Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI

Alejandro Barredo Arrieta[a], Natalia Díaz-Rodríguez[b], Javier Del Ser[a,c,d], Adrien Bennetot[b,e,f], Siham Tabik[g], Alberto Barbado[h], Salvador Garcia[g], Sergio Gil-Lopez[a], Daniel Molina[g], Richard Benjamins[h], Raja Chatila[f], and Francisco Herrera[g]

[a]*TECNALIA, 48160 Derio, Spain*
[b]*ENSTA, Institute Polytechnique Paris and INRIA Flowers Team, Palaiseau, France*
[c]*University of the Basque Country (UPV/EHU), 48013 Bilbao, Spain*
[d]*Basque Center for Applied Mathematics (BCAM), 48009 Bilbao, Bizkaia, Spain*
[e]*Segula Technologies, Parc d'activité de Pissaloup, Trappes, France*
[f]*Institut des Systèmes Intelligents et de Robotique, Sorbonne Universitè, France*
[g]*DaSCI Andalusian Institute of Data Science and Computational Intelligence, University of Granada, 18071 Granada, Spain*
[h]*Telefonica, 28050 Madrid, Spain*

Abstract

In the last few years, Artificial Intelligence (AI) has achieved a notable momentum that, if harnessed appropriately, may deliver the best of expectations over many application sectors across the field. For this to occur shortly in Machine Learning, the entire community stands in front of the barrier of explainability, an inherent problem of the latest techniques brought by sub-symbolism (e.g. ensembles or Deep Neural Networks) that were not present in the last hype of AI (namely, expert systems and rule based models). Paradigms underlying this problem fall within the so-called *eXplainable* AI (XAI) field, which is widely acknowledged as a crucial feature for the practical deployment of AI models. The overview presented in this article examines the existing literature and contributions already done in the field of XAI, including a prospect toward what is yet to be reached. For this purpose we summarize previous efforts made to define explainability in Machine Learning, establishing a novel definition of explainable Machine Learning that covers such prior conceptual propositions with a major focus on the audience for which the explainability is sought. Departing from this definition, we propose and discuss about a taxonomy of recent contributions related to the explainability of different Machine Learning models, including those aimed at explaining Deep Learning methods for which a second dedicated taxonomy is built and examined in detail. This critical literature analysis serves as the motivating background for a series of challenges faced by XAI, such as the interesting crossroads of data fusion and explainability. Our prospects lead toward the concept of *Responsible Artificial Intelligence*, namely, a methodology for the large-scale implementation of AI methods in real organizations with fairness, model explainability and accountability at its core. Our ultimate goal is to provide newcomers to the field of XAI with a thorough taxonomy that can serve as reference material in order to stimulate future research advances, but also to encourage experts and professionals from other disciplines to embrace the benefits of AI in their activity sectors, without any prior bias for its lack of interpretability.

Keywords: Explainable Artificial Intelligence, Machine Learning, Deep Learning, Data Fusion, Interpretability, Comprehensibility, Transparency, Privacy, Fairness, Accountability, Responsible Artificial Intelligence.

*Corresponding author. TECNALIA. P. Tecnologico, Ed. 700. 48170 Derio (Bizkaia), Spain. E-mail: javier.delser@tecnalia.com

1. Introduction

Artificial Intelligence (AI) lies at the core of many activity sectors that have embraced new information technologies [1]. While the roots of AI trace back to several decades ago, there is a clear consensus on the paramount importance featured nowadays by intelligent machines endowed with learning, reasoning and adaptation capabilities. It is by virtue of these capabilities that AI methods are achieving unprecedented levels of performance when learning to solve increasingly complex computational tasks, making them pivotal for the future development of the human society [2]. The sophistication of AI-powered systems has lately increased to such an extent that almost no human intervention is required for their design and deployment. When decisions derived from such systems ultimately affect humans' lives (as in e.g. medicine, law or defense), there is an emerging need for understanding how such decisions are furnished by AI methods [3].

While the very first AI systems were easily interpretable, the last years have witnessed the rise of opaque decision systems such as Deep Neural Networks (DNNs). The empirical success of Deep Learning (DL) models such as DNNs stems from a combination of efficient learning algorithms and their huge parametric space. The latter space comprises hundreds of layers and millions of parameters, which makes DNNs be considered as complex *black-box* models [4]. The opposite of *black-box-ness* is *transparency*, i.e., the search for a direct understanding of the mechanism by which a model works [5].

As black-box Machine Learning (ML) models are increasingly being employed to make important predictions in critical contexts, the demand for transparency is increasing from the various stakeholders in AI [6]. The danger is on creating and using decisions that are not justifiable, legitimate, or that simply do not allow obtaining detailed explanations of their behaviour [7]. Explanations supporting the output of a model are crucial, e.g., in precision medicine, where experts require far more information from the model than a simple binary prediction for supporting their diagnosis [8]. Other examples include autonomous vehicles in transportation, security, and finance, among others.

In general, humans are reticent to adopt techniques that are not directly interpretable, tractable and trustworthy [9], given the increasing demand for ethical AI [3]. It is customary to think that by focusing solely on performance, the systems will be increasingly opaque. This is true in the sense that there is a trade-off between the performance of a model and its transparency [10]. However, an improvement in the understanding of a system can lead to the correction of its deficiencies. When developing a ML model, the consideration of interpretability as an additional design driver can improve its implementability for 3 reasons:

- Interpretability helps ensure impartiality in decision-making, i.e. to detect, and consequently, correct from bias in the training dataset.
- Interpretability facilitates the provision of robustness by highlighting potential adversarial perturbations that could change the prediction.
- Interpretability can act as an insurance that only meaningful variables infer the output, i.e., guaranteeing that an underlying truthful causality exists in the model reasoning.

All these means that the interpretation of the system should, in order to be considered practical, provide either an understanding of the model mechanisms and predictions, a visualization of the model's discrimination rules, or hints on what could perturb the model [11].

In order to avoid limiting the effectiveness of the current generation of AI systems, *eXplainable AI* (XAI) [7] proposes creating a suite of ML techniques that 1) produce more explainable models while maintaining a high level of learning performance (e.g., prediction accuracy), and 2) enable humans to understand, appropriately trust, and effectively manage the emerging generation of artificially intelligent partners. XAI draws as well insights from the Social Sciences [12] and considers the psychology of explanation.

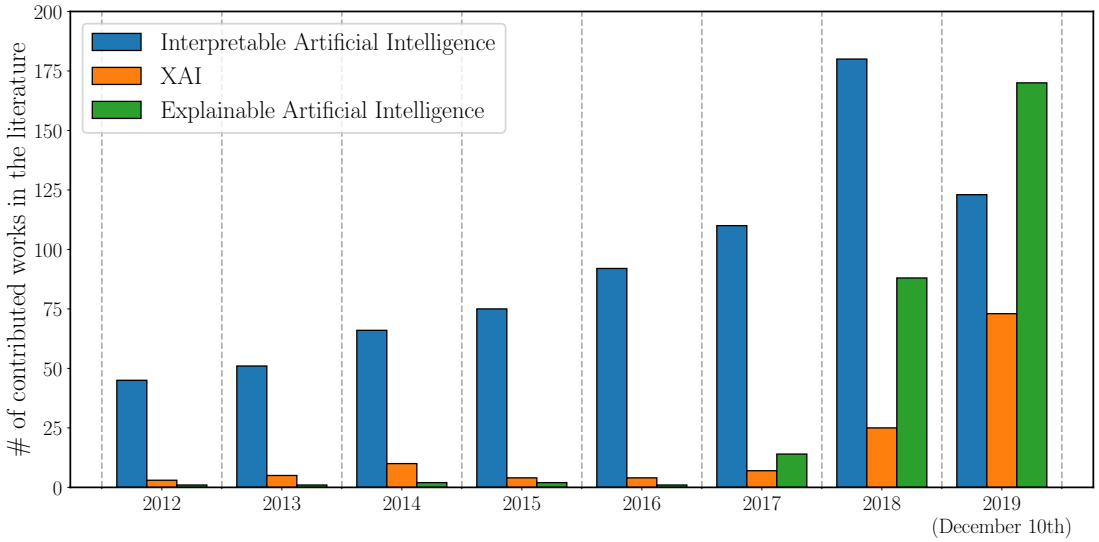


Figure 1: Evolution of the number of total publications whose title, abstract and/or keywords refer to the field of XAI during the last years. Data retrieved from Scopus® (December 10th, 2019) by using the search terms indicated in the legend when querying this database. It is interesting to note the latent need for interpretable AI models over time (which conforms to intuition, as interpretability is a requirement in many scenarios), yet it has not been until 2017 when the interest in techniques to explain AI models has permeated throughout the research community.

Figure 1 displays the rising trend of contributions on XAI and related concepts. This literature outbreak shares its rationale with the research agendas of national governments and agencies. Although some recent surveys [8, 13, 10, 14, 15, 16, 17] summarize the upsurge of activity in XAI across sectors and disciplines, this overview aims to cover the creation of a complete unified framework of categories and concepts that allow for scrutiny and understanding of the field of XAI methods. Furthermore, we pose intriguing thoughts around the explainability of AI models in data fusion contexts with regards to data privacy and model confidentiality. This, along with other research opportunities and challenges identified throughout our study, serve as the pull factor toward Responsible Artificial Intelligence, term by which we refer to a series of AI principles to be necessarily met when deploying AI in real applications. As we will later show in detail, model explainability is among the most crucial aspects to be ensured within this methodological framework. All in all, the novel contributions of this overview can be summarized as follows:

1. Grounded on a first elaboration of concepts and terms used in XAI-related research, we propose a novel definition of explainability that places *audience* (Figure 2) as a key aspect to be considered when explaining a ML model. We also elaborate on the diverse purposes sought when using XAI techniques, from trustworthiness to privacy awareness, which round up the claimed importance of purpose and targeted audience in model explainability.

2. We define and examine the different levels of transparency that a ML model can feature by itself, as well as the diverse approaches to post-hoc explainability, namely, the explanation of ML models that are not transparent by design.

3. We thoroughly analyze the literature on XAI and related concepts published to date, covering approximately 400 contributions arranged into two different taxonomies. The first taxonomy addresses the explainability of ML models using the previously made distinction between transparency and post-hoc explainability, including models that are transparent by themselves, Deep and non-Deep (i.e.,

shallow) learning models. The second taxonomy deals with XAI methods suited for the explanation of Deep Learning models, using classification criteria closely linked to this family of ML methods (e.g. layerwise explanations, representation vectors, attention).

4. We enumerate a series of challenges of XAI that still remain insufficiently addressed to date. Specifically, we identify research needs around the concepts and metrics to evaluate the explainability of ML models, and outline research directions toward making Deep Learning models more understandable. We further augment the scope of our prospects toward the implications of XAI techniques in regards to confidentiality, robustness in adversarial settings, data diversity, and other areas intersecting with explainability.

5. After the previous prospective discussion, we arrive at the concept of Responsible Artificial Intelligence, a manifold concept that imposes the systematic adoption of several AI principles for AI models to be of practical use. In addition to explainability, the guidelines behind Responsible AI establish that fairness, accountability and privacy should also be considered when implementing AI models in real environments.

6. Since Responsible AI blends together model explainability and privacy/security by design, we call for a profound reflection around the benefits and risks of XAI techniques in scenarios dealing with sensitive information and/or confidential ML models. As we will later show, the regulatory push toward data privacy, quality, integrity and governance demands more efforts to assess the role of XAI in this arena. In this regard, we provide an insight on the implications of XAI in terms of privacy and security under different data fusion paradigms.

The remainder of this overview is structured as follows: first, Section 2 and subsections therein open a discussion on the terminology and concepts revolving around explainability and interpretability in AI, ending up with the aforementioned novel definition of interpretability (Subsections 2.1 and 2.2), and a general criterion to categorize and analyze ML models from the XAI perspective. Sections 3 and 4 proceed by reviewing recent findings on XAI for ML models (on transparent models and post-hoc techniques respectively) that comprise the main division in the aforementioned taxonomy. We also include a review on hybrid approaches among the two, to attain XAI. Benefits and caveats of the synergies among the families of methods are discussed in Section 5, where we present a prospect of general challenges and some consequences to be cautious about. Finally, Section 6 elaborates on the concept of Responsible Artificial Intelligence. Section 7 concludes the survey with an outlook aimed at engaging the community around this vibrant research area, which has the potential to impact society, in particular those sectors that have progressively embraced ML as a core technology of their activity.

2. Explainability: What, Why, What For and How?

Before proceeding with our literature study, it is convenient to first establish a common point of understanding on what the term *explainability* stands for in the context of AI and, more specifically, ML. This is indeed the purpose of this section, namely, to pause at the numerous definitions that have been done in regards to this concept (what?), to argue why explainability is an important issue in AI and ML (why? what for?) and to introduce the general classification of XAI approaches that will drive the literature study thereafter (how?).

2.1. Terminology Clarification

One of the issues that hinders the establishment of common grounds is the interchangeable misuse of interpretability and explainability in the literature. There are notable differences among these concepts. To begin with, interpretability refers to a passive characteristic of a model referring to the level at which a given model makes sense for a human observer. This feature is also expressed as transparency. By

contrast, explainability can be viewed as an active characteristic of a model, denoting any action or procedure taken by a model with the intent of clarifying or detailing its internal functions.

To summarize the most commonly used nomenclature, in this section we clarify the distinction and similarities among terms often used in the ethical AI and XAI communities.

- **Understandability** (or equivalently, **intelligibility**) denotes the characteristic of a model to make a human understand its function – how the model works – without any need for explaining its internal structure or the algorithmic means by which the model processes data internally [18].
- **Comprehensibility**: when conceived for ML models, comprehensibility refers to the ability of a learning algorithm to represent its learned knowledge in a human understandable fashion [19, 20, 21]. This notion of model comprehensibility stems from the postulates of Michalski [22], which stated that *"the results of computer induction should be symbolic descriptions of given entities, semantically and structurally similar to those a human expert might produce observing the same entities. Components of these descriptions should be comprehensible as single 'chunks' of information, directly interpretable in natural language, and should relate quantitative and qualitative concepts in an integrated fashion"*. Given its difficult quantification, comprehensibility is normally tied to the evaluation of the model complexity [17].
- **Interpretability**: it is defined as the ability to explain or to provide the meaning in understandable terms to a human.
- **Explainability**: explainability is associated with the notion of explanation as an interface between humans and a decision maker that is, at the same time, both an accurate proxy of the decision maker and comprehensible to humans [17].
- **Transparency**: a model is considered to be transparent if by itself it is understandable. Since a model can feature different degrees of understandability, transparent models in Section 3 are divided into three categories: simulatable models, decomposable models and algorithmically transparent models [5].

In all the above definitions, *understandability* emerges as the most essential concept in XAI. Both transparency and interpretability are strongly tied to this concept: while transparency refers to the characteristic of a model to be, on its own, understandable for a human, understandability measures the degree to which a human can understand a decision made by a model. Comprehensibility is also connected to understandability in that it relies on the capability of the audience to understand the knowledge contained in the model. All in all, understandability is a two-sided matter: model understandability and human understandability. This is the reason why the definition of XAI given in Section 2.2 refers to the concept of *audience*, as the cognitive skills and pursued goal of the users of the model have to be taken into account jointly with the intelligibility and comprehensibility of the model in use. This prominent role taken by understandability makes the concept of *audience* the cornerstone of XAI, as we next elaborate in further detail.

2.2. What?

Although it might be considered to be beyond the scope of this paper, it is worth noting the discussion held around general theories of explanation in the realm of philosophy [23]. Many proposals have been done in this regard, suggesting the need for a general, unified theory that approximates the structure and intent of an explanation. However, nobody has stood the critique when presenting such a general theory. For the time being, the most agreed-upon thought blends together different approaches to explanation drawn from diverse knowledge disciplines. A similar problem is found when addressing interpretability in AI. It appears from the literature that there is not yet a common point of understanding on what interpretability or explainability are. However, many contributions claim the achievement of interpretable models and techniques that empower explainability.

To shed some light on this lack of consensus, it might be interesting to place the reference starting point at the definition of the term Explainable Artificial Intelligence (XAI) given by D. Gunning in [7]:

> *"XAI will create a suite of machine learning techniques that enables human users to understand, appropriately trust, and effectively manage the emerging generation of artificially intelligent partners"*

This definition brings together two concepts (understanding and trust) that need to be addressed in advance. However, it misses to consider other purposes motivating the need for interpretable AI models, such as causality, transferability, informativeness, fairness and confidence [5, 24, 25, 26]. We will later delve into these topics, mentioning them here as a supporting example of the incompleteness of the above definition.

As exemplified by the definition above, a thorough, complete definition of explainability in AI still slips from our fingers. A broader reformulation of this definition (e.g. *"An explainable Artificial Intelligence is one that produces explanations about its functioning"*) would fail to fully characterize the term in question, leaving aside important aspects such as its purpose. To build upon the completeness, a definition of explanation is first required.

As extracted from the Cambridge Dictionary of English Language, an explanation is *"the details or reasons that someone gives to make something clear or easy to understand"* [27]. In the context of an ML model, this can be rephrased as: *"the details or reasons a model gives to make its functioning clear or easy to understand"*. It is at this point where opinions start to diverge. Inherently stemming from the previous definitions, two ambiguities can be pointed out. First, the details or the reasons used to explain, are completely dependent of the audience to which they are presented. Second, whether the explanation has left the concept clear or easy to understand also depends completely on the audience. Therefore, the definition must be rephrased to reflect explicitly the dependence of the explainability of the model on the audience. To this end, a reworked definition could read as:

> *Given a certain audience, explainability refers to the details and reasons a model gives to make its functioning clear or easy to understand.*

Since explaining, as argumenting, may involve weighting, comparing or convincing an audience with logic-based formalizations of (counter) arguments [28], explainability might convey us into the realm of cognitive psychology and the *psychology of explanations* [7], since measuring whether something has been understood or put clearly is a hard task to be gauged objectively. However, measuring to which extent the internals of a model can be explained could be tackled objectively. Any means to reduce the complexity of the model or to simplify its outputs should be considered as an XAI approach. How big this leap is in terms of complexity or simplicity will correspond to how explainable the resulting model is. An underlying problem that remains unsolved is that the interpretability gain provided by such XAI approaches may not be straightforward to quantify: for instance, a model simplification can be evaluated based on the reduction of the number of architectural elements or number of parameters of the model itself (as often made, for instance, for DNNs). On the contrary, the use of visualization methods or natural language for the same purpose does not favor a clear quantification of the improvements gained in terms of interpretability. The derivation of general metrics to assess the quality of XAI approaches remain as an open challenge that should be under the spotlight of the field in forthcoming years. We will further discuss on this research direction in Section 5.

Explainability is linked to post-hoc explainability since it covers the techniques used to convert a non-interpretable model into a explainable one. In the remaining of this manuscript, explainability will be considered as the main design objective, since it represents a broader concept. A model can be explained, but the interpretability of the model is something that comes from the design of the model itself. Bearing these observations in mind, explainable AI can be defined as follows:

> *Given an audience, an **explainable** Artificial Intelligence is one that produces details or reasons to make its functioning clear or easy to understand.*

This definition is posed here as a first contribution of the present overview, implicitly assumes that the ease of understanding and clarity targeted by XAI techniques for the model at hand reverts on different application purposes, such as a better trustworthiness of the model's output by the audience.

2.3. Why?

As stated in the introduction, explainability is one of the main barriers AI is facing nowadays in regards to its practical implementation. The inability to explain or to fully understand the reasons by which state-of-the-art ML algorithms perform as well as they do, is a problem that find its roots in two different causes, which are conceptually illustrated in Figure 2.

Without a doubt, the first cause is the gap between the research community and business sectors, impeding the full penetration of the newest ML models in sectors that have traditionally lagged behind in the digital transformation of their processes, such as banking, finances, security and health, among many others. In general this issue occurs in strictly regulated sectors with some reluctance to implement techniques that may put at risk their assets.

The second axis is that of knowledge. AI has helped research across the world with the task of inferring relations that were far beyond the human cognitive reach. Every field dealing with huge amounts of reliable data has largely benefited from the adoption of AI and ML techniques. However, we are entering an era in which results and performance metrics are the only interest shown up in research studies. Although for certain disciplines this might be the fair case, science and society are far from being concerned just by performance. The search for understanding is what opens the door for further model improvement and its practical utility.

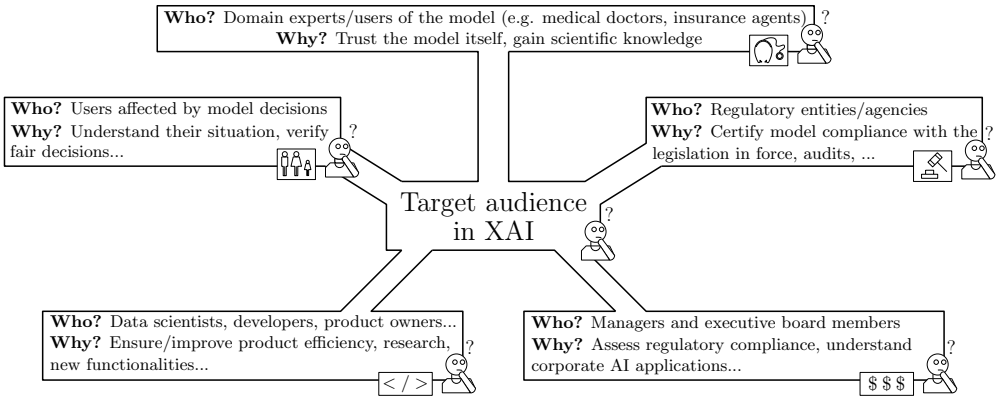


Figure 2: Diagram showing the different purposes of explainability in ML models sought by different audience profiles. Two goals occur to prevail across them: need for model understanding, and regulatory compliance. Image partly inspired by the one presented in [29], used with permission from IBM.

The following section develops these ideas further by analyzing the goals motivating the search for explainable AI models.

