Chandrasekaran, Christopher A Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. Machine unlearning. In *2021 IEEE Symposium on Security and Privacy (SP)*, pages 141–159. IEEE, 2021.

[263] Thanh Tam Nguyen, Thanh Trung Huynh, Phi Le Nguyen, Alan Wee-Chung Liew, Hongzhi Yin, and Quoc Viet Hung Nguyen. A survey of machine unlearning. *arXiv preprint arXiv:2209.02299*, 2022.

[264] Julie E Cohen. *Between truth and power*. Oxford University Press, 2019.

[265] Federico Cabitza, Andrea Campagner, Gianclaudio Malgieri, Chiara Natali, David Schneeberger, Karl Stoeger, and Andreas Holzinger. Quod erat demonstrandum?-towards a typology of the concept of explanation for the design of explainable ai. *Expert Systems with Applications*, 213:118888, 2023.

[266] Gianclaudio Malgieri. "just" algorithms: justification (beyond explanation) of automated decisions under the general data protection regulation. *Law and Business*, 1(1):16–28, 2021.

[267] Emre Bayamlioglu. Contesting automated decisions. *Eur. Data Prot. L. Rev.*, 4:433, 2018.

[268] Clément Henin and Daniel Le Métayer. Beyond explainability: justifiability and contestability of algorithmic decision systems. *AI & SOCIETY*, pages 1–14, 2021.

[269] Clément Henin and Daniel Le Métayer. A framework to contest and justify algorithmic decisions. *AI and Ethics*, 1(4):463–476, 2021.

[270] Lisa M Austin. Enough about me: why privacy is about power, not consent (or harm). *Forthcoming in Austin Sarat, ed., A World Without Privacy*, 2014.

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

[272] Sasha Costanza-Chock. *Design justice: Community-led practices to build the worlds we need*. The MIT Press, 2020.

[273] Margot E Kaminski and Gianclaudio Malgieri. Algorithmic impact assessments under the gdpr: producing multi-layered explanations. *International data privacy law*, pages 19–28, 2020.

[274] Judith Gregory. Scandinavian approaches to participatory design. *International Journal of Engineering Education*, 19(1):62–74, 2003.

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

[276] Gianclaudio Malgieri. In/acceptable marketing and consumers' privacy expectations: Four tests from eu data protection law. *Journal of Consumer Marketing*, 40(2):209–223, 2023.

[277] Keld Bodker, Finn Kensing, and Jesper Simonsen. *Participatory IT design: designing for business and workplace realities*. MIT press, 2009.