2.4. What for?

The research activity around XAI has so far exposed different goals to draw from the achievement of an explainable model. Almost none of the papers reviewed completely agrees in the goals required to describe what an explainable model should compel. However, all these different goals might help discriminate the purpose for which a given exercise of ML explainability is performed. Unfortunately, scarce contributions have attempted to define such goals from a conceptual perspective [5, 13, 24, 30]. We now synthesize and enumerate definitions for these XAI goals, so as to settle a first classification criteria for the full suit of papers covered in this review:

XAI Goal	**Main target audience** (Fig. 2)	**References**
Trustworthiness	Domain experts, users of the model affected by decisions	[5, 10, 24, 32, 33, 34, 35, 36, 37]
Causality	Domain experts, managers and executive board members, regulatory entities/agencies	[35, 38, 39, 40, 41, 42, 43]
Transferability	Domain experts, data scientists	[5, 44, 21, 26, 45, 30, 32, 37, 38, 39, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85]
Informativeness	All	[5, 44, 21, 25, 26, 45, 30, 32, 34, 35, 37, 38, 41, 46, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 63, 64, 65, 66, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 86, 87, 88, 89, 59, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100, 101, 102, 103, 104, 105, 106, 107, 108, 109, 110, 111, 112, 113, 114, 115, 116, 117, 118, 119, 120, 121, 122, 123, 124, 125, 126, 127, 128, 129, 130, 131, 132, 133, 134, 135, 136, 137, 138, 139, 140, 141, 142, 143, 144, 145, 146, 147, 148, 149, 150, 151, 152, 153, 154]
Confidence	Domain experts, developers, managers, regulatory entities/agencies	[5, 45, 35, 46, 48, 54, 61, 72, 88, 89, 96, 108, 117, 119, 155]
Fairness	Users affected by model decisions, regulatory entities/agencies	[5, 24, 45, 35, 47, 99, 100, 101, 120, 121, 128, 156, 157, 158]
Accessibility	Product owners, managers, users affected by model decisions	[21, 26, 30, 32, 37, 50, 53, 55, 62, 67, 68, 69, 70, 71, 74, 75, 76, 86, 93, 94, 103, 105, 107, 108, 111, 112, 113, 114, 115, 124, 129]
Interactivity	Domain experts, users affected by model decisions	[37, 50, 59, 65, 67, 74, 86, 124]
Privacy awareness	Users affected by model decisions, regulatory entities/agencies	[89]

Table 1: Goals pursued in the reviewed literature toward reaching explainability, and their main target audience.

- *Trustworthiness:* several authors agree upon the search for trustworthiness as the primary aim of an explainable AI model [31, 32]. However, declaring a model as explainable as per its capabilities of inducing trust might not be fully compliant with the requirement of model explainability. Trustworthiness might be considered as the confidence of whether a model will act as intended when facing a given problem. Although it should most certainly be a property of any explainable model, it does not imply that every trustworthy model can be considered explainable on its own, nor is trustworthiness a property easy to quantify. Trust might be far from being the only purpose of an explainable model since the relation among the two, if agreed upon, is not reciprocal. Part of the reviewed papers mention the concept of trust when stating their purpose for achieving explainability. However, as seen in Table 1, they do not amount to a large share of the recent contributions related to XAI.

- *Causality:* another common goal for explainability is that of finding causality among data variables. Several authors argue that explainable models might ease the task of finding relationships that, should they occur, could be tested further for a stronger causal link between the involved variables [159, 160]. The inference of causal relationships from observational data is a field that has been broadly studied over time [161]. As widely acknowledged by the community working on this topic, causality requires a wide frame of prior knowledge to prove that observed effects are causal. A ML model only discovers correlations among the data it learns from, and therefore might not suffice for unveiling a cause-effect relationship. However, causation involves correlation, so an explainable ML model could validate the results provided by causality inference techniques, or provide a first intuition of possible causal

relationships within the available data. Again, Table 1 reveals that causality is not among the most important goals if we attend to the amount of papers that state it explicitly as their goal.

- *Transferability:* models are always bounded by constraints that should allow for their seamless transferability. This is the main reason why a training-testing approach is used when dealing with ML problems [162, 163]. Explainability is also an advocate for transferability, since it may ease the task of elucidating the boundaries that might affect a model, allowing for a better understanding and implementation. Similarly, the mere understanding of the inner relations taking place within a model facilitates the ability of a user to reuse this knowledge in another problem. There are cases in which the lack of a proper understanding of the model might drive the user toward incorrect assumptions and fatal consequences [44, 164]. Transferability should also fall between the resulting properties of an explainable model, but again, not every transferable model should be considered as explainable. As observed in Table 1, the amount of papers stating that the ability of rendering a model explainable is to better understand the concepts needed to reuse it or to improve its performance is the second most used reason for pursuing model explainability.

- *Informativeness:* ML models are used with the ultimate intention of supporting decision making [92]. However, it should not be forgotten that the problem being solved by the model is not equal to that being faced by its human counterpart. Hence, a great deal of information is needed in order to be able to relate the user's decision to the solution given by the model, and to avoid falling in misconception pitfalls. For this purpose, explainable ML models should give information about the problem being tackled. Most of the reasons found among the papers reviewed is that of extracting information about the inner relations of a model. Almost all rule extraction techniques substantiate their approach on the search for a simpler understanding of what the model internally does, stating that the knowledge (information) can be expressed in these simpler proxies that they consider explaining the antecedent. This is the most used argument found among the reviewed papers to back up what they expect from reaching explainable models.

- *Confidence:* as a generalization of robustness and stability, confidence should always be assessed on a model in which reliability is expected. The methods to maintain confidence under control are different depending on the model. As stated in [165, 166, 167], stability is a must-have when drawing interpretations from a certain model. Trustworthy interpretations should not be produced by models that are not stable. Hence, an explainable model should contain information about the confidence of its working regime.

- *Fairness:* from a social standpoint, explainability can be considered as the capacity to reach and guarantee fairness in ML models. In a certain literature strand, an explainable ML model suggests a clear visualization of the relations affecting a result, allowing for a fairness or ethical analysis of the model at hand [3, 100]. Likewise, a related objective of XAI is highlighting bias in the data a model was exposed to [168, 169]. The support of algorithms and models is growing fast in fields that involve human lives, hence explainability should be considered as a bridge to avoid the unfair or unethical use of algorithm's outputs.

- *Accessibility:* a minor subset of the reviewed contributions argues for explainability as the property that allows end users to get more involved in the process of improving and developing a certain ML model [37, 86] . It seems clear that explainable models will ease the burden felt by non-technical or non-expert users when having to deal with algorithms that seem incomprehensible at first sight. This concept is expressed as the third most considered goal among the surveyed literature.

- *Interactivity:* some contributions [50, 59] include the ability of a model to be interactive with the user as one of the goals targeted by an explainable ML model. Once again, this goal is related to fields in

which the end users are of great importance, and their ability to tweak and interact with the models is what ensures success.

- *Privacy awareness:* almost forgotten in the reviewed literature, one of the byproducts enabled by explainability in ML models is its ability to assess privacy. ML models may have complex representations of their learned patterns. Not being able to understand what has been captured by the model [4] and stored in its internal representation may entail a privacy breach. Contrarily, the ability to explain the inner relations of a trained model by non-authorized third parties may also compromise the differential privacy of the data origin. Due to its criticality in sectors where XAI is foreseen to play a crucial role, confidentiality and privacy issues will be covered further in Subsections 5.4 and 6.3, respectively.

This subsection has reviewed the goals encountered among the broad scope of the reviewed papers. All these goals are clearly under the surface of the concept of explainability introduced before in this section. To round up this prior analysis on the concept of explainability, the last subsection deals with different strategies followed by the community to address explainability in ML models.

2.5. How?

The literature makes a clear distinction among models that are interpretable by design, and those that can be explained by means of external XAI techniques. This duality could also be regarded as the difference between interpretable models and model interpretability techniques; a more widely accepted classification is that of *transparent* models and post-hoc explainability. This same duality also appears in the paper presented in [17] in which the distinction its authors make refers to the methods to solve the transparent box design problem against the problem of explaining the black-box problem. This work, further extends the distinction made among transparent models including the different levels of transparency considered.

Within transparency, three levels are contemplated: algorithmic transparency, decomposability and simulatability[1]. Among post-hoc techniques we may distinguish among *text explanations*, *visualizations*, *local explanations*, *explanations by example*, *explanations by simplification* and *feature relevance*. In this context, there is a broader distinction proposed by [24] discerning between 1) opaque systems, where the mappings from input to output are invisible to the user; 2) interpretable systems, in which users can mathematically analyze the mappings; and 3) comprehensible systems, in which the models should output symbols or rules along with their specific output to aid in the understanding process of the rationale behind the mappings being made. This last classification criterion could be considered included within the one proposed earlier, hence this paper will attempt at following the more specific one.

2.5.1. Levels of Transparency in Machine Learning Models

Transparent models convey some degree of interpretability by themselves. Models belonging to this category can be also approached in terms of the domain in which they are interpretable, namely, algorithmic transparency, decomposability and simulatability. As we elaborate next in connection to Figure 3, each of these classes contains its predecessors, e.g. a *simulatable* model is at the same time a model that is decomposable and algorithmically transparent:

- *Simulatability* denotes the ability of a model of being simulated or thought about strictly by a human, hence complexity takes a dominant place in this class. This being said, simple but extensive (i.e., with *too large* amount of rules) rule based systems fall out of this characteristic, whereas a single perceptron neural network falls within. This aspect aligns with the claim that sparse linear models are more interpretable than dense ones [170], and that an interpretable model is one that can be easily presented

[1]The alternative term *simulability* is also used in the literature to refer to the capacity of a system or process to be simulated. However, we note that this term does not appear in current English dictionaries.

to a human by means of text and *visualizations* [32]. Again, endowing a decomposable model with simulatability requires that the model has to be self-contained enough for a human to think and reason about it as a whole.

- *Decomposability* stands for the ability to explain each of the parts of a model (input, parameter and calculation). It can be considered as intelligibility as stated in [171]. This characteristic might empower the ability to understand, interpret or explain the behavior of a model. However, as occurs with algorithmic transparency, not every model can fulfill this property. Decomposability requires every input to be readily interpretable (e.g. cumbersome features will not fit the premise). The added constraint for an algorithmically transparent model to become decomposable is that every part of the model must be understandable by a human without the need for additional tools.

- *Algorithmic Transparency* can be seen in different ways. It deals with the ability of the user to understand the process followed by the model to produce any given output from its input data. Put it differently, a linear model is deemed transparent because its error surface can be understood and reasoned about, allowing the user to understand how the model will act in every situation it may face [163]. Contrarily, it is not possible to understand it in deep architectures as the loss landscape might be opaque [172, 173] since it cannot be fully observed and the solution has to be approximated through heuristic optimization (e.g. through stochastic gradient descent). The main constraint for algorithmically transparent models is that the model has to be fully explorable by means of mathematical analysis and methods.

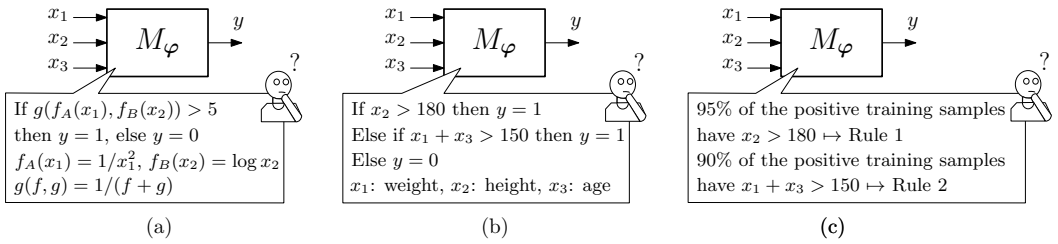


Figure 3: Conceptual diagram exemplifying the different levels of transparency characterizing a ML model M_{φ}, with φ denoting the parameter set of the model at hand: (a) simulatability; (b) decomposability; (c) algorithmic transparency. Without loss of generality, the example focuses on the ML model as the explanation target. However, other targets for explainability may include a given example, the output classes or the dataset itself.

2.5.2. Post-hoc Explainability Techniques for Machine Learning Models

Post-hoc explainability targets models that are not readily interpretable by design by resorting to diverse means to enhance their interpretability, such as *text explanations*, *visual explanations*, *local explanations*, *explanations by example*, *explanations by simplification* and *feature relevance explanations* techniques. Each of these techniques covers one of the most common ways humans explain systems and processes by themselves.

Further along this river, actual techniques, or better put, actual group of techniques are specified to ease the future work of any researcher that intends to look up for an specific technique that suits its knowledge. Not ending there, the classification also includes the type of data in which the techniques has been applied. Note that many techniques might be suitable for many different types of data, although the categorization only considers the type used by the authors that proposed such technique. Overall, post-hoc explainability techniques are divided first by the intention of the author (explanation technique e.g. Explanation by simplification), then, by the method utilized (actual technique e.g. sensitivity analysis) and finally by the type of data in which it was applied (e.g. images).

- *Text explanations* deal with the problem of bringing explainability for a model by means of learning to generate *text explanations* that help explaining the results from the model [169]. *Text explanations* also include every method generating symbols that represent the functioning of the model. These symbols may portrait the rationale of the algorithm by means of a semantic mapping from model to symbols.

- *Visual explanation* techniques for post-hoc explainability aim at visualizing the model's behavior. Many of the visualization methods existing in the literature come along with dimensionality reduction techniques that allow for a human interpretable simple visualization. Visualizations may be coupled with other techniques to improve their understanding, and are considered as the most suitable way to introduce complex interactions within the variables involved in the model to users not acquainted to ML modeling.

- *Local explanations* tackle explainability by segmenting the solution space and giving explanations to less complex solution subspaces that are relevant for the whole model. These explanations can be formed by means of techniques with the differentiating property that these only explain part of the whole system's functioning.

- *Explanations by example* consider the extraction of data examples that relate to the result generated by a certain model, enabling to get a better understanding of the model itself. Similarly to how humans behave when attempting to explain a given process, *explanations by example* are mainly centered in extracting representative examples that grasp the inner relationships and correlations found by the model being analyzed.

- *Explanations by simplification* collectively denote those techniques in which a whole new system is rebuilt based on the trained model to be explained. This new, simplified model usually attempts at optimizing its resemblance to its antecedent functioning, while reducing its complexity, and keeping a similar performance score. An interesting byproduct of this family of post-hoc techniques is that the simplified model is, in general, easier to be implemented due to its reduced complexity with respect to the model it represents.

- Finally, *feature relevance explanation* methods for post-hoc explainability clarify the inner functioning of a model by computing a relevance score for its managed variables. These scores quantify the affection (sensitivity) a feature has upon the output of the model. A comparison of the scores among different variables unveils the importance granted by the model to each of such variables when producing its output. *Feature relevance* methods can be thought to be an indirect method to explain a model.

The above classification (portrayed graphically in Figure 4) will be used when reviewing specific/agnostic XAI techniques for ML models in the following sections (Table 2). For each ML model, a distinction of the propositions to each of these categories is presented in order to pose an overall image of the field's trends.

3. Transparent Machine Learning Models

The previous section introduced the concept of *transparent* models. A model is considered to be transparent if by itself it is understandable. The models surveyed in this section are a suit of transparent models that can fall in one or all of the levels of model transparency described previously (namely, simulatability, decomposability and algorithmic transparency). In what follows we provide reasons for this statement, with graphical support given in Figure 5.

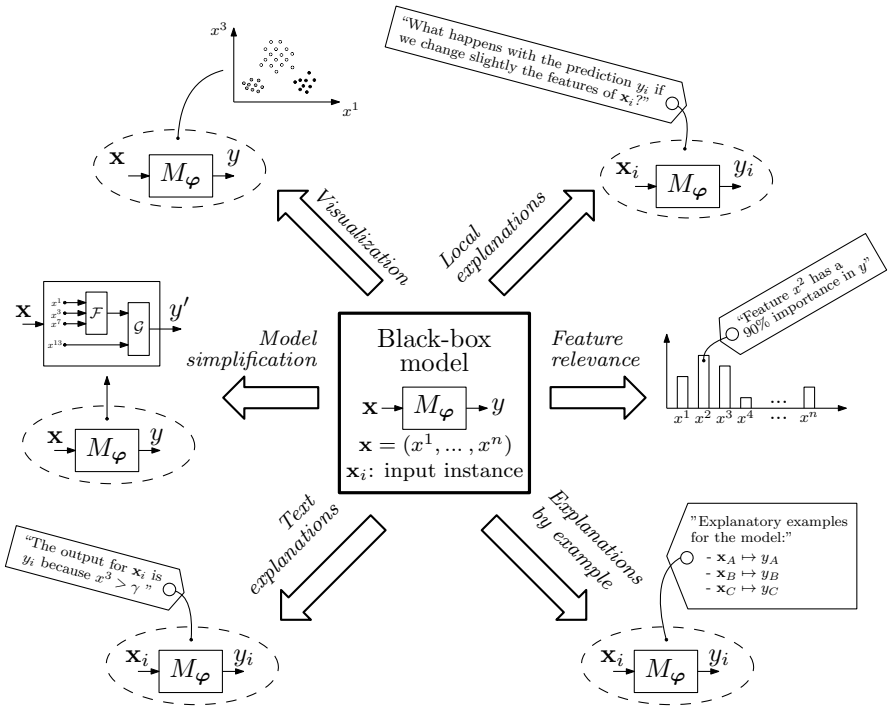


Figure 4: Conceptual diagram showing the different post-hoc explainability approaches available for a ML model M_{φ}.

3.1. Linear/Logistic Regression

Logistic Regression (LR) is a classification model to predict a dependent variable (category) that is dichotomous (binary). However, when the dependent variable is continuous, linear regression would be its homonym. This model takes the assumption of linear dependence between the predictors and the predicted variables, impeding a flexible fit to the data. This specific reason (stiffness of the model) is the one that maintains the model under the umbrella of transparent methods. However, as stated in Section 2, explainability is linked to a certain audience, which makes a model fall under both categories depending who is to interpret it. This way, logistic and linear regression, although clearly meeting the characteristics of transparent models (algorithmic transparency, decomposability and simulatability), may also demand post-hoc explainability techniques (mainly, visualization), particularly when the model is to be explained to non-expert audiences.

The usage of this model has been largely applied within Social Sciences for quite a long time, which has pushed researchers to create ways of explaining the results of the models to non-expert users. Most authors agree on the different techniques used to analyze and express the soundness of LR [174, 175, 176, 177], including the overall model evaluation, statistical tests of individual predictors, goodness-of-fit statistics and validation of the predicted probabilities. The overall model evaluation shows the improvement of the applied model over a baseline, showing if it is in fact improving the model without predictions. The statistical significance of single predictors is shown by calculating the Wald chi-square statistic. The goodness-of-fit statistics show the quality of fitness of the model to the data and how significant this is. This can be achieved by resorting to different techniques e.g. the so-called Hosmer-Lemeshow (H-L) statistic. The validation of predicted probabilities involves testing whether the output of the model corresponds to what is shown by the data. These techniques show mathematical ways of representing the fitness of the model and its behavior.

Other techniques from other disciplines besides Statistics can be adopted for explaining these re-

Model	Transparent ML Models			Post-hoc analysis
	Simulatability	**Decomposability**	**Algorithmic Transparency**	
Linear/Logistic Regression	Predictors are human readable and interactions among them are kept to a minimum	Variables are still readable, but the number of interactions and predictors involved in them have grown to force decomposition	Variables and interactions are too complex to be analyzed without mathematical tools	Not needed
Decision Trees	A human can simulate and obtain the prediction of a decision tree on his/her own, without requiring any mathematical background	The model comprises rules that do not alter data whatsoever, and preserves their readability	Human-readable rules that explain the knowledge learned from data and allows for a direct understanding of the prediction process	Not needed
K-Nearest Neighbors	The complexity of the model (number of variables, their understandability and the similarity measure under use) matches human naive capabilities for simulation	The amount of variables is too high and/or the similarity measure is too complex to be able to simulate the model completely, but the similarity measure and the set of variables can be decomposed and analyzed separately	The similarity measure cannot be decomposed and/or the number of variables is so high that the user has to rely on mathematical and statistical tools to analyze the model	Not needed
Rule Based Learners	Variables included in rules are readable, and the size of the rule set is manageable by a human user without external help	The size of the rule set becomes too large to be analyzed without decomposing it into small rule chunks	Rules have become so complicated (and the rule set size has grown so much) that mathematical tools are needed for inspecting the model behaviour	Not needed
General Additive Models	Variables and the interaction among them as per the smooth functions involved in the model must be constrained within human capabilities for understanding	Interactions become too complex to be simulated, so decomposition techniques are required for analyzing the model	Due to their complexity, variables and interactions cannot be analyzed without the application of mathematical and statistical tools	Not needed
Bayesian Models	Statistical relationships modeled among variables and the variables themselves should be directly understandable by the target audience	Statistical relationships involve so many variables that they must be decomposed in marginals so as to ease their analysis	Statistical relationships cannot be interpreted even if already decomposed, and predictors are so complex that model can be only analyzed with mathematical tools	Not needed
Tree Ensembles	✗	✗	✗	Needed: Usually *Model simplification* or *Feature relevance* techniques
Support Vector Machines	✗	✗	✗	Needed: Usually *Model simplification* or *Local explanations* techniques
Multi–layer Neural Network	✗	✗	✗	Needed: Usually *Model simplification*, *Feature relevance* or *Visualization* techniques
Convolutional Neural Network	✗	✗	✗	Needed: Usually *Feature relevance* or *Visualization* techniques
Recurrent Neural Network	✗	✗	✗	Needed: Usually *Feature relevance* techniques

Table 2: Overall picture of the classification of ML models attending to their level of explainability.

gression models. Visualization techniques are very powerful when presenting statistical conclusions to users not well-versed in statistics. For instance, the work in [178] shows that the usage of probabilities to communicate the results, implied that the users where able to estimate the outcomes correctly in 10% of the cases, as opposed to 46% of the cases when using natural frequencies. Although logistic regression is among the simplest classification models in supervised learning, there are concepts that must be taken care of.

In this line of reasoning, the authors of [179] unveil some concerns with the interpretations derived from LR. They first mention how dangerous it might be to interpret log odds ratios and odd ratios as substantive effects, since they also represent unobserved heterogeneity. Linked to this first concern, [179] also states that a comparison between these ratios across models with different variables might be problematic, since the unobserved heterogeneity is likely to vary, thereby invalidating the comparison. Finally they also mention that the comparison of these odds across different samples, groups and time is also risky, since the variation of the heterogeneity is not known across samples, groups and time points. This last paper serves the purpose of visualizing the problems a model's interpretation might entail, even when its construction is as simple as that of LR.

Also interesting is to note that, for a model such as logistic or linear regression to maintain decomposability and simulatability, its size must be limited, and the variables used must be understandable by their users. As stated in Section 2, if inputs to the model are highly engineered features that are complex or difficult to understand, the model at hand will be far from being *decomposable*. Similarly, if the model is so large that a human cannot think of the model as a whole, its simulatability will be put to question.

3.2. Decision Trees

Decision trees are another example of a model that can easily fulfill every constraint for transparency. Decision trees are hierarchical structures for decision making used to support regression and classification problems [132, 180]. In the simplest of their flavors, decision trees are *simulatable* models. However, their properties can render them *decomposable* or *algorithmically transparent*.

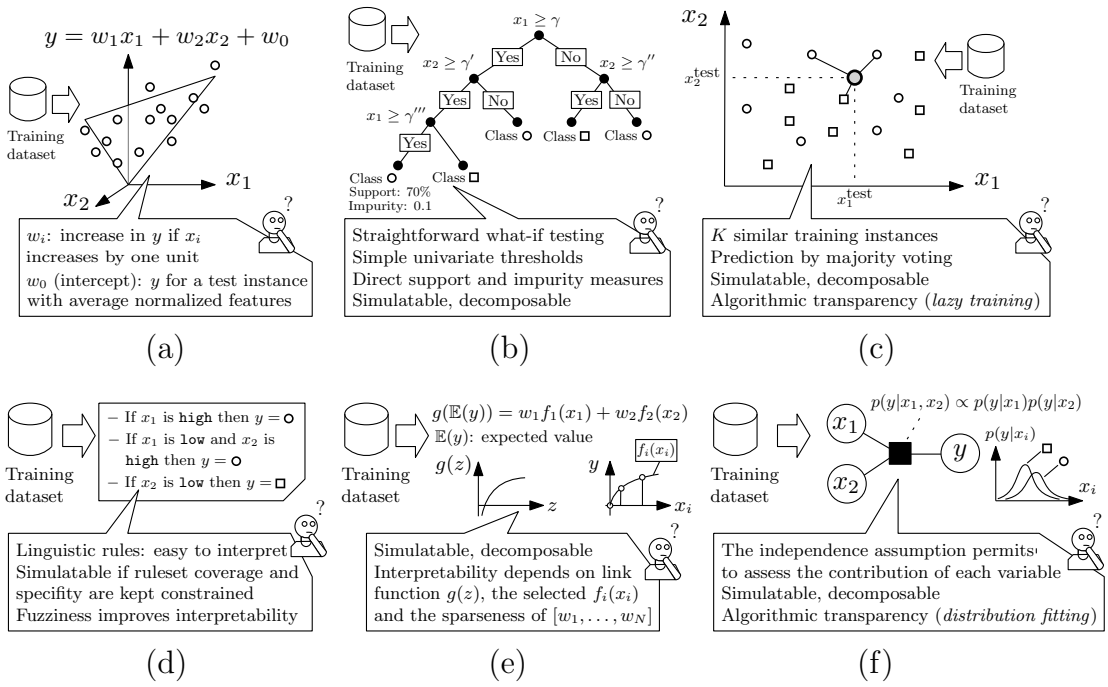


Figure 5: Graphical illustration of the levels of transparency of different ML models considered in this overview: (a) Linear regression; (b) Decision trees; (c) K-Nearest Neighbors; (d) Rule-based Learners; (e) Generalized Additive Models; (f) Bayesian Models.

Decision trees have always lingered in between the different categories of transparent models. Their utilization has been closely linked to decision making contexts, being the reason why their complexity and understandability have always been considered a paramount matter. A proof of this relevance can be found in the upsurge of contributions to the literature dealing with decision tree simplification and generation [132, 180, 181, 182]. As noted above, although being capable of fitting every category within transparent models, the individual characteristics of decision trees can push them toward the category of algorithmically transparent models. A *simulatable* decision tree is one that is manageable by a human user. This means its size is somewhat small and the amount of features and their meaning are easily understandable. An increment in size transforms the model into a *decomposable* one since its size impedes its full evaluation (simulation) by a human. Finally, further increasing its size and using complex feature relations will make the model *algorithmically transparent* loosing the previous characteristics.

Decision trees have long been used in decision support contexts due to their off-the-shelf transparency. Many applications of these models fall out of the fields of computation and AI (even information technologies), meaning that experts from other fields usually feel comfortable interpreting the outputs of these models [183, 184, 185]. However, their poor generalization properties in comparison with other models make this model family less interesting for their application to scenarios where a balance between predictive performance is a design driver of utmost importance. Tree ensembles aim at overcoming such a poor performance by aggregating the predictions performed by trees learned on different subsets of training data. Unfortunately, the combination of decision trees looses every transparent property, calling for the adoption of post-hoc explainability techniques as the ones reviewed later in the manuscript.

3.3. K-Nearest Neighbors

Another method that falls within transparent models is that of K-Nearest Neighbors (KNN), which deals with classification problems in a methodologically simple way: it predicts the class of a test sample by voting the classes of its K nearest neighbors (where the neighborhood relation is induced by a measure of distance between samples). When used in the context of regression problems, the voting is replaced by an aggregation (e.g. average) of the target values associated with the nearest neighbors.

In terms of model explainability, it is important to observe that predictions generated by KNN models rely on the notion of distance and similarity between examples, which can be tailored depending on the specific problem being tackled. Interestingly, this prediction approach resembles that of experience-based human decision making, which decides upon the result of past similar cases. There lies the rationale of why KNN has also been adopted widely in contexts in which model interpretability is a requirement [186, 187, 188, 189]. Furthermore, aside from being simple to explain, the ability to inspect the reasons by which a new sample has been classified inside a group and to examine how these predictions evolve when the number of neighbors K is increased or decreased empowers the interaction between the users and the model.

One must keep in mind that as mentioned before, KNN's class of transparency depends on the features, the number of neighbors and the distance function used to measure the similarity between data instances. A very high K impedes a full simulation of the model performance by a human user. Similarly, the usage of complex features and/or distance functions would hinder the decomposability of the model, restricting its interpretability solely to the transparency of its algorithmic operations.

3.4. Rule-based Learning

Rule-based learning refers to every model that generates rules to characterize the data it is intended to learn from. Rules can take the form of simple conditional *if-then* rules or more complex combinations of simple rules to form their knowledge. Also connected to this general family of models, fuzzy rule based systems are designed for a broader scope of action, allowing for the definition of verbally formulated rules over imprecise domains. Fuzzy systems improve two main axis relevant for this paper. First, they empower more understandable models since they operate in linguistic terms. Second, they perform better that classic rule systems in contexts with certain degrees of uncertainty. Rule based learners are clearly transparent models that have been often used to explain complex models by generating rules that explain their predictions [126, 127, 190, 191].

Rule learning approaches have been extensively used for knowledge representation in expert systems [192]. However, a central problem with rule generation approaches is the coverage (amount) and the specificity (length) of the rules generated. This problem relates directly to the intention for their use in the first place. When building a rule database, a typical design goal sought by the user is to be able to analyze and understand the model. The amount of rules in a model will clearly improve the performance of the model at the stake of compromising its intepretability. Similarly, the specificity of the rules plays also against interpretability, since a rule with a high number of antecedents an/or consequences might become difficult to interpret. In this same line of reasoning, these two features of a rule based learner play along with the classes of transparent models presented in Section 2. The greater the coverage or the specificity is, the closer the model will be to being just *algorithmically transparent*. Sometimes, the reason to transition from classical rules to fuzzy rules is to relax the constraints of rule sizes, since a greater range can be covered with less stress on interpretability.

Rule based learners are great models in terms of interpretability across fields. Their natural and seamless relation to human behaviour makes them very suitable to understand and explain other models. If a certain threshold of coverage is acquired, a rule wrapper can be thought to contain enough information about a model to explain its behavior to a non-expert user, without forfeiting the possibility of using the generated rules as an standalone prediction model.

3.5. General Additive Models

In statistics, a Generalized Additive Model (GAM) is a linear model in which the value of the variable to be predicted is given by the aggregation of a number of unknown smooth functions defined for the predictor variables. The purpose of such model is to infer the smooth functions whose aggregate composition approximates the predicted variable. This structure is easily interpretable, since it allows the user to verify the importance of each variable, namely, how it affects (through its corresponding function) the predicted output.

Similarly to every other transparent model, the literature is replete with case studies where GAMs are in use, specially in fields related to risk assessment. When compared to other models, these are understandable enough to make users feel confident on using them for practical applications in finance [193, 194, 195], environmental studies [196], geology [197], healthcare [44], biology [198, 199] and energy [200]. Most of these contributions use visualization methods to further ease the interpretation of the model. GAMs might be also considered as *simulatable* and *decomposable* models if the properties mentioned in its definitions are fulfilled, but to an extent that depends roughly on eventual modifications to the baseline GAM model, such as the introduction of link functions to relate the aggregation with the predicted output, or the consideration of interactions between predictors.

All in all, applications of GAMs like the ones exemplified above share one common factor: understandability. The main driver for conducting these studies with GAMs is to understand the underlying relationships that build up the cases for scrutiny. In those cases the research goal is not accuracy for its own sake, but rather the need for understanding the problem behind and the relationship underneath the variables involved in data. This is why GAMs have been accepted in certain communities as their *de facto* modeling choice, despite their acknowledged misperforming behavior when compared to more complex counterparts.

3.6. Bayesian Models

A Bayesian model usually takes the form of a probabilistic directed acyclic graphical model whose links represent the conditional dependencies between a set of variables. For example, a Bayesian network could represent the probabilistic relationships between diseases and symptoms. Given symptoms, the network can be used to compute the probabilities of the presence of various diseases. Similar to GAMs, these models also convey a clear representation of the relationships between features and the target, which in this case are given explicitly by the connections linking variables to each other.

Once again, Bayesian models fall below the ceiling of Transparent models. Its categorization leaves it under *simulatable*, *decomposable* and *algorithmically transparent*. However, it is worth noting that under certain circumstances (overly complex or cumbersome variables), a model may loose these first two properties. Bayesian models have been shown to lead to great insights in assorted applications such as cognitive modeling [201, 202], fishery [196, 203], gaming [204], climate [205], econometrics [206] or robotics [207]. Furthermore, they have also been utilized to explain other models, such as averaging tree ensembles [208].

4. Post-hoc Explainability Techniques for Machile Learning Models: Taxonomy, Shallow Models and Deep Learning

When ML models do not meet any of the criteria imposed to declare them transparent, a separate method must be devised and applied to the model to explain its decisions. This is the purpose of post-hoc explainability techniques (also referred to as post-modeling explainability), which aim at communicating understandable information about how an already developed model produces its predictions for any given input. In this section we categorize and review different algorithmic approaches for post-hoc explainability, discriminating among 1) those that are designed for their application to ML models of any kind; and 2) those that are designed for a specific ML model and thus, can not be directly extrapolated to any other

learner. We now elaborate on the trends identified around post-hoc explainability for different ML models, which are illustrated in Figure 6 in the form of hierarchical bibliographic categories and summarized next:

- Model-agnostic techniques for post-hoc explainability (Subsection 4.1), which can be applied seamlessly to any ML model disregarding its inner processing or internal representations.
- Post-hoc explainability that are tailored or specifically designed to explain certain ML models. We divide our literature analysis into two main branches: contributions dealing with post-hoc explainability of *shallow* ML models, which collectively refers to all ML models that do not hinge on layered structures of neural processing units (Subsection 4.2); and techniques devised for *deep* learning models, which correspondingly denote the family of neural networks and related variants, such as convolutional neural networks, recurrent neural networks (Subsection 4.3) and hybrid schemes encompassing deep neural networks and transparent models. For each model we perform a thorough review of the latest post-hoc methods proposed by the research community, along with a identification of trends followed by such contributions.
- We end our literature analysis with Subsection 4.4, where we present a second taxonomy that complements the more general one in Figure 6 by classifying contributions dealing with the post-hoc explanation of Deep Learning models. To this end we focus on particular aspects related to this family of black-box ML methods, and expose how they link to the classification criteria used in the first taxonomy.

4.1. Model-agnostic Techniques for Post-hoc Explainability

Model-agnostic techniques for post-hoc explainability are designed to be plugged to any model with the intent of extracting some information from its prediction procedure. Sometimes, simplification techniques are used to generate proxies that mimic their antecedents with the purpose of having something tractable and of reduced complexity. Other times, the intent focuses on extracting knowledge directly from the models or simply visualizing them to ease the interpretation of their behavior. Following the taxonomy introduced in Section 2, model-agnostic techniques may rely on *model simplification*, *feature relevance* estimation and *visualization* techniques:

- *Explanation by simplification*. They are arguably the broadest technique under the category of model agnostic post-hoc methods. *Local explanations* are also present within this category, since sometimes, simplified models are only representative of certain sections of a model. Almost all techniques taking this path for *model simplification* are based on rule extraction techniques. Among the most known contributions to this approach we encounter the technique of Local Interpretable Model-Agnostic Explanations (LIME) [32] and all its variations [214, 216]. LIME builds locally linear models around the predictions of an opaque model to explain it. These contributions fall under explanations by simplification as well as under *local explanations*. Besides LIME and related flavors, another approach to rule extraction is G-REX [212]. Although it was not originally intended for extracting rules from opaque models, the generic proposition of G-REX has been extended to also account for model explainability purposes [190, 211]. In line with rule extraction methods, the work in [215] presents a novel approach to learn rules in CNF (Conjunctive Normal Form) or DNF (Disjunctive Normal Form) to bridge from a complex model to a human-interpretable model. Another contribution that falls off the same branch is that in [218], where the authors formulate *model simplification* as a model extraction process by approximating a transparent model to the complex one. Simplification is approached from a different perspective in [120], where an approach to distill and audit black box models is presented. In it, two main ideas are exposed: a method for model distillation and comparison to audit black-box risk scoring models; and an statistical test to check if the auditing data is missing key features it was trained with. The popularity of *model simplification* is evident, given it temporally coincides with the most

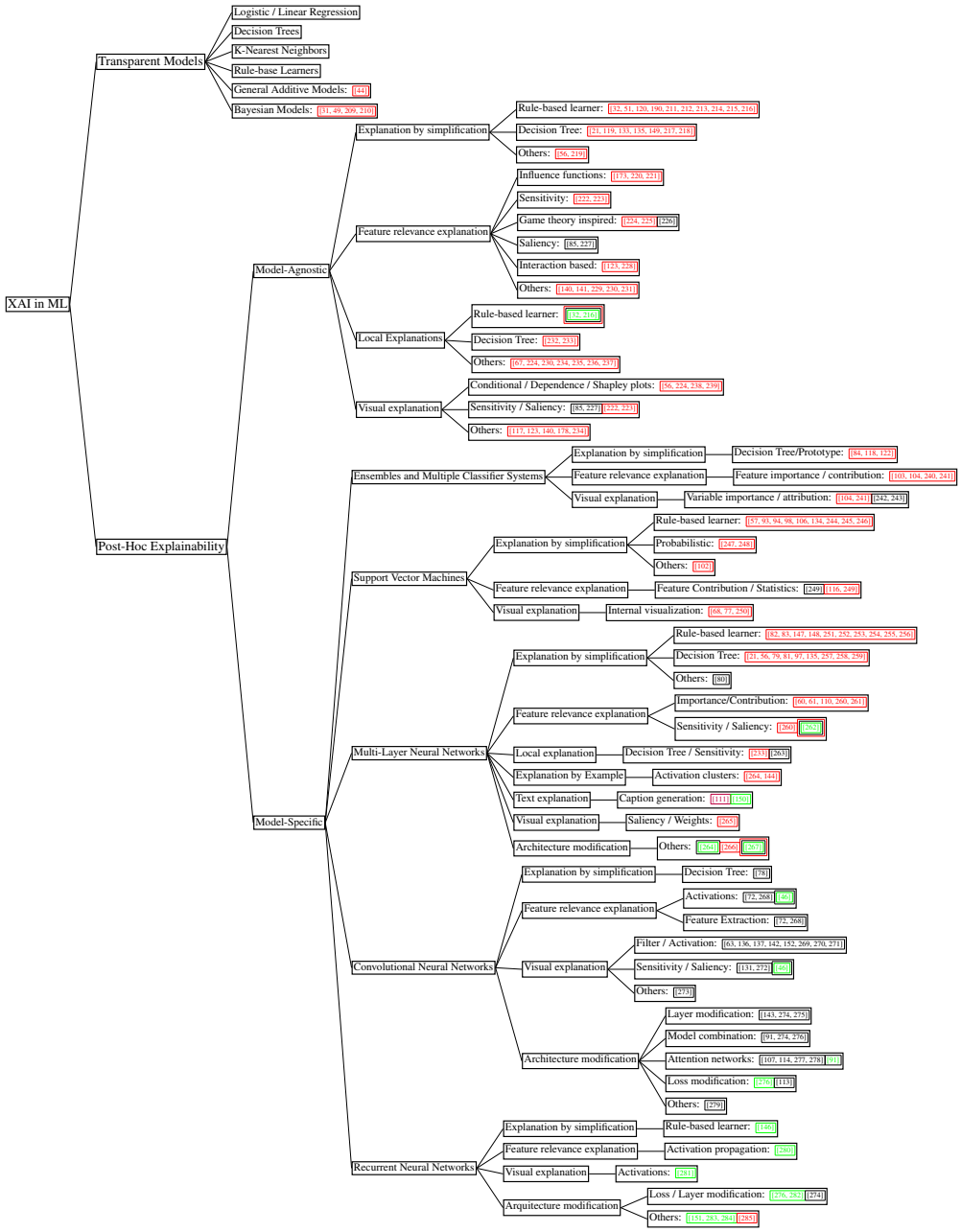


Figure 6: Taxonomy of the reviewed literature and trends identified for explainability techniques related to different ML models. References boxed in blue, green and red correspond to XAI techniques using image, text or tabular data, respectively. In order to build this taxonomy, the literature has been analyzed in depth to discriminate whether a post-hoc technique can be seamlessly applied to any ML model, even if, e.g., explicitly mentions *Deep Learning* in its title and/or abstract.

recent literature on XAI, including techniques such as LIME or G-REX. This symptomatically reveals that this post-hoc explainability approach is envisaged to continue playing a central role on XAI.

- *Feature relevance explanation* techniques aim to describe the functioning of an opaque model by

ranking or measuring the influence, relevance or importance each feature has in the prediction output by the model to be explained. An amalgam of propositions are found within this category, each resorting to different algorithmic approaches with the same targeted goal. One fruitful contribution to this path is that of [224] called SHAP (SHapley Additive exPlanations). Its authors presented a method to calculate an additive feature importance score for each particular prediction with a set of desirable properties (local accuracy, *missingness* and consistency) that its antecedents lacked. Another approach to tackle the contribution of each feature to predictions has been coalitional Game Theory [225] and local gradients [234]. Similarly, by means of local gradients [230] test the changes needed in each feature to produce a change in the output of the model. In [228] the authors analyze the relations and dependencies found in the model by grouping features, that combined, bring insights about the data. The work in [173] presents a broad variety of measures to tackle the quantification of the degree of influence of inputs on outputs of systems. Their QII (Quantitative Input Influence) measures account for correlated inputs while measuring influence. In contrast, in [222] the authors build upon the existing SA (Sensitivity Analysis) to construct a Global SA which extends the applicability of the existing methods. In [227] a real-time image saliency method is proposed, which is applicable to differentiable image classifiers. The study in [123] presents the so-called Automatic STRucture IDentification method (ASTRID) to inspect which attributes are exploited by a classifier to generate a prediction. This method finds the largest subset of features such that the accuracy of a classifier trained with this subset of features cannot be distinguished in terms of accuracy from a classifier built on the original feature set. In [221] the authors use influence functions to trace a model's prediction back to the training data, by only requiring an oracle version of the model with access to gradients and Hessian-vector products. Heuristics for creating counterfactual examples by modifying the input of the model have been also found to contribute to its explainability [236, 237]. Compared to those attempting explanations by simplification, a similar amount of publications were found tackling explainability by means of *feature relevance* techniques. Many of the contributions date from 2017 and some from 2018, implying that as with *model simplification* techniques, *feature relevance* has also become a vibrant subject study in the current XAI landscape.

- *Visual explanation* techniques are a vehicle to achieve model-agnostic explanations. Representative works in this area can be found in [222], which present a portfolio of visualization techniques to help in the explanation of a black-box ML model built upon the set of extended techniques mentioned earlier (Global SA). Another set of visualization techniques is presented in [223]. The authors present three novel SA methods (data based SA, Monte-Carlo SA, cluster-based SA) and one novel input importance measure (Average Absolute Deviation). Finally, [238] presents ICE (Individual Conditional Expectation) plots as a tool for visualizing the model estimated by any supervised learning algorithm. Visual explanations are less common in the field of model-agnostic techniques for post-hoc explainability. Since the design of these methods must ensure that they can be seamlessly applied to any ML model disregarding its inner structure, creating *visualizations* from just inputs and outputs from an opaque model is a complex task. This is why almost all visualization methods falling in this category work along with *feature relevance* techniques, which provide the information that is eventually displayed to the end user.

Several trends emerge from our literature analysis. To begin with, rule extraction techniques prevail in model-agnostic contributions under the umbrella of post-hoc explainability. This could have been intuitively expected if we bear in mind the wide use of rule based learning as explainability wrappers anticipated in Section 3.4, and the complexity imposed by not being able to *get into* the model itself. Similarly, another large group of contributions deals with *feature relevance*. Lately these techniques are gathering much attention by the community when dealing with DL models, with hybrid approaches that utilize particular aspects of this class of models and therefore, compromise the independence of the *feature relevance* method on the model being explained. Finally, visualization techniques propose interesting

ways for visualizing the output of *feature relevance* techniques to ease the task of model's interpretation. By contrast, visualization techniques for other aspects of the trained model (e.g. its structure, operations, etc) are tightly linked to the specific model to be explained.

4.2. Post-hoc Explainability in Shallow ML Models

Shallow ML covers a diversity of supervised learning models. Within these models, there are strictly interpretable (transparent) approaches (e.g. KNN and Decision Trees, already discussed in Section 3). However, other shallow ML models rely on more sophisticated learning algorithms that require additional layers of explanation. Given their prominence and notable performance in predictive tasks, this section concentrates on two popular shallow ML models (tree ensembles and Support Vector Machines, SVMs) that require the adoption of post-hoc explainability techniques for explaining their decisions.

4.2.1. Tree Ensembles, Random Forests and Multiple Classifier Systems

Tree ensembles are arguably among the most accurate ML models in use nowadays. Their advent came as an efficient means to improve the generalization capability of single decision trees, which are usually prone to overfitting. To circumvent this issue, tree ensembles combine different trees to obtain an aggregated prediction/regression. While it results to be effective against overfitting, the combination of models makes the interpretation of the overall ensemble more complex than each of its compounding tree learners, forcing the user to draw from post-hoc explainability techniques. For tree ensembles, techniques found in the literature are explanation by simplification and *feature relevance* techniques; we next examine recent advances in these techniques.

To begin with, many contributions have been presented to simplify tree ensembles while maintaining part of the accuracy accounted for the added complexity. The author from [119] poses the idea of training a single albeit less complex model from a set of random samples from the data (ideally following the real data distribution) labeled by the ensemble model. Another approach for simplification is that in [118], in which authors create a Simplified Tree Ensemble Learner (STEL). Likewise, [122] presents the usage of two models (simple and complex) being the former the one in charge of interpretation and the latter of prediction by means of Expectation-Maximization and Kullback-Leibler divergence. As opposed to what was seen in model-agnostic techniques, not that many techniques to board explainability in tree ensembles by means of *model simplification*. It derives from this that either the proposed techniques are good enough, or model-agnostic techniques do cover the scope of simplification already.

Following simplification procedures, *feature relevance* techniques are also used in the field of tree ensembles. Breiman [286] was the first to analyze the variable importance within Random Forests. His method is based on measuring MDA (Mean Decrease Accuracy) or MIE (Mean Increase Error) of the forest when a certain variable is randomly permuted in the out-of-bag samples. Following this contribution [241] shows, in an real setting, how the usage of variable importance reflects the underlying relationships of a complex system modeled by a Random Forest. Finally, a crosswise technique among post-hoc explainability, [240] proposes a framework that poses recommendations that, if taken, would convert an example from one class to another. This idea attempts to disentangle the variables importance in a way that is further descriptive. In the article, the authors show how these methods can be used to elevate recommendations to improve malicious online ads to make them rank higher in paying rates.

Similar to the trend shown in model-agnostic techniques, for tree ensembles again, simplification and *feature relevance* techniques seem to be the most used schemes. However, contrarily to what was observed before, most papers date back from 2017 and place their focus mostly on bagging ensembles. When shifting the focus towards other ensemble strategies, scarce activity has been recently noted around the explainability of boosting and stacking classifiers. Among the latter, it is worth highlighting the connection between the reason why a compounding learner of the ensemble produces an specific prediction on a given data, and its contribution to the output of the ensemble. The so-called Stacking With Auxiliary Features (SWAF) approach proposed in [242] points in this direction by harnessing and integrating explanations in

stacking ensembles to improve their generalization. This strategy allows not only relying on the output of the compounding learners, but also on the origin of that output and its consensus across the entire ensemble. Other interesting studies on the explainability of ensemble techniques include model-agnostic schemes such as DeepSHAP [226], put into practice with stacking ensembles and multiple classifier systems in addition to Deep Learning models; the combination of explanation maps of multiple classifiers to produce improved explanations of the ensemble to which they belong [243]; and recent insights dealing with traditional and gradient boosting ensembles [287, 288].

4.2.2. Support Vector Machines

Another shallow ML model with historical presence in the literature is the SVM. SVM models are more complex than tree ensembles, with a much opaquer structure. Many implementations of post-hoc explainability techniques have been proposed to relate what is mathematically described internally in these models, to what different authors considered explanations about the problem at hand. Technically, an SVM constructs a hyper-plane or set of hyper-planes in a high or infinite-dimensional space, which can be used for classification, regression, or other tasks such as outlier detection. Intuitively, a good separation is achieved by the hyperplane that has the largest distance (so-called functional margin) to the nearest training-data point of any class, since in general, the larger the margin, the lower the generalization error of the classifier. SVMs are among the most used ML models due to their excellent prediction and generalization capabilities. From the techniques stated in Section 2, post-hoc explainability applied to SVMs covers explanation by *simplification*, *local explanations*, *visualizations* and *explanations by example*.

Among explanation by simplification, four classes of simplifications are made. Each of them differentiates from the other by how deep they go into the algorithm inner structure. First, some authors propose techniques to build rule based models only from the support vectors of a trained model. This is the approach of [93], which proposes a method that extracts rules directly from the support vectors of a trained SVM using a modified sequential covering algorithm. In [57] the same authors propose eclectic rule extraction, still considering only the support vectors of a trained model. The work in [94] generates fuzzy rules instead of classical propositional rules. Here, the authors argue that long antecedents reduce comprehensibility, hence, a fuzzy approach allows for a more linguistically understandable result. The second class of simplifications can be exemplified by [98], which proposed the addition of the SVM's hyperplane, along with the support vectors, to the components in charge of creating the rules. His method relies on the creation of hyper-rectangles from the intersections between the support vectors and the hyper-plane. In a third approach to *model simplification*, another group of authors considered adding the actual training data as a component for building the rules. In [126, 244, 246] the authors proposed a clustering method to group prototype vectors for each class. By combining them with the support vectors, it allowed defining ellipsoids and hyper-rectangles in the input space. Similarly in [106], the authors proposed the so-called Hyper-rectangle Rule Extraction, an algorithm based on SVC (Support Vector Clustering) to find prototype vectors for each class and then define small hyper-rectangles around. In [105], the authors formulate the rule extraction problem as a multi-constrained optimization to create a set of non-overlapping rules. Each rule conveys a non-empty hyper-cube with a shared edge with the hyper-plane. In a similar study conducted in [245], extracting rules for gene expression data, the authors presented a novel technique as a component of a multi-kernel SVM. This multi-kernel method consists of feature selection, prediction modeling and rule extraction. Finally, the study in [134] makes use of a growing SVC to give an interpretation to SVM decisions in terms of linear rules that define the space in Voronoi sections from the extracted prototypes.

Leaving aside rule extraction, the literature has also contemplated some other techniques to contribute to the interpretation of SVMs. Three of them (visualization techniques) are clearly used toward explaining SVM models when used for concrete applications. For instance, [77] presents an innovative approach to visualize trained SVM to extract the information content from the kernel matrix. They center the study

on Support Vector Regression models. They show the ability of the algorithm to visualize which of the input variables are actually related with the associated output data. In [68] a visual way combines the output of the SVM with heatmaps to guide the modification of compounds in late stages of drug discovery. They assign colors to atoms based on the weights of a trained linear SVM that allows for a much more comprehensive way of debugging the process. In [116] the authors argue that many of the presented studies for interpreting SVMs only account for the weight vectors, leaving the margin aside. In their study they show how this margin is important, and they create an statistic that explicitly accounts for the SVM margin. The authors show how this statistic is specific enough to explain the multivariate patterns shown in neuroimaging.

Noteworthy is also the intersection between SVMs and Bayesian systems, the latter being adopted as a post-hoc technique to explain decisions made by the SVM model. This is the case of [248] and [247], which are studies where SVMs are interpreted as MAP (Maximum A Posteriori) solutions to inference problems with Gaussian Process priors. This framework makes tuning the hyper-parameters comprehensible and gives the capability of predicting class probabilities instead of the classical binary classification of SVMs. Interpretability of SVM models becomes even more involved when dealing with non-CPD (Conditional Positive Definite) kernels that are usually harder to interpret due to missing geometrical and theoretical understanding. The work in [102] revolves around this issue with a geometrical interpretation of indefinite kernel SVMs, showing that these do not classify by hyper-plane margin optimization. Instead, they minimize the distance between convex hulls in pseudo-Euclidean spaces.

A difference might be appreciated between the post-hoc techniques applied to other models and those noted for SVMs. In previous models, *model simplification* in a broad sense was the prominent method for post-hoc explainability. In SVMs, *local explanations* have started to take some weight among the propositions. However, simplification based methods are, on average, much older than local explanations.

As a final remark, none of the reviewed methods treating SVM explainability are dated beyond 2017, which might be due to the progressive proliferation of DL models in almost all disciplines. Another plausible reason is that these models are already understood, so it is hard to improve upon what has already been done.

4.3. Explainability in Deep Learning

Post-hoc *local explanations* and *feature relevance* techniques are increasingly the most adopted methods for explaining DNNs. This section reviews explainability studies proposed for the most used DL models, namely multi-layer neural networks, Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN).

4.3.1. Multi-layer Neural Networks

From their inception, multi-layer neural networks (also known as multi-layer perceptrons) have been warmly welcomed by the academic community due to their huge ability to infer complex relations among variables. However, as stated in the introduction, developers and engineers in charge of deploying these models in real-life production find in their questionable explainability a common reason for reluctance. That is why neural networks have been always considered as black-box models. The fact that explainability is often a must for the model to be of practical value, forced the community to generate multiple explainability techniques for multi-layer neural networks, including *model simplification* approaches, *feature relevance* estimators, *text explanations*, *local explanations* and model *visualizations*.

Several *model simplification* techniques have been proposed for neural networks with one single hidden layer, however very few works have been presented for neural networks with multiple hidden layers. One of these few works is DeepRED algorithm [257], which extends the decompositional approach to rule extraction (splitting at neuron level) presented in [259] for multi-layer neural network by adding more decision trees and rules.

Some other works use *model simplification* as a post-hoc explainability approach. For instance, [56] presents a simple distillation method called *Interpretable Mimic Learning* to extract an interpretable model

by means of gradient boosting trees. In the same direction, the authors in [135] propose a hierarchical partitioning of the feature space that reveals the iterative rejection of unlikely class labels, until association is predicted. In addition, several works addressed the distillation of knowledge from an ensemble of models into a single model [80, 289, 290] .

Given the fact that the simplification of multi-layer neural networks is more complex as the number of layers increases, explaining these models by *feature relevance* methods has become progressively more popular. One of the representative works in this area is [60], which presents a method to decompose the network classification decision into contributions of its input elements. They consider each neuron as an object that can be decomposed and expanded then aggregate and back-propagate these decompositions through the network, resulting in a *deep* Taylor decomposition. In the same direction, the authors in [110] proposed DeepLIFT, an approach for computing importance scores in a multi-layer neural network. Their method compares the activation of a neuron to the reference activation and assigns the score according to the difference.

On the other hand, some works try to verify the theoretical soundness of current explainability methods. For example, the authors in [262], bring up a fundamental problem of most *feature relevance* techniques, designed for multi-layer networks. They showed that two axioms that such techniques ought to fulfill namely, *sensitivity* and *implementation invariance*, are violated in practice by most approaches. Following these axioms, the authors of [262] created *integrated gradients*, a new *feature relevance* method proven to meet the aforementioned axioms. Similarly, the authors in [61] analyzed the correctness of current *feature relevance* explanation approaches designed for Deep Neural Networks, e,g., DeConvNet, Guided BackProp and LRP, on simple linear neural networks. Their analysis showed that these methods do not produce the theoretically correct explanation and presented two new explanation methods *PatternNet* and *PatternAttribution* that are more theoretically sound for both, simple and deep neural networks.

4.3.2. Convolutional Neural Networks

Currently, CNNs constitute the state-of-art models in all fundamental computer vision tasks, from image classification and object detection to instance segmentation. Typically, these models are built as a sequence of convolutional layers and pooling layers to automatically learn increasingly higher level features. At the end of the sequence, one or multiple fully connected layers are used to map the output features map into scores. This structure entails extremely complex internal relations that are very difficult to explain. Fortunately, the road to explainability for CNNs is easier than for other types of models, as the human cognitive skills favors the understanding of visual data.

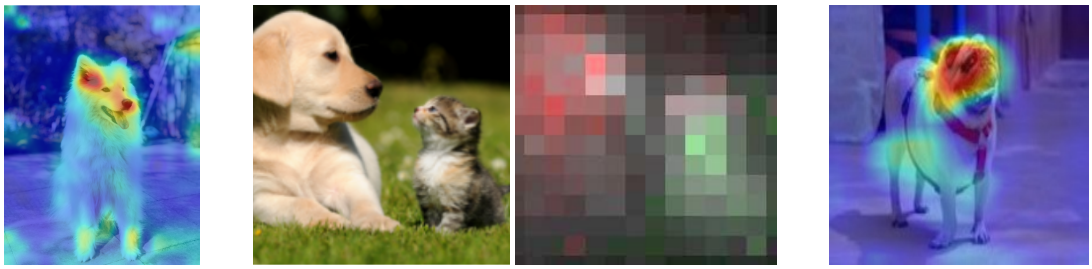
Existing works that aim at understanding what CNNs learn can be divided into two broad categories: 1) those that try to understand the decision process by mapping back the output in the input space to see which parts of the input were discriminative for the output; and 2) those that try to delve inside the network and interpret how the intermediate layers see the external world, not necessarily related to any specific input, but in general.

One of the seminal works in the first category was [291]. When an input image runs feed-forward through a CNN, each layer outputs a number of feature maps with strong and soft activations. The authors in [291] used Deconvnet, a network designed previously by the same authors [142] that, when fed with a feature map from a selected layer, reconstructs the maximum activations. These reconstructions can give an idea about the parts of the image that produced that effect. To visualize these strongest activations in the input image, the same authors used the occlusion sensitivity method to generate a saliency map [136], which consists of iteratively forwarding the same image through the network occluding a different region at a time.

To improve the quality of the mapping on the input space, several subsequent papers proposed simplifying both the CNN architecture and the visualization method. In particular, [96] included a global average pooling layer between the last convolutional layer of the CNN and the fully-connected layer that predicts the object class. With this simple architectural modification of the CNN, the authors built a class

activation map that helps identify the image regions that were particularly important for a specific object class by projecting back the weights of the output layer on the convolutional feature maps. Later, in [143], the authors showed that max-pooling layers can be used to replace convolutional layers with a large stride without loss in accuracy on several image recognition benchmarks. They obtained a cleaner visualization than Deconvnet by using a guided backpropagation method.

To increase the interpretability of classical CNNs, the authors in [113] used a loss for each filter in high level convolutional layers to force each filter to learn very specific object components. The obtained activation patterns are much more interpretable for their exclusiveness with respect to the different labels to be predicted. The authors in [72] proposed visualizing the contribution to the prediction of each single pixel of the input image in the form of a heatmap. They used a Layer-wise Relevance Propagation (LRP) technique, which relies on a Taylor series close to the prediction point rather than partial derivatives at the prediction point itself. To further improve the quality of the visualization, attribution methods such as heatmaps, saliency maps or class activation methods (*GradCAM* [292]) are used (see Figure 7). In particular, the authors in [292] proposed a Gradient-weighted Class Activation Mapping (Grad-CAM), which uses the gradients of any target concept, flowing into the final convolutional layer to produce a coarse localization map, highlighting the important regions in the image for predicting the concept.



(a) Heatmap [168] (b) Attribution [293] (c) Grad-CAM [292]

Figure 7: Examples of rendering for different XAI visualization techniques on images.

In addition to the aforementioned *feature relevance* and *visual* explanation methods, some works proposed generating *text explanations* of the visual content of the image. For example, the authors in [91] combined a CNN feature extractor with an RNN attention model to automatically learn to describe the content of images. In the same line, [278] presented a three-level attention model to perform a fine-grained classification task. The general model is a pipeline that integrates three types of attention: the object level attention model proposes candidate image regions or patches from the input image, the part-level attention model filters out non-relevant patches to a certain object, and the last attention model localizes discriminative patches. In the task of video captioning, the authors in [111] use a CNN model combined with a bi-directional LSTM model as encoder to extract video features and then feed these features to an LSTM decoder to generate textual descriptions.

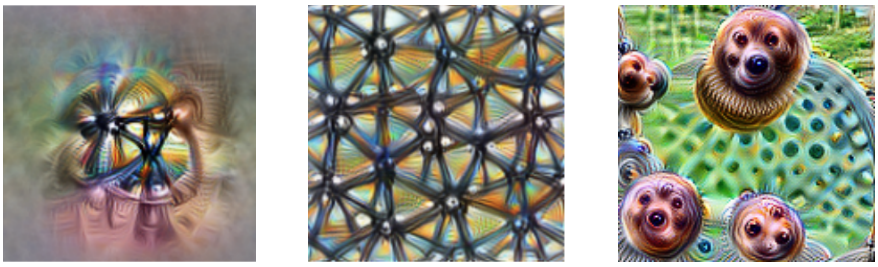
One of the seminal works in the second category is [137]. In order to analyse the visual information contained inside the CNN, the authors proposed a general framework that reconstruct an image from the CNN internal representations and showed that several layers retain photographically accurate information about the image, with different degrees of geometric and photometric invariance. To visualize the notion of a class captured by a CNN, the same authors created an image that maximizes the class score based on computing the gradient of the class score with respect to the input image [272]. In the same direction, the authors in [268] introduced a Deep Generator Network (DGN) that generates the most representative image for a given output neuron in a CNN.

For quantifying the interpretability of the latent representations of CNNs, the authors in [125] used a different approach called network dissection. They run a large number of images through a CNN and then analyze the top activated images by considering each unit as a concept detector to further evaluate each

unit for semantic segmentation. This paper also examines the effects of classical training techniques on the interpretability of the learned model.

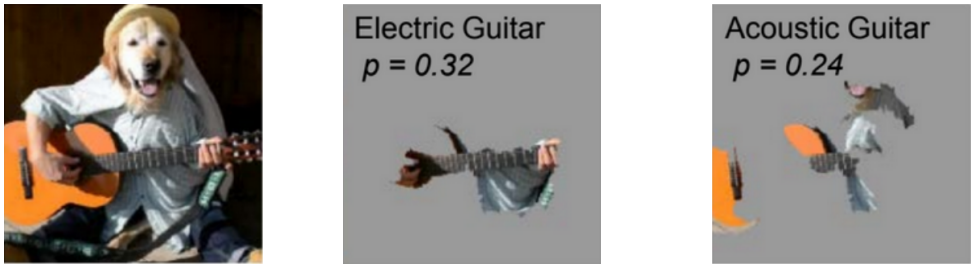
Although many of the techniques examined above utilize *local explanations* to achieve an overall explanation of a CNN model, others explicitly focus on building global explanations based on locally found prototypes. In [263, 294], the authors empirically showed how *local explanations* in deep networks are strongly dominated by their lower level features. They demonstrated that deep architectures provide strong priors that prevent the altering of how these low-level representations are captured. All in all, *visualization* mixed with *feature relevance* methods are arguably the most adopted approach to explainability in CNNs.

Instead of using one single interpretability technique, the framework proposed in [295] combines several methods to provide much more information about the network. For example, combining feature visualization (*what is a neuron looking for?*) with attribution (*how does it affect the output?*) allows exploring how the network decides between labels. This visual interpretability interface displays different blocks such as feature visualization and attribution depending on the visualization goal. This interface can be thought of as a union of individual elements that belong to layers (input, hidden, output), atoms (a neuron, channel, spatial or neuron group), content (activations – the amount a neuron fires, attribution – which classes a spatial position most contributes to, which tends to be more meaningful in later layers), and presentation (information visualization, feature visualization). Figure 8 shows some examples. Attribution methods normally rely on pixel association, displaying what part of an input example is responsible for the network activating in a particular way [293].



(a) Neuron (b) Channel (c) Layer

Figure 8: Feature visualization at different levels of a certain network [293].



(a) Original image (b) Explaining *electric guitar* (c) Explaining *acoustic guitar*

Figure 9: Examples of explanation when using LIME on images [71].

A much simpler approach to all the previously cited methods was proposed in LIME framework [71], as was described in Subsection 4.1 LIME perturbs the input and sees how the predictions change. In image classification, LIME creates a set of perturbed instances by dividing the input image into interpretable components (contiguous *superpixels*), and runs each perturbed instance through the model

to get a probability. A simple linear model learns on this data set, which is locally weighted. At the end of the process, LIME presents the superpixels with highest positive weights as an explanation (see Figure 9).

A completely different explainability approach is proposed in adversarial detection. To understand model failures in detecting adversarial examples, the authors in [264] apply the k-nearest neighbors algorithm on the representations of the data learned by each layer of the CNN. A test input image is considered as adversarial if its representations are far from the representations of the training images.

4.3.3. Recurrent Neural Networks

As occurs with CNNs in the visual domain, RNNs have lately been used extensively for predictive problems defined over inherently sequential data, with a notable presence in natural language processing and time series analysis. These types of data exhibit long-term dependencies that are complex to be captured by a ML model. RNNs are able to retrieve such time-dependent relationships by formulating the retention of knowledge in the neuron as another parametric characteristic that can be learned from data.

Few contributions have been made for explaining RNN models. These studies can be divided into two groups: 1) explainability by understanding what a RNN model has learned (mainly via *feature relevance* methods); and 2) explainability by modifying RNN architectures to provide insights about the decisions they make (*local explanations*).

In the first group, the authors in [280] extend the usage of LRP to RNNs. They propose a specific propagation rule that works with multiplicative connections as those in LSTMs (Long Short Term Memory) units and GRUs (Gated Recurrent Units). The authors in [281] propose a visualization technique based on finite horizon n-grams that discriminates interpretable cells within LSTM and GRU networks. Following the premise of not altering the architecture, [296] extends the interpretable mimic learning distillation method used for CNN models to LSTM networks, so that interpretable features are learned by fitting Gradient Boosting Trees to the trained LSTM network under focus.

Aside from the approaches that do not change the inner workings of the RNNs, [285] presents RETAIN (REverse Time AttentIoN) model, which detects influential past patterns by means of a two-level neural attention model. To create an interpretable RNN, the authors in [283] propose an RNN based on SISTA (Sequential Iterative Soft-Thresholding Algorithm) that models a sequence of correlated observations with a sequence of sparse latent vectors, making its weights interpretable as the parameters of a principled statistical model. Finally, [284] constructs a combination of an HMM (Hidden Markov Model) and an RNN, so that the overall model approach harnesses the interpretability of the HMM and the accuracy of the RNN model.

4.3.4. Hybrid Transparent and Black-box Methods

The use of background knowledge in the form of logical statements or constraints in Knowledge Bases (KBs) has shown to not only improve explainability but also performance with respect to purely data-driven approaches [297, 298, 299]. A positive side effect shown is that this hybrid approach provides robustness to the learning system when errors are present in the training data labels. Other approaches have shown to be able to jointly learn and reason with both symbolic and sub-symbolic representations and inference. The interesting aspect is that this blend allows for expressive probabilistic-logical reasoning in an end-to-end fashion [300]. A successful use case is on dietary recommendations, where explanations are extracted from the reasoning behind (non-deep but KB-based) models [301].

Future data fusion approaches may thus consider endowing DL models with explainability by externalizing other domain information sources. Deep formulation of classical ML models has been done, e.g. in Deep Kalman filters (DKFs) [302], Deep Variational Bayes Filters (DVBFs) [303], Structural Variational Autoencoders (SVAE) [304], or conditional random fields as RNNs [305]. These approaches provide deep models with the interpretability inherent to probabilistic graphical models. For instance, SVAE combines probabilistic graphical models in the embedding space with neural networks to enhance the interpretability of DKFs. A particular example of classical ML model enhanced with its DL counterpart is

Deep Nearest Neighbors DkNN [264], where the neighbors constitute human-interpretable explanations of predictions. The intuition is based on the rationalization of a DNN prediction based on evidence. This evidence consists of a characterization of confidence termed *credibility* that spans the hierarchy of representations within a DNN, that must be supported by the training data [264].

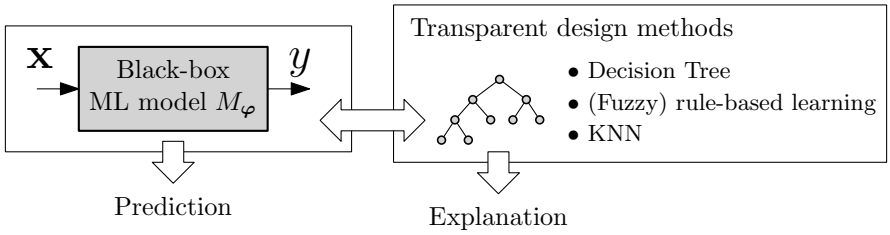


Figure 10: Pictorial representation of a hybrid model. A neural network considered as a black-box can be explained by associating it to a more interpretable model such as a Decision Tree [306], a (fuzzy) rule-based system [19] or KNN [264].

A different perspective on hybrid XAI models consists of enriching black-box models knowledge with that one of transparent ones, as proposed in [24] and further refined in [169] and [307]. In particular, this can be done by constraining the neural network thanks to a semantic KB and bias-prone concepts [169], or by stacking ensembles jointly encompassing white- and black-box models [307].

Other examples of hybrid symbolic and sub-symbolic methods where a knowledge-base tool or graph-perspective enhances the neural (e.g., language [308]) model are in [309, 310]. In reinforcement learning, very few examples of symbolic (graphical [311] or relational [75, 312]) hybrid models exist, while in recommendation systems, for instance, explainable autoencoders are proposed [313]. A specific transformer architecture symbolic visualization method (applied to music) pictorially shows how soft-max attention works [314]. By visualizing self-reference, i.e., the last layer of attention weights, arcs show which notes in the past are informing the future and how attention is skip over less relevant sections. Transformers can also help explain image captions visually [315].

Another hybrid approach consists of mapping an uninterpretable black-box system to a white-box *twin* that is more interpretable. For example, an opaque neural network can be combined with a transparent Case Based Reasoning (CBR) system [316, 317]. In [318], the DNN and the CBR (in this case a kNN) are paired in order to improve interpretability while keeping the same accuracy. The *explanation by example* consists of analyzing the feature weights of the DNN which are then used in the CBR, in order to retrieve nearest-neighbor cases to explain the DNN's prediction.

4.4. Alternative Taxonomy of Post-hoc Explainability Techniques for Deep Learning

DL is the model family where most research has been concentrated in recent times and they have become central for most of the recent literature on XAI. While the division between model-agnostic and model-specific is the most common distinction made, the community has not only relied on this criteria to classify XAI methods. For instance, some model-agnostic methods such as *SHAP* [224] are widely used to explain DL models. That is why several XAI methods can be easily categorized in different taxonomy branches depending on the angle the method is looked at. An example is LIME which can also be used over CNNs, despite not being exclusive to deal with images. Searching within the alternative DL taxonomy shows us that LIME can explicitly be used for *Explaining a Deep Network Processing*, as a kind of *Linear Proxy Model*. Another type of classification is indeed proposed in [13] with a segmentation based on 3 categories. The first category groups methods explaining the processing of data by the network, thus answering to the question *"why does this particular input leads to this particular output?"*. The second one concerns methods explaining the representation of data inside the network, i.e., answering to the question *"what information does the network contain?"*. The third approach concerns

models specifically designed to simplify the interpretation of their own behavior. Such a multiplicity of classification possibilities leads to different ways of constructing XAI taxonomies.

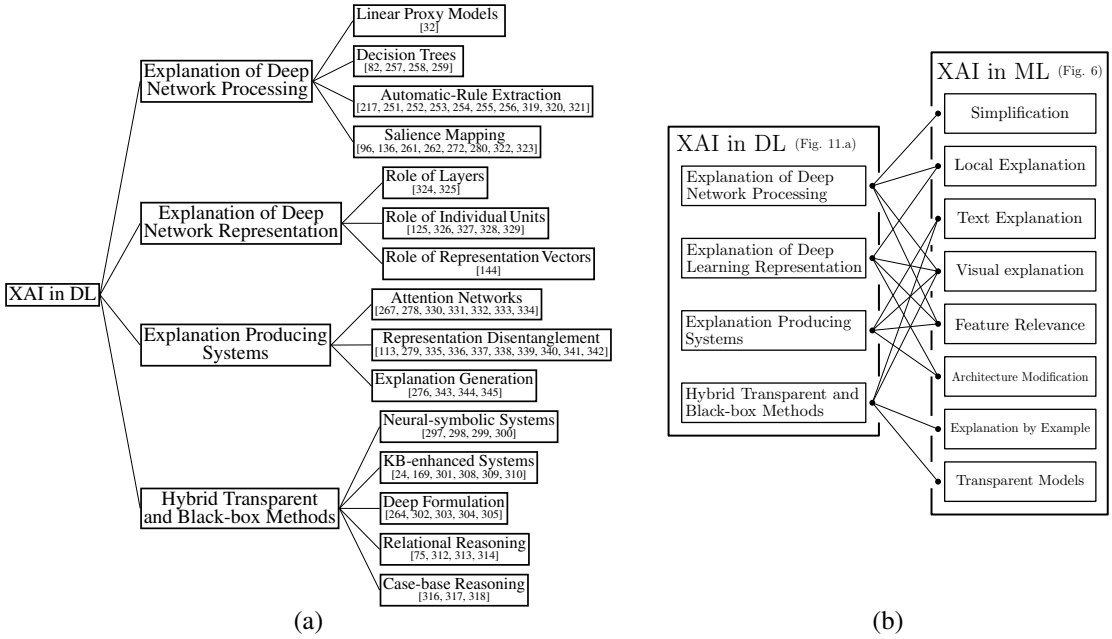


Figure 11: (a) Alternative Deep Learning specific taxonomy extended from the categorization from [13]; and (b) its connection to the taxonomy in Figure 6.

Figure 11 shows the alternative Deep Learning taxonomy inferred from [13]. From the latter, it can be deduced the complementarity and overlapping of this taxonomy to Figure 6 as:

- Some methods [272, 280] classified in distinct categories (namely *feature relevance for CNN* and *feature relevance for RNN*) in Figure 6 are included in a single category (*Explanation of Deep Network Processing with Salience Mapping*) when considering the classification from [13].
- Some methods [82, 144] are classified on a single category (*Explanation by simplification for Multi-Layer Neural Network*) in Figure 6 while being in 2 different categories (namely, *Explanation of Deep Network Processing with Decision Trees* and *Explanation of Deep Network Representation with the Role of Representation Vectors*) in [13], as shown in Figure 11.

A classification based on explanations of model processing and explanations of model representation is relevant, as it leads to a differentiation between the execution trace of the model and its internal data structure. This means that depending of the failure reasons of a complex model, it would be possible to pick-up the right XAI method according to the information needed: the execution trace or the data structure. This idea is analogous to testing and debugging methods used in regular programming paradigms [346].

5. XAI: Opportunities, Challenges and Future Research Needs

We now capitalize on the performed literature review to put forward a critique of the achievements, trends and challenges that are still to be addressed in the field of explainability of ML and data fusion models. Actually our discussion on the advances taken so far in this field has already anticipated some of these challenges. In this section we revisit them and explore new research opportunities for XAI, identifying possible research paths that can be followed to address them effectively in years to come:

- When introducing the overview in Section 1 we already mentioned the existence of a tradeoff between model interpretability and performance, in the sense that making a ML model more understandable could eventually degrade the quality of its produced decisions. In Subsection 5.1 we will stress on the potential of XAI developments to effectively achieve an optimal balance between the interpretability and performance of ML models.
- In Subsection 2.2 we stressed on the imperative need for reaching a consensus on *what* explainability entails within the AI realm. Reasons for pursuing explainability are also assorted and, under our own assessment of the literature so far, not unambiguously mentioned throughout related works. In Subsection 5.2 we will further delve into this important issue.
- Given its notable prevalence in the XAI literature, Subsections 4.3 and 4.4 revolved on the explainability of Deep Learning models, examining advances reported so far around a specific bibliographic taxonomy. We go in this same direction with Subsection 5.3, which exposes several challenges that hold in regards to the explainability of this family of models.
- Finally, we close up this prospective discussion with Subsections 5.4 to 5.8, which place on the table several research niches that despite its connection to model explainability, remain insufficiently studied by the community.

Before delving into these identified challenges, it is important to bear in mind that this prospective section is complemented by Section 6, which enumerates research needs and open questions related to XAI within a broader context: the need for responsible AI.

5.1. On the Tradeoff between Interpretability and Performance

The matter of interpretability versus performance is one that repeats itself through time, but as any other big statement, has its surroundings filled with myths and misconceptions.

As perfectly stated in [347], it is not necessarily true that models that are more complex are inherently more accurate. This statement is false in cases in which the data is well structured and features at our disposal are of great quality and value. This case is somewhat common in some industry environments, since features being analyzed are constrained within very controlled physical problems, in which all of the features are highly correlated, and not much of the possible landscape of values can be explored in the data [348]. What can be hold as true, is that more complex models enjoy much more flexibility than their simpler counterparts, allowing for more complex functions to be approximated. Now, returning to the statement *"models that are more complex are more accurate"*, given the premise that the function to be approximated entails certain complexity, that the data available for study is greatly widespread among the world of suitable values for each variable and that there is enough data to harness a complex model, the statement presents itself as a true statement. It is in this situation that the trade-off between performance and interpretability can be observed. It should be noted that the attempt at solving problems that do not respect the aforementioned premises will fall on the trap of attempting to solve a problem that does not provide enough data diversity (variance). Hence, the added complexity of the model will only fight against the task of accurately solving the problem.

In this path toward performance, when the performance comes hand in hand with complexity, interpretability encounters itself on a downwards slope that until now appeared unavoidable. However, the apparition of more sophisticated methods for explainability could invert or at least cancel that slope. Figure 12 shows a tentative representation inspired by previous works [7], in which XAI shows its power to improve the common trade-off between model interpretability and performance. Another aspect worth mentioning at this point due to its close link to model interpretability and performance is the *approximation dilemma*: explanations made for a ML model must be made drastic and approximate enough to match the requirements of the audience for which they are sought, ensuring that explanations are representative of the studied model and do not oversimplify its essential features.

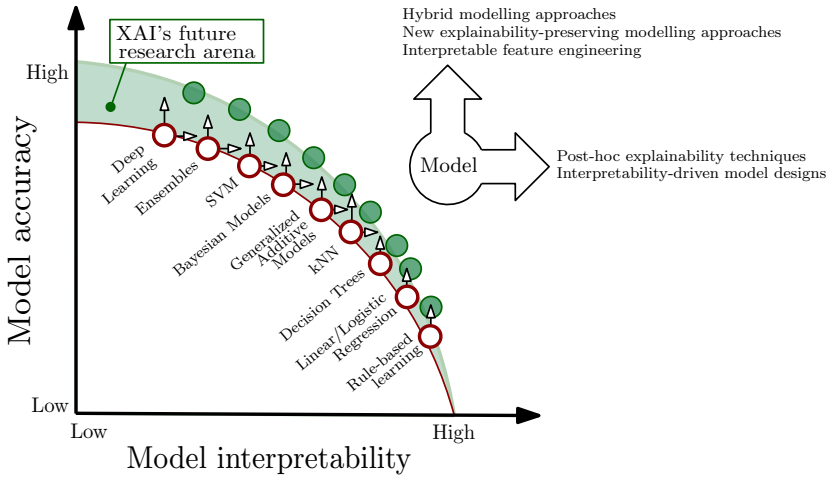


Figure 12: Trade-off between model interpretability and performance, and a representation of the area of improvement where the potential of XAI techniques and tools resides.

5.2. On the Concept and Metrics

The literature clearly asks for an unified concept of explainability. In order for the field to thrive, it is imperative to place a common ground upon which the community is enabled to contribute new techniques and methods. A common concept must convey the needs expressed in the field. It should propose a common structure for every XAI system. This paper attempted a new proposition of a concept of explainability that is built upon that from Gunning [7]. In that proposition and the following strokes to complete it (Subsection 2.2), explainability is defined as the ability a model has to make its functioning clearer to an audience. To address it, post-hoc type methods exist. The concept portrayed in this survey might not be complete but as it stands, allows for a first common ground and reference point to sustain a profitable discussion in this matter. It is paramount that the field of XAI reaches an agreement in this respect combining the shattered efforts of a widespread field behind the same banner.

Another key feature needed to relate a certain model to this concrete concept is the existence of a metric. A metric, or group of them should allow for a meaningful comparison of how well a model fits the definition of explainable. Without such tool, any claim in this respect dilutes among the literature, not providing a solid ground on which to stand. These metrics, as the classic ones (accuracy, F1, sensitivity...), should express how well the model performs in a certain aspect of explainability. Some attempts have been done recently around the measurement of XAI, as reviewed thoroughly in [349, 350]. In general, XAI measurements should evaluate the goodness, usefulness and satisfaction of explanations, the improvement of the mental model of the audience induced by model explanations, and the impact of explanations on the performance of the model and on the trust and reliance of the audience. Measurement techniques surveyed in [349] and [350] (e.g., goodness checklist, explanation satisfaction scale, elicitation methods for mental models, computational measures for explainer fidelity, explanation trustworthiness and model reliability) seem to be a good push in the direction of evaluating XAI techniques. Unfortunately, conclusions drawn from these overviews are aligned with our prospects on the field: more quantifiable, general XAI metrics are really needed to support the existing measurement procedures and tools proposed by the community.

This survey does not tackle the problem of designing such a suite of metrics, since such a task should be approached by the community as a whole prior acceptance of the broader concept of explainability, which on the other hand, is one of the aims of the current work. Nevertheless, we advocate for further efforts towards new proposals to evaluate the performance of XAI techniques, as well as comparison methodologies among XAI approaches that allow contrasting them quantitatively under different

application context, models and purposes.

5.3. Challenges to achieve Explainable Deep Learning

While many efforts are currently being made in the area of XAI, there are still many challenges to be faced before being able to obtain explainability in DL models. First, as explained in Subsection 2.2, there is a lack of agreement on the vocabulary and the different definitions surrounding XAI. As an example, we often see the terms *feature importance* and *feature relevance* referring to the same concept. This is even more obvious for visualization methods, where there is absolutely no consistency behind what is known as saliency maps, salient masks, heatmaps, neuron activations, attribution, and other approaches alike. As XAI is a relatively young field, the community does not have a standardized terminology yet.

As it has been commented in Subsection 5.1, there is a trade-off between interpretability and accuracy [13], i.e., between the simplicity of the information given by the system on its internal functioning, and the exhaustiveness of this description. Whether the observer is an expert in the field, a policy-maker or a user without machine learning knowledge, intelligibility does not have to be at the same level in order to provide the *audience* an understanding [6]. This is one of the reasons why, as mentioned above, a challenge in XAI is establishing objective metrics on what constitutes a good explanation. A possibility to reduce this subjectivity is taking inspiration from experiments on human psychology, sociology or cognitive sciences to create objectively convincing explanations. Relevant findings to be considered when creating an explainable AI model are highlighted in [12]: First, explanations are better when *constrictive*, meaning that a prerequisite for a good explanation is that it does not only indicate why the model made a decision X, but also why it made decision X rather than decision Y. It is also explained that probabilities are not as important as causal links in order to provide a satisfying explanation. Considering that black box models tend to process data in a quantitative manner, it would be necessary to translate the probabilistic results into qualitative notions containing causal links. In addition, they state that explanations are *selective*, meaning that focusing solely on the main causes of a decision-making process is sufficient. It was also shown that the use of counterfactual explanations can help the user to understand the decision of a model [40, 42, 351].

Combining connectionist and symbolic paradigms seems a favourable way to address this challenge [169, 299, 312, 352, 353]. On one hand, connectionist methods are more precise but opaque. On the other hand, symbolic methods are popularly considered less efficient, while they offer a greater explainability thus respecting the conditions mentioned above:

- The ability to refer to established reasoning rules allows symbolic methods to be constrictive.
- The use of a KB formalized e.g. by an ontology can allow data to be processed directly in a qualitative way.
- Being selective is less straightforward for connectionist models than for symbolic ones.

Recalling that a good explanation needs to influence the mental model of the user, i.e. the representation of the external reality using, among other things, symbols, it seems obvious that the use of the symbolic learning paradigm is appropriate to produce an explanation. Therefore, neural-symbolic interpretability could provide convincing explanations while keeping or improving generic performance [297].

As stated in [24], a truly explainable model should not leave explanation generation to the users as different explanations may be deduced depending on their background knowledge. Having a semantic representation of the knowledge can help a model to have the ability to produce explanations (e.g., in natural language [169]) combining common sense reasoning and human-understandable features.

Furthermore, until an objective metric has been adopted, it appears necessary to make an effort to rigorously formalize evaluation methods. One way may be drawing inspiration from the social sciences, e.g., by being consistent when choosing the evaluation questions and the population sample used [354].

A final challenge XAI methods for DL need to address is providing explanations that are accessible for society, policy makers and the law as a whole. In particular, conveying explanations that require non-technical expertise will be paramount to both handle ambiguities, and to develop the social right to the (not-yet available) right for explanation in the EU General Data Protection Regulation (GDPR) [355].

5.4. Explanations for AI Security: XAI and Adversarial Machine Learning

Nothing has been said about confidentiality concerns linked to XAI. One of the last surveys very briefly introduced the idea of algorithm property and trade secrets [14]. However, not much attention has been payed to these concepts. If *confidential* is the property that makes something *secret*, in the AI context many aspects involved in a model may hold this property. For example, imagine a model that some company has developed through many years of research in a specific field. The knowledge synthesized in the model built might be considered to be confidential, and it may be compromised even by providing only input and output access [356]. The latter shows that, under minimal assumptions, *data model functionality stealing* is possible. An approach that has served to make DL models more robust against intellectual property exposure based on a sequence of non accessible queries is in [357]. This recent work exposes the need for further research toward the development of XAI tools capable of explaining ML models while keeping the model's confidentiality in mind.

Ideally, XAI should be able to explain the knowledge within an AI model and it should be able to reason about what the model acts upon. However, the information revealed by XAI techniques can be used both to generate more effective attacks in adversarial contexts aimed at confusing the model, at the same time as to develop techniques to better protect against private content exposure by using such information. Adversarial attacks [358] try to manipulate a ML algorithm after learning what is the specific information that should be fed to the system so as to lead it to a specific output. For instance, regarding a supervised ML classification model, adversarial attacks try to discover the minimum changes that should be applied to the input data in order to cause a different classification. This has happened regarding computer vision systems of autonomous vehicles; a minimal change in a stop signal, imperceptible to the human eye, led vehicles to detect it as a 45 mph signal [359]. For the particular case of DL models, available solutions such as Cleverhans [360] seek to detect adversarial vulnerabilities, and provide different approaches to harden the model against them. Other examples include AlfaSVMLib [361] for SVM models, and AdversarialLib [362] for evasion attacks. There are even available solutions for unsupervised ML, like clustering algorithms [363].

While XAI techniques can be used to furnish more effective adversarial attacks or to reveal confidential aspects of the model itself, some recent contributions have capitalized on the possibilities of Generative Adversarial Networks (GANs [364]), Variational Autoencoders [365] and other generative models towards explaining data-based decisions. Once trained, generative models can generate instances of what they have learned based on a noise input vector that can be interpreted as a latent representation of the data at hand. By manipulating this latent representation and examining its impact on the output of the generative model, it is possible to draw insights and discover specific patterns related to the class to be predicted. This generative framework has been adopted by several recent studies [366, 367] mainly as an attribution method to relate a particular output of a Deep Learning model to their input variables. Another interesting research direction is the use of generative models for the creation of counterfactuals, i.e., modifications to the input data that could eventually alter the original prediction of the model [368]. Counterfactual prototypes help the user understand the performance boundaries of the model under consideration for his/her improved trust and informed criticism. In light of this recent trend, we definitely believe that there is road ahead for generative ML models to take their part in scenarios demanding understandable machine decisions.

5.5. XAI and Output Confidence

Safety issues have also been studied in regards to processes that depend on the output of AI models, such as vehicular perception and self-driving in autonomous vehicles, automated surgery, data-based

support for medical diagnosis, insurance risk assessment and cyber-physical systems in manufacturing, among others [369]. In all these scenarios erroneous model outputs can lead to harmful consequences, which has yielded comprehensive regulatory efforts aimed at ensuring that no decision is made solely on the basis of data processing [3].

In parallel, research has been conducted towards minimizing both risk and uncertainty of harms derived from decisions made on the output of a ML model. As a result, many techniques have been reported to reduce such a risk, among which we pause at the evaluation of the model's output confidence to decide upon. In this case, the inspection of the share of epistemic uncertainty (namely, the uncertainty due to lack of knowledge) of the input data and its correspondence with the model's output confidence can inform the user and eventually trigger his/her rejection of the model's output [370, 371]. To this end, explaining via XAI techniques which region of the input data the model is focused on when producing a given output can discriminate possible sources of epistemic uncertainty within the input domain.

5.6. XAI, Rationale Explanation, and Critical Data Studies

When shifting the focus to the research practices seen in Data Science, it has been noted that reproducibility is stringently subject not only to the mere sharing of data, models and results to the community, but also to the availability of information about the full discourse around data collection, understanding, assumptions held and insights drawn from model construction and results' analyses [372]. In other words, in order to transform data into a valuable actionable asset, individuals must engage in collaborative sense-making by sharing the context producing their findings, wherein context refers to sets of narrative stories around how data were processed, cleaned, modeled and analyzed. In this discourse we find also an interesting space for the adoption of XAI techniques due to their powerful ability to describe black-box models in an understandable, hence conveyable fashion towards colleagues from Social Science, Politics, Humanities and Legal fields.

XAI can effectively ease the process of explaining the reasons why a model reached a decision in an accessible way to non-expert users, i.e. the *rationale explanation*. This confluence of multi-disciplinary teams in projects related to Data Science and the search for methodologies to make them appraise the ethical implications of their data-based choices has been lately coined as Critical Data studies [373]. It is in this field where XAI can significantly boost the exchange of information among heterogeneous audiences about the knowledge learned by models.

5.7. XAI and Theory-guided Data Science

We envision an exciting synergy between the XAI realm and *Theory-guided Data Science*, a paradigm exposed in [374] that merges both Data Science and the classic theoretical principles underlying the application/context where data are produced. The rationale behind this rising paradigm is the need for data-based models to generate knowledge that is the prior knowledge brought by the field in which it operates. This means that the model type should be chosen according to the type of relations we intend to encounter. The structure should also follow what is previously known. Similarly, the training approach should not allow for the optimization process to enter regions that are not plausible. Accordingly, regularization terms should stand the prior premises of the field, avoiding the elimination of badly represented true relations for spurious and deceptive false relations. Finally, the output of the model should inform about everything the model has come to learn, allowing to reason and merge the new knowledge with what was already known in the field.

Many examples of the implementation of this approach are currently available with promising results. The studies in [375]-[382] were carried out in diverse fields, showcasing the potential of this new paradigm for data science. Above all, it is relevant to notice the resemblance that all concepts and requirements of Theory-guided Data Science share with XAI. All the additions presented in [374] push toward techniques that would eventually render a model explainable, and furthermore, knowledge consistent. The concept of *knowledge from the beginning*, central to Theory-guided Data Science, must also consider how

the knowledge captured by a model should be explained for assessing its compliance with theoretical principles known beforehand. This, again, opens a magnificent window of opportunity for XAI.

5.8. Guidelines for ensuring Interpretable AI Models

Recent surveys have emphasized on the multidisciplinary, inclusive nature of the process of making an AI-based model interpretable. Along this process, it is of utmost importance to scrutinize and take into proper account the interests, demands and requirements of all stakeholders interacting with the system to be explained, from the designers of the system to the decision makers consuming its produced outputs and users undergoing the consequences of decisions made therefrom.

Given the confluence of multiple criteria and the need for having the human in the loop, some attempts at establishing the procedural guidelines to implement and explain AI systems have been recently contributed. Among them, we pause at the thorough study in [383], which suggests that the incorporation and consideration of explainability in practical AI design and deployment workflows should comprise four major methodological steps:

1. Contextual factors, potential impacts and domain-specific needs must be taken into account when devising an approach to interpretability: These include a thorough understanding of the purpose for which the AI model is built, the complexity of explanations that are required by the audience, and the performance and interpretability levels of existing technology, models and methods. The latter pose a reference point for the AI system to be deployed in lieu thereof.

2. Interpretable techniques should be preferred when possible: when considering explainability in the development of an AI system, the decision of which XAI approach should be chosen should gauge domain-specific risks and needs, the available data resources and existing domain knowledge, and the suitability of the ML model to meet the requirements of the computational task to be addressed. It is in the confluence of these three design drivers where the guidelines postulated in [383] (and other studies in this same line of thinking [384]) recommend first the consideration of standard interpretable models rather than sophisticated yet opaque modeling methods. In practice, the aforementioned aspects (contextual factors, impacts and domain-specific needs) can make transparent models preferable over complex modeling alternatives whose interpretability require the application of post-hoc XAI techniques. By contrast, black-box models such as those reviewed in this work (namely, support vector machines, ensemble methods and neural networks) should be selected only when their superior modeling capabilities fit best the characteristics of the problem at hand.

3. If a black-box model has been chosen, the third guideline establishes that ethics-, fairness- and safety-related impacts should be weighed. Specifically, responsibility in the design and implementation of the AI system should be ensured by checking whether such identified impacts can be mitigated and counteracted by supplementing the system with XAI tools that provide the level of explainability required by the domain in which it is deployed. To this end, the third guideline suggests 1) a detailed articulation, examination and evaluation of the applicable explanatory strategies, 2) the analysis of whether the coverage and scope of the available explanatory approaches match the requirements of the domain and application context where the model is to be deployed; and 3) the formulation of an interpretability action plan that sets forth the explanation delivery strategy, including a detailed time frame for the execution of the plan, and a clearance of the roles and responsibilities of the team involved in the workflow.

4. Finally, the fourth guideline encourages to rethink interpretability in terms of the cognitive skills, capacities and limitations of the individual human. This is an important question on which studies on measures of explainability are intensively revolving by considering human mental models, the accessibility of the audience to vocabularies of explanatory outcomes, and other means to involve the expertise of the audience into the decision of what explanations should provide.

We foresee that the set of guidelines proposed in [383] and summarized above will be complemented and enriched further by future methodological studies, ultimately heading to a more *responsible* use of AI. Methodological principles ensure that the purpose for which explainability is pursued is met by bringing the manifold of requirements of all participants into the process, along with other universal aspects of equal relevance such as no discrimination, sustainability, privacy or accountability. A challenge remains in harnessing the potential of XAI to realize a *Responsible AI*, as we discuss in the next section.

6. Toward Responsible AI: Principles of Artificial Intelligence, Fairness, Privacy and Data Fusion

Over the years many organizations, both private and public, have published guidelines to indicate how AI should be developed and used. These guidelines are commonly referred to as AI *principles*, and they tackle issues related to potential AI threats to both individuals and to the society as a whole. This section presents some of the most important and widely recognized principles in order to link XAI – which normally appears inside its own principle – to all of them. Should a responsible implementation and use of AI models be sought in practice, it is our firm claim that XAI does not suffice on its own. Other important principles of Artificial Intelligence such as privacy and fairness must be carefully addressed in practice. In the following sections we elaborate on the concept of Responsible AI, along with the implications of XAI and data fusion in the fulfillment of its postulated principles.

6.1. Principles of Artificial Intelligence

A recent review of some of the main AI principles published since 2016 appears in [385]. In this work, the authors show a visual framework where different organizations are classified according to the following parameters:

- Nature, which could be private sector, government, inter-governmental organization, civil society or multistakeholder.
- Content of the principles: eight possible principles such as privacy, explainability, or fairness. They also consider the coverage that the document grants for each of the considered principles.
- Target audience: to whom the principles are aimed. They are normally for the organization that developed them, but they could also be destined for another audience (see Figure 2).
- Whether or not they are rooted on the International Human Rights, as well as whether they explicitly talk about them.

For instance, [386] is an illustrative example of a document of AI principles for the purpose of this overview, since it accounts for some of the most common principles, and deals explicitly with explainability. Here, the authors propose five principles mainly to guide the development of AI within their company, while also indicating that they could also be used within other organizations and businesses.

The authors of those principles aim to develop AI in a way that it directly reinforces inclusion, gives equal opportunities for everyone, and contributes to the common good. To this end, the following aspects should be considered:

- The outputs after using AI systems should not lead to any kind of discrimination against individuals or collectives in relation to race, religion, gender, sexual orientation, disability, ethnic, origin or any other personal condition. Thus, a fundamental criteria to consider while optimizing the results of an AI system is not only their outputs in terms of error optimization, but also how the system deals with those groups. This defines the principle of *Fair AI*.

- People should always know when they are communicating with a person, and when they are communicating with an AI system. People should also be aware if their personal information is being used by the AI system and for what purpose. It is crucial to ensure a certain level of understanding about the decisions taken by an AI system. This can be achieved through the usage of XAI techniques. It is important that the generated explanations consider the profile of the user that will receive those explanations (the so-called *audience* as per the definition given in Subsection 2.2) in order to adjust the transparency level, as indicated in [45]. This defines the principle of *Transparent and Explainable AI*.

- AI products and services should always be aligned with the United Nation's Sustainable Development Goals [387] and contribute to them in a positive and tangible way. Thus, AI should always generate a benefit for humanity and the common good. This defines the principle of *Human-centric AI* (also referred to as *AI for Social Good* [388]).

- AI systems, specially when they are fed by data, should always consider privacy and security standards during all of its life cycle. This principle is not exclusive of AI systems since it is shared with many other software products. Thus, it can be inherited from processes that already exist within a company. This defines the principle of *Privacy and Security by Design*, which was also identified as one of the core ethical and societal challenges faced by Smart Information Systems under the Responsible Research and Innovation paradigm (RRI, [389]). RRI refers to a package of methodological guidelines and recommendations aimed at considering a wider context for scientific research, from the perspective of the lab to global societal challenges such as sustainability, public engagement, ethics, science education, gender equality, open access, and governance. Interestingly, RRI also requires openness and transparency to be ensured in projects embracing its principles, which links directly to the principle of Transparent and Explainable AI mentioned previously.

- The authors emphasize that all these principles should always be extended to any third-party (providers, consultants, partners...).

Going beyond the scope of these five AI principles, the European Commission (EC) has recently published ethical guidelines for Trustworthy AI [390] through an assessment checklist that can be completed by different profiles related to AI systems (namely, product managers, developers and other roles). The assessment is based in a series of principles: 1) human agency and oversight; 2) technical robustness and safety; 3) privacy and data governance; 4) transparency, diversity, non-discrimination and fairness; 5) societal and environmental well-being; 6) accountability. These principles are aligned with the ones detailed in this section, though the scope for the EC principles is more general, including any type of organization involved in the development of AI.

It is worth mentioning that most of these AI principles guides directly approach XAI as a key aspect to consider and include in AI systems. In fact, the overview for these principles introduced before [385], indicates that 28 out of the 32 AI principles guides covered in the analysis, explicitly include XAI as a crucial component. Thus, the work and scope of this article deals directly with one of the most important aspects regarding AI at a worldwide level.

6.2. Fairness and Accountability

As mentioned in the previous section, there are many critical aspects, beyond XAI, included within the different AI principles guidelines published during the last decade. However, those aspects are not completely detached from XAI; in fact, they are intertwined. This section presents two key components with a huge relevance within the AI principles guides, Fairness and Accountability. It also highlights how they are connected to XAI.

6.2.1. Fairness and Discrimination

It is in the identification of implicit correlations between protected and unprotected features where XAI techniques find their place within discrimination-aware data mining methods. By analyzing how the output of the model behaves with respect to the input feature, the model designer may unveil hidden correlations between the input variables amenable to cause discrimination. XAI techniques such as SHAP [224] could be used to generate counterfactual outcomes explaining the decisions of a ML model when fed with protected and unprotected variables.

Recalling the Fair AI principle introduced in the previous section, [386] reminds that fairness is a discipline that generally includes proposals for bias detection within datasets regarding sensitive data that affect protected groups (through variables like gender, race...). Indeed, ethical concerns with black-box models arise from their tendency to unintentionally create unfair decisions by considering sensitive factors such as the individual's race, age or gender [391]. Unfortunately, such unfair decisions can give rise to discriminatory issues, either by explicitly considering sensitive attributes or implicitly by using factors that correlate with sensitive data. In fact, an attribute may implicitly encode a protected factor, as occurs with postal code in credit rating [392]. The aforementioned proposals centered on fairness aspects permit to discover correlations between non-sensitive variables and sensitive ones, detect imbalanced outcomes from the algorithms that penalize a specific subgroup of people (*discrimination*), and mitigate the effect of bias on the model's decisions. These approaches can deal with:

- Individual fairness: here, fairness is analyzed by modeling the differences between each subject and the rest of the population.
- Group fairness: it deals with fairness from the perspective of all individuals.
- Counterfactual fairness: it tries to interpret the causes of bias using, for example, causal graphs.

The sources for bias, as indicated in [392], can be traced to:

- Skewed data: bias within the data acquisition process.
- Tainted data: errors in the data modelling definition, wrong feature labelling, and other possible causes.
- Limited features: using too few features could lead to an inference of false feature relationships that can lead to bias.
- Sample size disparities: when using sensitive features, disparities between different subgroups can induce bias.
- Proxy features: there may be correlated features with sensitive ones that can induce bias even when the sensitive features are not present in the dataset.

The next question that can be asked is what criteria could be used to define when AI is not biased. For supervised ML, [393] presents a framework that uses three criteria to evaluate group fairness when there is a sensitive feature present within the dataset:

- Independence: this criterion is fulfilled when the model predictions are independent of the sensitive feature. Thus, the proportion of positive samples (namely, those ones belonging to the class of interest) given by the model is the same for all the subgroups within the sensitive feature.
- Separation: it is met when the model predictions are independent of the sensitive feature given the target variable. For instance, in classification models, the True Positive (TP) rate and the False Positive (FP) rate are the same in all the subgroups within the sensitive feature. This criteria is also known as *Equalized Odds*.

- Sufficiency: it is accomplished when the target variable is independent of the sensitive feature given the model output. Thus, the Positive Predictive Value is the same for all subgroups within the sensitive feature. This criteria is also known as Predictive Rate Parity.

Although not all of the criteria can be fulfilled at the same time, they can be optimized together in order to minimize the bias within the ML model.

There are two possible actions that could be used in order to achieve those criteria. On one hand, evaluation includes measuring the amount of bias present within the model (regarding one of the criteria aforementioned). There are many different metrics that can be used, depending on the criteria considered. Regarding independence criterion, possible metrics are *statistical parity difference* or *disparate impact*. In case of the separation criterion, possible metrics are *equal opportunity difference* and *average odds difference* [393]. Another possible metric is the *Theil index* [394], which measures inequality both in terms of individual and group fairness.

On the other hand, mitigation refers to the process of fixing some aspects in the model in order to remove the effect of the bias in terms of one or several sensitive features. Several techniques exist within the literature, classified in the following categories:

- Pre-processing: these groups of techniques are applied before the ML model is trained, looking to remove the bias at the first step of the learning process. An example is Reweighing [395], which modifies the weights of the features in order to remove discrimination in sensitive attributes. Another example is [396], which hinges on transforming the input data in order to find a good representation that obfuscates information about membership in sensitive features.

- In-processing: these techniques are applied during the training process of the ML model. Normally, they include Fairness optimization constraints along with cost functions of the ML model. An example is Adversarial Debiasing, [397]. This technique optimizes jointly the ability of predicting the target variable while minimizing the ability of predicting sensitive features using a GAN.

- Post-processing: these techniques are applied after the ML model is trained. They are less intrusive because they do not modify the input data or the ML model. An example is Equalized Odds [393]. This techniques allows to adjust the thresholds in the classification model in order to reduce the differences between the TP rate and the FP rate for each sensitive subgroup.

Even though these references apparently address an AI principle that appears to be independent of XAI, the literature shows that they are intertwined. For instance, the survey in [385] evinces that 26 out of the 28 AI principles that deal with XAI, also talk about fairness explicitly. This fact elucidates that organizations usually consider both aspects together when implementing Responsible AI.

The literature also exploses that XAI proposals can be used for bias detection. For example, [398] proposes a framework to visually analyze the bias present in a model (both for individual and group fairness). Thus, the fairness report is shown just like the visual summaries used within XAI. This explainability approach eases the understanding and measurement of bias. The system must report that there is bias, justify it quantitatively, indicate the degree of fairness, and explain why a user or group would be treated unfairly with the available data. Similarly, XAI techniques such as SHAP [224] could be used to generate counterfactual outcomes explaining the decisions of a ML model when fed with protected and unprotected variables. By identifying implicit correlations between protected and unprotected features through XAI techniques, the model designer may unveil hidden correlations between the input variables amenable to cause discrimination.

Another example is [399], where the authors propose a fair-by-design approach in order to develop ML models that jointly have less bias and include as explanations human comprehensible rules. The proposal is based in self-learning locally generative models that use only a small part of the whole dataset available (weak supervision). It first finds recursively relevant prototypes within the dataset, and

extracts the empirical distribution and density of the points around them. Then it generates rules in an IF/THEN format that explain that a data point is classified within a specific category because it is *similar* to some prototypes. The proposal then includes an algorithm that both generates explanations and reduces bias, as it is demonstrated for the use case of recidivism using the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) dataset [400]. The same goal has been recently pursued in [401], showing that post-hoc XAI techniques can forge fairer explanations from truly unfair black-box models. Finally, CERTIFAI (Counterfactual Explanations for Robustness, Transparency, Interpretability, and Fairness of Artificial Intelligence models) [402] uses a customized genetic algorithm to generate counterfactuals that can help to see the robustness of a ML model, generate explanations, and examine fairness (both at the individual level and at the group level) at the same time.

Strongly linked to the concept of fairness, much attention has been lately devoted to the concept of *data diversity*, which essentially refers to the capability of an algorithmic model to ensure that all different types of objects are represented in its output [403]. Therefore, diversity can be thought to be an indicator of the quality of a collection of items that, when taking the form of a model's output, can quantify the proneness of the model to produce diverse results rather than highly accurate predictions. Diversity comes into play in human-centered applications with ethical restrictions that permeate to the AI modeling phase [404]. Likewise, certain AI problems (such as content recommendation or information retrieval) also aim at producing diverse recommendations rather than highly-scoring yet similar results [405, 406]. In these scenarios, dissecting the internals of a black-box model via XAI techniques can help identifying the capability of the model to maintain the input data diversity at its output. Learning strategies to endow a model with diversity keeping capabilities could be complemented with XAI techniques in order to shed transparency over the model internals, and assess the effectiveness of such strategies with respect to the diversity of the data from which the model was trained. Conversely, XAI could help to discriminate which parts of the model are compromising its overall ability to preserve diversity.

6.2.2. Accountability

Regarding accountability, the EC [390] defines the following aspects to consider:

- Auditability: it includes the assessment of algorithms, data and design processes, but preserving the intellectual property related to the AI systems. Performing the assessment by both internal and external auditors, and making the reports available, could contribute to the trustworthiness of the technology. When the AI system affects fundamental rights, including safety-critical applications, it should always be audited by an external third party.

- Minimization and reporting of negative impacts: it consists of reporting actions or decisions that yield a certain outcome by the system. It also comprises the assessment of those outcomes and how to respond to them. To address that, the development of AI systems should also consider the identification, assessment, documentation and minimization of their potential negative impacts. In order to minimize the potential negative impact, impact assessments should be carried out both prior to and during the development, deployment and use of AI systems. It is also important to guarantee protection for anyone who raises concerns about an AI system (e.g., *whistle-blowers*). All assessments must be proportionate to the risk that the AI systems pose.

- Trade-offs: in case any tension arises due to the implementation of the above requirements, trade-offs could be considered but only if they are ethically acceptable. Such trade-offs should be reasoned, explicitly acknowledged and documented, and they must be evaluated in terms of their risk to ethical principles. The decision maker must be accountable for the manner in which the appropriate trade-off is being made, and the trade-off decided should be continually reviewed to ensure the appropriateness of the decision. If there is no ethically acceptable trade-off, the development, deployment and use of the AI system should not proceed in that form.

- Redress: it includes mechanisms that ensure an adequate redress for situations when unforeseen unjust adverse impacts take place. Guaranteeing a redress for those non-predicted scenarios is a key to ensure trust. Special attention should be paid to vulnerable persons or groups.

These aspects addressed by the EC highlight different connections of XAI with accountability. First, XAI contributes to auditability as it can help explaining AI systems for different profiles, including regulatory ones. Also, since there is a connection between fairness and XAI as stated before, XAI can also contribute to the minimization and report of negative impacts.

6.3. Privacy and Data Fusion

The ever-growing number of information sources that nowadays coexist in almost all domains of activity calls for data fusion approaches aimed at exploiting them simultaneously toward solving a learning task. By merging heterogeneous information, data fusion has been proven to improve the performance of ML models in many applications, such as industrial prognosis [348], cyber-physical social systems [407] or the Internet of Things [408], among others. This section speculates with the potential of data fusion techniques to enrich the explainability of ML models, and to compromise the privacy of the data from which ML models are learned. To this end, we briefly overview different data fusion paradigms, and later analyze them from the perspective of data privacy. As we will later, despite its relevance in the context of Responsible AI, the confluence between XAI and data fusion is an uncharted research area in the current research mainstream.

6.3.1. Basic Levels of Data Fusion

We depart from the different levels of data fusion that have been identified in comprehensive surveys on the matter [409, 410, 411, 412]. In the context of this subsection, we will distinguish among fusion at data level, fusion at model level and fusion at knowledge level. Furthermore, a parallel categorization can be established depending on where such data is processed and fused, yielding centralized and distributed methods for data fusion. In a centralized approach, nodes deliver their locally captured data to a centralized processing system to merge them together. In contrast, in a distributed approach, each of the nodes merges its locally captured information, eventually sharing the result of the local fusion with its counterparts.

Fusion through the information generation process has properties and peculiarities depending on the level at which the fusion is performed. At the so-called *data level*, fusion deals with raw data. As schematically shown in Figure 13, a fusion model at this stage receives raw data from different information sources, and combines them to create a more coherent, compliant, robust or simply representative data flow. On the other hand, fusion at the *model level* aggregates models, each learned from a subset of the data sets that were to be fused. Finally, at the *knowledge level* the fusion approach deals with knowledge in the form of rules, ontologies or other knowledge representation techniques with the intention of merging them to create new, better or more complete knowledge from what was originally provided. Structured knowledge information is extracted from each data source and for every item in the data set using multiple *knowledge extractors* (e.g. a reasoning engine operating on an open semantic database). All produced information is then fused to further ensure the quality, correctness and manageability of the produced knowledge about the items in the data set.

Other data fusion approaches exist beyonds the ones represented in Figure 13. As such, data-level fusion can be performed either by a technique specifically devoted to this end (as depicted in Figure 13.b) or, instead, performed along the learning process of the ML model (as done in e.g. DL models). Similarly, model-level data fusion can be made by combining the decisions of different models (as done in tree ensembles).

6.3.2. Emerging Data Fusion Approaches

In the next subsection we examine other data fusion approaches that have recently come into scene due to their implications in terms of data privacy:

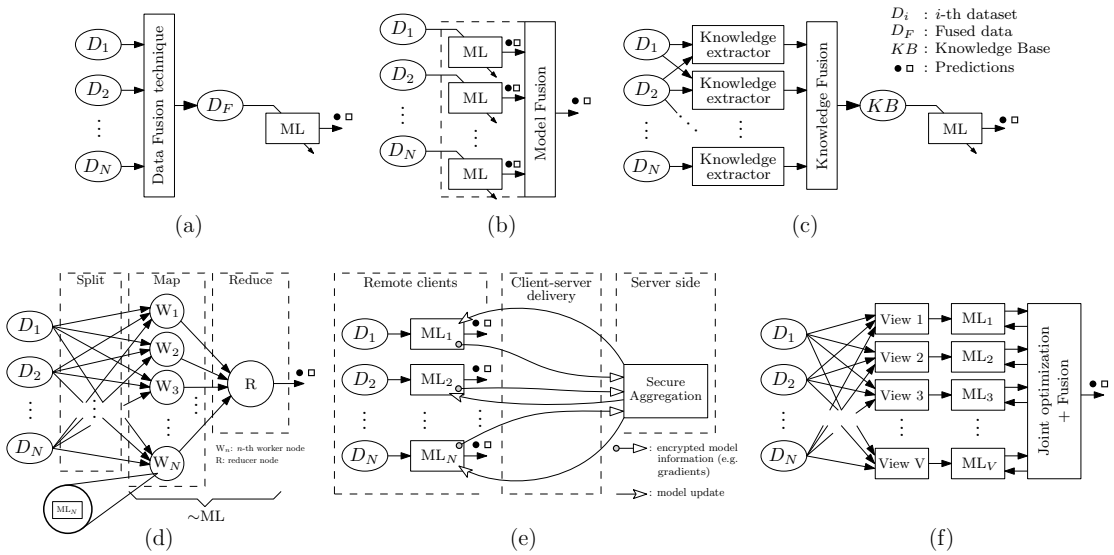


Figure 13: Diagrams showing different levels at which data fusion can be performed: (a) data level; (b) model level; (c) knowledge level; (d) Big Data fusion; (e) Federated Learning and (f) Multiview Learning.

- In Big Data fusion (Figure 13.d), local models are learned on a split of the original data sources, each submitted to a Worker node in charge of performing this learning process (*Map* task). Then, a *Reduce* node (or several *Reduce* nodes, depending on the application) combines the outputs produced by each *Map* task. Therefore, Big Data fusion can be conceived as a means to distribute the complexity of learning a ML model over a pool of Worker nodes, wherein the strategy to design how information/models are fused together between the *Map* and the *Reduce* tasks is what defines the quality of the finally generated outcome [413].

- By contrast, in Federated Learning [414, 415, 416], the computation of ML models is made on data captured locally by remote client devices (Figure 13.e). Upon local model training, clients transmit encrypted information about their learned knowledge to a central server, which can take the form of layer-wise gradients (in the case of neural ML models) or any other model-dependent content alike. The central server aggregates (fuses) the knowledge contributions received from all clients to yield a shared model harnessing the collected information from the pool of clients. It is important to observe that no client data is delivered to the central server, which elicits the privacy-preserving nature of Federated Learning. Furthermore, computation is set closer to the collected data, which reduces the processing latency and alleviates the computational burden of the central server.

- Finally, Multiview Learning [417] constructs different *views* of the object as per the information contained in the different data sources (Figure 13.f). These views can be produced from multiple sources of information and/or different feature subsets [418]. Multiview Learning devises strategies to jointly optimize ML models learned from the aforementioned views to enhance the generalization performance, specially in those applications with weak data supervision and hence, prone to model overfitting. This joint optimization resorts to different algorithmic means, from co-training to co-regularization [419].

6.3.3. Opportunities and Challenges in Privacy and Data Fusion under the Responsible AI Paradigm

AI systems, specially when dealing with multiple data sources, need to explicitly include privacy considerations during the system's life cycle. This is specially critical when working with personal data,

because respecting people's right to privacy should always be addressed. The EC highlights that privacy should also address data governance, covering the quality and integrity of the used data [390]. It should also include the definition of access protocols and the capability to process data in a way that ensures privacy. The EC guide breaks down the privacy principle into three aspects:

- Privacy and data protection: they should be guaranteed in AI systems throughout its entire lifecycle. It includes both information provided by users and information generated about those users derived from their interactions with the system. Since digital information about a user could be used in a negative way against them (discrimination due to sensitive features, unfair treatment...), it is crucial to ensure proper usage of all the data collected.

- Quality and integrity of data: quality of data sets is fundamental to reach good performance with AI systems that are fueled with data, like ML. However, sometimes the data collected contains socially constructed biases, inaccuracies, errors and mistakes. This should be tackled before training any model with the data collected. Additionally, the integrity of the data sets should be ensured.

- Access to data: if there is individual personal data, there should always be data protocols for data governance. These protocols should indicate who may access data and under which circumstances.

The aforementioned examples from the EC shows how data fusion is directly intertwined with privacy and with fairness, regardless of the technique employed for it.

Notwithstanding this explicit concern from regulatory bodies, loss of privacy has been compromised by DL methods in scenarios where no data fusion is performed. For instance, a few images are enough to threaten users' privacy even in the presence of image obfuscation [420], and the model parameters of a DNN can be exposed by simply performing input queries on the model [356, 357]. An approach to explain loss of privacy is by using *privacy loss* and *intent loss* subjective scores. The former provides a subjective measure of the severity of the privacy violation depending on the role of a face in the image, while the latter captures the intent of the bystanders to appear in the picture. These kind of explanations have motivated, for instance, secure matching cryptographic protocols for photographer and bystanders to preserve privacy [356, 421, 422]. We definite advocate for more efforts invested in this direction, namely, in ensuring that XAI methods do not pose a threat in regards to the privacy of the data used for training the ML model under target.

When data fusion enters the picture, different implications arise with the context of explainability covered in this survey. To begin with, classical techniques for fusion at the data level only deal with data and have no connection to the ML model, so they have little to do with explainability. However, the advent of DL models has blurred the distinction between information fusion and predictive modeling. The first layers of DL architectures are in charge of learning high-level features from raw data that possess relevance for the task at hand. This learning process can be thought to aim at solving a data level fusion problem, yet in a directed learning fashion that makes the fusion process tightly coupled to the task to be solved.

In this context, many techniques in the field of XAI have been proposed to deal with the analysis of correlation between features. This paves the way to explaining how data sources are actually fused through the DL model, which can yield interesting insights on how the predictive task at hand induces correlations among the data sources over the spatial and/or time domain. Ultimately, this gained information on the fusion could not only improve the usability of the model as a result of its enhanced understanding by the user, but could also help identifying other data sources of potential interest that could be incorporated to the model, or even contribute to a more efficient data fusion in other contexts.

Unfortunately, this previously mentioned concept of fusion at data level contemplates data under certain constraints of known form and source origin. As presented in [423], the Big Data era presents an environment in which these premises cannot be taken for granted, and methods to board Big Data fusion (as that illustrated in Figure 13.d) have to be thought. Conversely, a concern with model fusion

context emerges in the possibility that XAI techniques could be explanatory enough to compromise the confidentiality of private data. This could eventually occur if sensitive information (e.g. ownership) could be inferred from the explained fusion among protected and unprotected features.

When turning our prospects to data fusion at model level, we have already argued that the fusion of the outputs of several transparent models (as in tree ensembles) could make the overall model opaque, thereby making it necessary to resort to post-hoc explainability solutions. However, model fusion may entail other drawbacks when endowed with powerful post-hoc XAI techniques. Let us imagine that relationships of a model's input features have been discovered by means of a post-hoc technique) and that one of those features is hidden or unknown. Will it be possible to infer another model's features if that previous feature was known to be used in that model? Would this possibility uncover a problem as privacy breaches in cases in which related protected input variables are not even shared in the first place?

To get the example clearer, in [424] a multiview perspective is utilized in which different single views (representing the sources they attend to) models are fused. These models contain among others, cell-phone data, transportation data, etc. which might introduce the problem that information that is not even shared can be discovered through other sources that are actually shared. In the example above, what if instead of features, a model shares with another a layer or part of its architecture as in Federated Learning? Would this sharing make possible to infer information from that exchanged part of its model, to the extent of allowing for the design of adversarial attacks with better success rate upon the antecedent model?

If focused at knowledge level fusion, a similar reasoning holds: XAI comprises techniques that extract knowledge from ML model(s). This ability to explain models could have an impact on the necessity of discovering new knowledge through the complex interactions formed within ML models. If so, XAI might enrich knowledge fusion paradigms, bringing the possibility of discovering new knowledge extractors of relevance for the task at hand. For this purpose, it is of paramount importance that the knowledge extracted from a model by means of XAI techniques can be understood and extrapolated to the domain in which knowledge extractors operate. The concept matches with ease with that of transfer learning portrayed in [425]. Although XAI is not contemplated in the surveyed processes of extracting knowledge from models trained in certain feature spaces and distributions, to then be utilized in environments where previous conditions do not hold, when deployed, XAI can pose a threat if the explanations given about the model can be reversely engineered through the knowledge fusion paradigm to eventually compromise, for instance, the differential privacy of the overall model.

The distinction between centralized and distributed data fusion also spurs further challenges in regards to privacy and explainability. The centralized approach does not bring any further concerns that those presented above. However, distributed fusion does arise new problems. Distributed fusion might be applied for different reasons, mainly due to environmental constraints or due to security or privacy issues. The latter context may indulge some dangers. Among other goals (e.g. computational efficiency), model-level data fusion is performed in a distributed fashion to ensure that no actual data is actually shared, but rather parts of an ML model trained on local data. This rationale lies at the heart of Federated Learning, where models exchange locally learned information among nodes. Since data do not leave the local device, only the transmission of model updates is required across distributed devices. This lightens the training process for network-compromised settings and guarantees data privacy [416]. Upon the use of post-hoc explainability techniques, a node could disguise sensitive information about the local context in which the received ML model part was trained. In fact, it was shown that a black-box model based on a DNN from which an input/output query interface is given can be used to accurately predict every single hyperparameter value used for training, allowing for potential privacy-related consequences [357, 420, 421]. This relates to studies showing that blurring images does not guarantee privacy preservation.

Data fusion, privacy and model explainability are concepts that have not been analysed together so far. From the above discussion it is clear that there are unsolved concerns and caveats that demand further study by the community in forthcoming times.

6.4. Implementing Responsible AI Principles in an Organization

While increasingly more organizations are publishing AI principles to declare that they care about avoiding unintended negative consequences, there is much less experience on how to actually implement the principles into an organization. Looking at several examples of principles declared by different organizations [385], we can divide them into two groups:

- AI-specific principles that focus on aspects that are specific to AI, such as explainability, fairness and human agency.
- End-to-end principles that cover all aspects involved in AI, including also privacy, security and safety.

The EC Guidelines for Trustworthy AI are an example of end-to-end principles [390], while those of Telefonica (a large Spanish ICT company operating worldwide) are more AI-specific [386]. For example, safety and security are relevant for any connected IT system, and therefore also for AI systems. The same holds for privacy, but it is probably true that privacy in the context of AI systems is even more important than for general IT systems, due to the fact that ML models need huge amounts of data and most importantly, because XAI tools and data fusion techniques pose new challenges to preserve the privacy of protected records.

When it comes to implement the AI Principles into an organization, it is important to operationalize the AI-specific parts and, at the same time, leverage the processes already existing for the more generic principles. Indeed, in many organizations there already exist norms and procedures for privacy, security and safety. Implementing AI principles requires a methodology such as that presented in [386] that breaks down the process into different parts. The ingredients of such a methodology should include, at least:

- AI principles (already discussed earlier), which set the values and boundaries.
- Awareness and training about the potential issues, both technical and non-technical.
- A questionnaire that forces people to think about certain impacts of the AI system (*impact explanation*). This questionnaire should give concrete guidance on what to do if certain undesired impacts are detected.
- Tools that help answering some of the questions, and help mitigating any problems identified. XAI tools and fairness tools fall in this category, as well as other recent proposals such as *model cards* [426].
- A governance model assigning responsibilities and accountabilities (*responsibility explanation*). There are two philosophies for governance: 1) based on committees that review and approve AI developments, and 2) based on the self-responsibility of the employees. While both are possible, given the fact that agility is key for being successful in the digital world, it seems wiser to focus on awareness and employee responsibility, and only use committees when there are specific, but important issues.

From the above elaborations, it is clear that the implementation of Responsible AI principles in companies should balance between two requirements: 1) major cultural and organizational changes needed to enforce such principles over processes endowed with AI functionalities; and 2) the feasibility and compliance of the implementation of such principles with the IT assets, policies and resources already available at the company. It is in the gradual process of rising corporate awareness around the principles and values of Responsible AI where we envision that XAI will make its place and create huge impact.

7. Conclusions and Outlook

This overview has revolved around eXplainable Artificial Intelligence (XAI), which has been identified in recent times as an utmost need for the adoption of ML methods in real-life applications. Our study

has elaborated on this topic by first clarifying different concepts underlying model explainability, as well as by showing the diverse purposes that motivate the search for more interpretable ML methods. These conceptual remarks have served as a solid baseline for a systematic review of recent literature dealing with explainability, which has been approached from two different perspectives: 1) ML models that feature some degree of transparency, thereby interpretable to an extent by themselves; and 2) post-hoc XAI techniques devised to make ML models more interpretable. This literature analysis has yielded a global taxonomy of different proposals reported by the community, classifying them under uniform criteria. Given the prevalence of contributions dealing with the explainability of Deep Learning models, we have inspected in depth the literature dealing with this family of models, giving rise to an alternative taxonomy that connects more closely with the specific domains in which explainability can be realized for Deep Learning models.

We have moved our discussions beyond what has been made so far in the XAI realm toward the concept of Responsible AI, a paradigm that imposes a series of AI principles to be met when implementing AI models in practice, including fairness, transparency, and privacy. We have also discussed the implications of adopting XAI techniques in the context of data fusion, unveiling the potential of XAI to compromise the privacy of protected data involved in the fusion process. Implications of XAI in fairness have also been discussed in detail. This vision of XAI as a core concept to ensure the aforementioned principles for Responsible AI is summarized graphically in Figure 14.

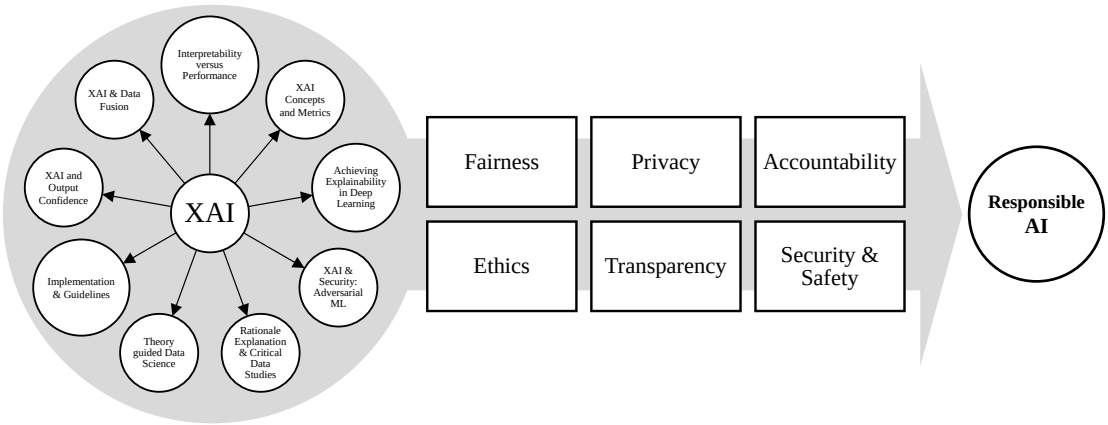


Figure 14: Summary of XAI challenges discussed in this overview and its impact on the principles for Responsible AI.

Our reflections about the future of XAI, conveyed in the discussions held throughout this work, agree on the compelling need for a proper understanding of the potentiality and caveats opened up by XAI techniques. It is our vision that model interpretability must be addressed jointly with requirements and constraints related to data privacy, model confidentiality, fairness and accountability. A responsible implementation and use of AI methods in organizations and institutions worldwide will be only guaranteed if all these AI principles are studied jointly.

Acknowledgments

Alejandro Barredo-Arrieta, Javier Del Ser and Sergio Gil-Lopez would like to thank the Basque Government for the funding support received through the EMAITEK and ELKARTEK programs. Javier Del Ser also acknowledges funding support from the Consolidated Research Group MATHMODE (IT1294-19) granted by the Department of Education of the Basque Government. Siham Tabik, Salvador Garcia, Daniel Molina and Francisco Herrera would like to thank the Spanish Government for its funding support (SMART-DaSCI project, TIN2017-89517-P), as well as the BBVA Foundation through its *Ayudas*

Fundación BBVA a Equipos de Investigación Científica 2018 call (DeepSCOP project). This work was also funded in part by the European Union's Horizon 2020 research and innovation programme AI4EU under grant agreement 825619. We also thank Chris Olah, Alexander Mordvintsev and Ludwig Schubert for borrowing images for illustration purposes. Part of this overview is inspired by a preliminary work of the concept of Responsible AI: R. Benjamins, A. Barbado, D. Sierra, *"Responsible AI by Design"*, to appear in the Proceedings of the Human-Centered AI: Trustworthiness of AI Models & Data (HAI) track at AAAI Fall Symposium, DC, November 7-9, 2019 [386].

References

[1] S. J. Russell, P. Norvig, Artificial intelligence: a modern approach, Malaysia; Pearson Education Limited,, 2016.

[2] D. M. West, The future of work: robots, AI, and automation, Brookings Institution Press, 2018.

[3] B. Goodman, S. Flaxman, European union regulations on algorithmic decision-making and a "right to explanation", AI Magazine 38 (3) (2017) 50–57.

[4] D. Castelvecchi, Can we open the black box of AI?, Nature News 538 (7623) (2016) 20.

[5] Z. C. Lipton, The mythos of model interpretability, Queue 16 (3) (2018) 30:31–30:57.

[6] A. Preece, D. Harborne, D. Braines, R. Tomsett, S. Chakraborty, Stakeholders in Explainable AI (2018). arXiv:1810.00184.

[7] D. Gunning, Explainable artificial intelligence (xAI), Tech. rep., Defense Advanced Research Projects Agency (DARPA) (2017).

[8] E. Tjoa, C. Guan, A survey on explainable artificial intelligence (XAI): Towards medical XAI (2019). arXiv:1907.07374.

[9] J. Zhu, A. Liapis, S. Risi, R. Bidarra, G. M. Youngblood, Explainable AI for designers: A human-centered perspective on mixed-initiative co-creation, 2018 IEEE Conference on Computational Intelligence and Games (CIG) (2018) 1–8.

[10] F. K. Došilović, M. Brčić, N. Hlupić, Explainable artificial intelligence: A survey, in: 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2018, pp. 210–215.

[11] P. Hall, On the Art and Science of Machine Learning Explanations (2018). arXiv:1810.02909.

[12] T. Miller, Explanation in artificial intelligence: Insights from the social sciences, Artif. Intell. 267 (2019) 1–38.

[13] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, L. Kagal, Explaining Explanations: An Overview of Interpretability of Machine Learning (2018). arXiv:1806.00069.

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

[15] O. Biran, C. Cotton, Explanation and justification in machine learning: A survey, in: IJCAI-17 workshop on explainable AI (XAI), Vol. 8, 2017, p. 1.

[16] S. T. Shane T. Mueller, R. R. Hoffman, W. Clancey, G. Klein, Explanation in Human-AI Systems: A Literature Meta-Review Synopsis of Key Ideas and Publications and Bibliography for Explainable AI, Tech. rep., Defense Advanced Research Projects Agency (DARPA) XAI Program (2019).

[17] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, D. Pedreschi, A survey of methods for explaining black box models, ACM Computing Surveys 51 (5) (2018) 93:1–93:42.

[18] G. Montavon, W. Samek, K.-R. Müller, Methods for interpreting and understanding deep neural networks, Digital Signal Processing 73 (2018) 1–15. doi:10.1016/j.dsp.2017.10.011.

[19] A. Fernandez, F. Herrera, O. Cordon, M. Jose del Jesus, F. Marcelloni, Evolutionary fuzzy systems for explainable artificial intelligence: Why, when, what for, and where to?, IEEE Computational Intelligence Magazine 14 (1) (2019) 69–81.

[20] M. Gleicher, A framework for considering comprehensibility in modeling, Big data 4 (2) (2016) 75–88.

[21] M. W. Craven, Extracting comprehensible models from trained neural networks, Tech. rep., University of Wisconsin-Madison Department of Computer Sciences (1996).

[22] R. S. Michalski, A theory and methodology of inductive learning, in: Machine learning, Springer, 1983, pp. 83–134.

[23] J. Díez, K. Khalifa, B. Leuridan, General theories of explanation: buyer beware, Synthese 190 (3) (2013) 379–396.

[24] D. Doran, S. Schulz, T. R. Besold, What does explainable AI really mean? a new conceptualization of perspectives (2017). arXiv:1710.00794.

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

[26] A. Vellido, J. D. Martín-Guerrero, P. J. Lisboa, Making machine learning models interpretable., in: European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN), Vol. 12, Citeseer, 2012, pp. 163–172.

[27] E. Walter, Cambridge advanced learner's dictionary, Cambridge University Press, 2008.

[28] P. Besnard, A. Hunter, Elements of Argumentation, The MIT Press, 2008.

[29] F. Rossi, AI Ethics for Enterprise AI (2019).
URL https://economics.harvard.edu/files/economics/files/rossi-francesca_4-22-19_ai-ethics-for-enterprise-ai_ec3118-hbs.pdf

[30] A. Holzinger, C. Biemann, C. S. Pattichis, D. B. Kell, What do we need to build explainable Ai systems for the medical domain? (2017). arXiv:1712.09923.

[31] B. Kim, E. Glassman, B. Johnson, J. Shah, iBCM: Interactive bayesian case model empowering humans via intuitive interaction, Tech. rep., MIT-CSAIL-TR-2015-010 (2015).

[32] M. T. Ribeiro, S. Singh, C. Guestrin, Why should I trust you?: Explaining the predictions of any classifier, in: ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2016, pp. 1135–1144.

[33] M. Fox, D. Long, D. Magazzeni, Explainable planning (2017). arXiv:1709.10256.

[34] H. C. Lane, M. G. Core, M. Van Lent, S. Solomon, D. Gomboc, Explainable artificial intelligence for training and tutoring, Tech. rep., University of Southern California (2005).

[35] W. J. Murdoch, C. Singh, K. Kumbier, R. Abbasi-Asl, B. Yu, Interpretable machine learning: definitions, methods, and applications (2019). arXiv:1901.04592.

[36] J. Haspiel, N. Du, J. Meyerson, L. P. Robert Jr, D. Tilbury, X. J. Yang, A. K. Pradhan, Explanations and expectations: Trust building in automated vehicles, in: Companion of the ACM/IEEE International Conference on Human-Robot Interaction, ACM, 2018, pp. 119–120.

[37] A. Chander, R. Srinivasan, S. Chelian, J. Wang, K. Uchino, Working with beliefs: AI transparency in the enterprise., in: Workshops of the ACM Conference on Intelligent User Interfaces, 2018.

[38] A. B. Tickle, R. Andrews, M. Golea, J. Diederich, The truth will come to light: Directions and challenges in extracting the knowledge embedded within trained artificial neural networks, IEEE Transactions on Neural Networks 9 (6) (1998) 1057–1068.

[39] C. Louizos, U. Shalit, J. M. Mooij, D. Sontag, R. Zemel, M. Welling, Causal effect inference with deep latent-variable models, in: Advances in Neural Information Processing Systems, 2017, pp. 6446–6456.

[40] O. Goudet, D. Kalainathan, P. Caillou, I. Guyon, D. Lopez-Paz, M. Sebag, Learning functional causal models with generative neural networks, in: Explainable and Interpretable Models in Computer Vision and Machine Learning, Springer, 2018, pp. 39–80.

[41] S. Athey, G. W. Imbens, Machine learning methods for estimating heterogeneous causal effects, stat 1050 (5) (2015).

[42] D. Lopez-Paz, R. Nishihara, S. Chintala, B. Scholkopf, L. Bottou, Discovering causal signals in images, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 6979–6987.

[43] C. Barabas, K. Dinakar, J. Ito, M. Virza, J. Zittrain, Interventions over predictions: Reframing the ethical debate for actuarial risk assessment (2017). arXiv:1712.08238.

[44] R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, N. Elhadad, Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission, in: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '15, 2015, pp. 1721–1730.

[45] A. Theodorou, R. H. Wortham, J. J. Bryson, Designing and implementing transparency for real time inspection of autonomous robots, Connection Science 29 (3) (2017) 230–241.

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

[47] C. Wadsworth, F. Vera, C. Piech, Achieving fairness through adversarial learning: an application to recidivism prediction (2018). arXiv:1807.00199.

[48] X. Yuan, P. He, Q. Zhu, X. Li, Adversarial examples: Attacks and defenses for deep learning, IEEE Transactions on Neural Networks and Learning Systems 30 (9) (2019) 2805–2824.

[49] B. Letham, C. Rudin, T. H. McCormick, D. Madigan, et al., Interpretable classifiers using rules and bayesian analysis: Building a better stroke prediction model, The Annals of Applied Statistics 9 (3) (2015) 1350–1371.

[50] M. Harbers, K. van den Bosch, J.-J. Meyer, Design and evaluation of explainable BDI agents, in: IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, Vol. 2, IEEE, 2010, pp. 125–132.

[51] M. H. Aung, P. G. Lisboa, T. A. Etchells, A. C. Testa, B. Van Calster, S. Van Huffel, L. Valentin, D. Timmerman, Comparing analytical decision support models through boolean rule extraction: A case study of ovarian tumour malignancy, in: International Symposium on Neural Networks, Springer, 2007, pp. 1177–1186.

[52] A. Weller, Challenges for transparency (2017). arXiv:1708.01870.

[53] A. A. Freitas, Comprehensible classification models: a position paper, ACM SIGKDD explorations newsletter 15 (1) (2014) 1–10.

[54] V. Schetinin, J. E. Fieldsend, D. Partridge, T. J. Coats, W. J. Krzanowski, R. M. Everson, T. C. Bailey, A. Hernandez, Confident interpretation of bayesian decision tree ensembles for clinical applications, IEEE Transactions on Information Technology in Biomedicine 11 (3) (2007) 312–319.

[55] D. Martens, J. Vanthienen, W. Verbeke, B. Baesens, Performance of classification models from a user perspective, Decision Support Systems 51 (4) (2011) 782–793.

[56] Z. Che, S. Purushotham, R. Khemani, Y. Liu, Interpretable deep models for ICU outcome prediction, in: AMIA Annual Symposium Proceedings, Vol. 2016, American Medical Informatics Association, 2016, p. 371.

[57] N. Barakat, J. Diederich, Eclectic rule-extraction from support vector machines, International Journal of Computer, Electrical, Automation, Control and Information Engineering 2 (5) (2008) 1672–1675.

[58] F. J. C. Garcia, D. A. Robb, X. Liu, A. Laskov, P. Patron, H. Hastie, Explain yourself: A natural language interface for scrutable autonomous robots (2018). arXiv:1803.02088.

[59] P. Langley, B. Meadows, M. Sridharan, D. Choi, Explainable agency for intelligent autonomous systems, in: AAAI Conference on Artificial Intelligence, 2017, pp. 4762–4763.

[60] G. Montavon, S. Lapuschkin, A. Binder, W. Samek, K.-R. Müller, Explaining nonlinear classification decisions with deep taylor decomposition, Pattern Recognition 65 (2017) 211–222.

[61] P.-J. Kindermans, K. T. Schütt, M. Alber, K.-R. Müller, D. Erhan, B. Kim, S. Dähne, Learning how to explain neural networks: Patternnet and patternattribution (2017). arXiv:1705.05598.

[62] G. Ras, M. van Gerven, P. Haselager, Explanation methods in deep learning: Users, values, concerns and challenges, in: Explainable and Interpretable Models in Computer Vision and Machine Learning, Springer, 2018, pp. 19–36.

[63] S. Bach, A. Binder, K.-R. Müller, W. Samek, Controlling explanatory heatmap resolution and semantics via decomposition depth, in: IEEE International Conference on Image Processing (ICIP), IEEE, 2016, pp. 2271–2275.

[64] G. J. Katuwal, R. Chen, Machine learning model interpretability for precision medicine (2016). arXiv:1610.09045.

[65] M. A. Neerincx, J. van der Waa, F. Kaptein, J. van Diggelen, Using perceptual and cognitive explanations for enhanced human-agent team performance, in: International Conference on Engineering Psychology and Cognitive Ergonomics, Springer, 2018, pp. 204–214.

[66] J. D. Olden, D. A. Jackson, Illuminating the "black box": a randomization approach for understanding variable contributions in artificial neural networks, Ecological modelling 154 (1-2) (2002) 135–150.

[67] J. Krause, A. Perer, K. Ng, Interacting with predictions: Visual inspection of black-box machine learning models, in: CHI Conference on Human Factors in Computing Systems, ACM, 2016, pp. 5686–5697.

[68] L. Rosenbaum, G. Hinselmann, A. Jahn, A. Zell, Interpreting linear support vector machine models with heat map molecule coloring, Journal of Cheminformatics 3 (1) (2011) 11.

[69] J. Tan, M. Ung, C. Cheng, C. S. Greene, Unsupervised feature construction and knowledge extraction from genome-wide assays of breast cancer with denoising autoencoders, in: Pacific Symposium on Biocomputing Co-Chairs, World Scientific, 2014, pp. 132–143.

[70] S. Krening, B. Harrison, K. M. Feigh, C. L. Isabell, M. Riedl, A. Thomaz, Learning from explanations using sentiment and advice in RL, IEEE Transactions on Cognitive and Developmental Systems 9 (1) (2017) 44–55.

[71] M. T. Ribeiro, S. Singh, C. Guestrin, Model-agnostic interpretability of machine learning (2016). arXiv:1606.05386.

[72] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, W. Samek, On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation, PloS one 10 (7) (2015) e0130140.

[73] T. A. Etchells, P. J. Lisboa, Orthogonal search-based rule extraction (OSRE) for trained neural networks: a practical and efficient approach, IEEE Transactions on Neural Networks 17 (2) (2006) 374–384.

[74] Y. Zhang, S. Sreedharan, A. Kulkarni, T. Chakraborti, H. H. Zhuo, S. Kambhampati, Plan explicability and predictability for robot task planning, in: 2017 IEEE International Conference on Robotics and Automation (ICRA), IEEE, 2017, pp. 1313–1320.

[75] A. Santoro, D. Raposo, D. G. Barrett, M. Malinowski, R. Pascanu, P. Battaglia, T. Lillicrap, A simple neural network module for relational reasoning, in: Advances in Neural Information Processing Systems, 2017, pp. 4967–4976.

[76] C.-Y. J. Peng, T.-S. H. So, F. K. Stage, E. P. S. John, The use and interpretation of logistic regression in higher education journals: 1988–1999, Research in Higher Education 43 (3) (2002) 259–293.

[77] B. Üstün, W. Melssen, L. Buydens, Visualisation and interpretation of support vector regression models, Analytica Chimica Acta 595 (1-2) (2007) 299–309.

[78] Q. Zhang, Y. Yang, H. Ma, Y. N. Wu, Interpreting CNNs via decision trees, in: IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 6261–6270.

[79] M. Wu, M. C. Hughes, S. Parbhoo, M. Zazzi, V. Roth, F. Doshi-Velez, Beyond sparsity: Tree regularization of deep models for interpretability, in: AAAI Conference on Artificial Intelligence, 2018, pp. 1670–1678.

[80] G. Hinton, O. Vinyals, J. Dean, Distilling the knowledge in a neural network (2015). arXiv:1503.02531.

[81] N. Frosst, G. Hinton, Distilling a neural network into a soft decision tree (2017). arXiv:1711.09784.

[82] M. G. Augasta, T. Kathirvalavakumar, Reverse engineering the neural networks for rule extraction in classification problems, Neural Processing Letters 35 (2) (2012) 131–150.

[83] Z.-H. Zhou, Y. Jiang, S.-F. Chen, Extracting symbolic rules from trained neural network ensembles, AI Communications 16 (1) (2003) 3–15.

[84] H. F. Tan, G. Hooker, M. T. Wells, Tree space prototypes: Another look at making tree ensembles interpretable (2016). arXiv:1611.07115.

[85] R. C. Fong, A. Vedaldi, Interpretable explanations of black boxes by meaningful perturbation, in: IEEE International Conference on Computer Vision, 2017, pp. 3429–3437.

[86] T. Miller, P. Howe, L. Sonenberg, Explainable AI: Beware of inmates running the asylum, in: International Joint Conference on Artificial Intelligence, Workshop on Explainable AI (XAI), Vol. 36, 2017, pp. 36–40.

[87] R. Goebel, A. Chander, K. Holzinger, F. Lecue, Z. Akata, S. Stumpf, P. Kieseberg, A. Holzinger, Explainable AI: the new 42?, in: International Cross-Domain Conference for Machine Learning and Knowledge Extraction, Springer, 2018, pp. 295–303.

[88] V. Belle, Logic meets probability: Towards explainable AI systems for uncertain worlds, in: International Joint Conference on Artificial Intelligence, 2017, pp. 5116–5120.

[89] L. Edwards, M. Veale, Slave to the algorithm: Why a right to an explanation is probably not the remedy you are looking for, Duke L. & Tech. Rev. 16 (2017) 18.

[90] Y. Lou, R. Caruana, J. Gehrke, G. Hooker, Accurate intelligible models with pairwise interactions, in: ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2013, pp. 623–631.

[91] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, Y. Bengio, Show, attend and tell: Neural image caption generation with visual attention, in: International Conference on Machine Learning, 2015, pp. 2048–2057.

[92] J. Huysmans, K. Dejaeger, C. Mues, J. Vanthienen, B. Baesens, An empirical evaluation of the comprehensibility of decision table, tree and rule based predictive models, Decision Support Systems 51 (1) (2011) 141–154.

[93] N. H. Barakat, A. P. Bradley, Rule extraction from support vector machines: A sequential covering approach, IEEE Transactions on Knowledge and Data Engineering 19 (6) (2007) 729–741.

[94] F. C. Adriana da Costa, M. M. B. Vellasco, R. Tanscheit, Fuzzy rule extraction from support vector machines, in: International Conference on Hybrid Intelligent Systems, IEEE, 2005, pp. 335–340.

[95] D. Martens, B. Baesens, T. Van Gestel, J. Vanthienen, Comprehensible credit scoring models using rule extraction from support vector machines, European Journal of Operational Research 183 (3) (2007) 1466–1476.

[96] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, A. Torralba, Learning deep features for discriminative localization, in: IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2921–2929.

[97] R. Krishnan, G. Sivakumar, P. Bhattacharya, Extracting decision trees from trained neural networks, Pattern Recognition 32 (12) (1999) 1999–2009.

[98] X. Fu, C. Ong, S. Keerthi, G. G. Hung, L. Goh, Extracting the knowledge embedded in support vector machines, in: IEEE International Joint Conference on Neural Networks, Vol. 1, IEEE, 2004, pp. 291–296.

[99] B. Green, "Fair" risk assessments: A precarious approach for criminal justice reform, in: 5th Workshop on Fairness, Accountability, and Transparency in Machine Learning, 2018.

[100] A. Chouldechova, Fair prediction with disparate impact: A study of bias in recidivism prediction instruments, Big Data 5 (2) (2017) 153–163.

[101] M. Kim, O. Reingold, G. Rothblum, Fairness through computationally-bounded awareness, in: Advances in Neural Information Processing Systems, 2018, pp. 4842–4852.

[102] B. Haasdonk, Feature space interpretation of SVMs with indefinite kernels, IEEE Transactions on Pattern Analysis and Machine Intelligence 27 (4) (2005) 482–492.

[103] A. Palczewska, J. Palczewski, R. M. Robinson, D. Neagu, Interpreting random forest classification models using a feature contribution method, in: Integration of Reusable Systems, Springer, 2014, pp. 193–218.

[104] S. H. Welling, H. H. Refsgaard, P. B. Brockhoff, L. H. Clemmensen, Forest floor visualizations of random forests (2016). arXiv:1605.09196.

[105] G. Fung, S. Sandilya, R. B. Rao, Rule extraction from linear support vector machines, in: ACM SIGKDD International Conference on Knowledge Discovery in Data Mining, ACM, 2005, pp. 32–40.

[106] Y. Zhang, H. Su, T. Jia, J. Chu, Rule extraction from trained support vector machines, in: Pacific-Asia Conference on Knowledge Discovery and Data Mining, Springer, 2005, pp. 61–70.

[107] D. Linsley, D. Shiebler, S. Eberhardt, T. Serre, Global-and-local attention networks for visual recognition (2018). arXiv:1805.08819.

[108] S.-M. Zhou, J. Q. Gan, Low-level interpretability and high-level interpretability: a unified view of data-driven interpretable fuzzy system modelling, Fuzzy Sets and Systems 159 (23) (2008) 3091–3131.

[109] J. Burrell, How the machine 'thinks': Understanding opacity in machine learning algorithms, Big Data & Society 3 (1) (2016) 1–12.

[110] A. Shrikumar, P. Greenside, A. Shcherbina, A. Kundaje, Not just a black box: Learning important features through propagating activation differences (2016). arXiv:1605.01713.

[111] Y. Dong, H. Su, J. Zhu, B. Zhang, Improving interpretability of deep neural networks with semantic information, in: IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4306–4314.

[112] G. Ridgeway, D. Madigan, T. Richardson, J. O'Kane, Interpretable boosted naïve bayes classification., in: ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 1998, pp. 101–104.

[113] Q. Zhang, Y. Nian Wu, S.-C. Zhu, Interpretable convolutional neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 8827–8836.

[114] S. Seo, J. Huang, H. Yang, Y. Liu, Interpretable convolutional neural networks with dual local and global attention for review rating prediction, in: Proceedings of the Eleventh ACM Conference on Recommender Systems, ACM, 2017, pp. 297–305.

[115] K. Larsen, J. H. Petersen, E. Budtz-Jørgensen, L. Endahl, Interpreting parameters in the logistic regression model with random effects, Biometrics 56 (3) (2000) 909–914.

[116] B. Gaonkar, R. T. Shinohara, C. Davatzikos, A. D. N. Initiative, et al., Interpreting support vector machine models for multivariate group wise analysis in neuroimaging, Medical image analysis 24 (1) (2015) 190–204.

[117] K. Xu, D. H. Park, C. Yi, C. Sutton, Interpreting deep classifier by visual distillation of dark knowledge (2018). arXiv:1803.04042.

[118] H. Deng, Interpreting tree ensembles with intrees (2014). arXiv:1408.5456.

[119] P. Domingos, Knowledge discovery via multiple models, Intelligent Data Analysis 2 (1-4) (1998) 187–202.

[120] S. Tan, R. Caruana, G. Hooker, Y. Lou, Distill-and-compare: Auditing black-box models using transparent model distillation, in: AAAI/ACM Conference on AI, Ethics, and Society, ACM, 2018, pp. 303–310.

[121] R. A. Berk, J. Bleich, Statistical procedures for forecasting criminal behavior: A comparative assessment, Criminology & Public Policy 12 (3) (2013) 513–544.

[122] S. Hara, K. Hayashi, Making tree ensembles interpretable (2016). arXiv:1606.05390.

[123] A. Henelius, K. Puolamäki, A. Ukkonen, Interpreting classifiers through attribute interactions in datasets (2017). arXiv:1707.07576.

[124] H. Hastie, F. J. C. Garcia, D. A. Robb, P. Patron, A. Laskov, MIRIAM: a multimodal chat-based interface for autonomous systems, in: ACM International Conference on Multimodal Interaction, ACM, 2017, pp. 495–496.

[125] D. Bau, B. Zhou, A. Khosla, A. Oliva, A. Torralba, Network dissection: Quantifying interpretability of deep visual representations, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 6541–6549.

[126] H. Núñez, C. Angulo, A. Català, Rule extraction from support vector machines., in: European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN), 2002, pp. 107–112.

[127] H. Núñez, C. Angulo, A. Català, Rule-based learning systems for support vector machines, Neural Processing Letters 24 (1) (2006) 1–18.

[128] M. Kearns, S. Neel, A. Roth, Z. S. Wu, Preventing fairness gerrymandering: Auditing and learning for subgroup fairness (2017). arXiv:1711.05144.

[129] E. Akyol, C. Langbort, T. Basar, Price of transparency in strategic machine learning (2016). arXiv:1610.08210.

[130] D. Erhan, A. Courville, Y. Bengio, Understanding representations learned in deep architectures, Department dInformatique et Recherche Operationnelle, University of Montreal, QC, Canada, Tech. Rep 1355 (2010) 1.

[131] Y. Zhang, B. Wallace, A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification (2015). arXiv:1510.03820.

[132] J. R. Quinlan, Simplifying decision trees, International journal of man-machine studies 27 (3) (1987) 221–234.

[133] Y. Zhou, G. Hooker, Interpreting models via single tree approximation (2016). arXiv:1610.09036.

[134] A. Navia-Vázquez, E. Parrado-Hernández, Support vector machine interpretation, Neurocomputing 69 (13-15) (2006) 1754–1759.

[135] J. J. Thiagarajan, B. Kailkhura, P. Sattigeri, K. N. Ramamurthy, Treeview: Peeking into deep neural networks via feature-space partitioning (2016). arXiv:1611.07429.

[136] M. D. Zeiler, R. Fergus, Visualizing and understanding convolutional networks, in: European conference on computer vision, Springer, 2014, pp. 818–833.

[137] A. Mahendran, A. Vedaldi, Understanding deep image representations by inverting them, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 5188–5196.

[138] J. Wagner, J. M. Kohler, T. Gindele, L. Hetzel, J. T. Wiedemer, S. Behnke, Interpretable and fine-grained visual explanations for convolutional neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 9097–9107.

[139] A. Kanehira, T. Harada, Learning to explain with complemental examples, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 8603–8611.

[140] D. W. Apley, Visualizing the effects of predictor variables in black box supervised learning models (2016). arXiv:1612.08468.

[141] M. Staniak, P. Biecek, Explanations of Model Predictions with live and breakDown Packages, The R Journal 10 (2) (2018) 395–409.

[142] M. D. Zeiler, D. Krishnan, G. W. Taylor, R. Fergus, Deconvolutional networks., in: CVPR, Vol. 10, 2010, p. 7.

[143] J. T. Springenberg, A. Dosovitskiy, T. Brox, M. Riedmiller, Striving for simplicity: The all convolutional net (2014). arXiv:1412.6806.

[144] B. Kim, M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, F. Viegas, R. Sayres, Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV) (2017). arXiv:1711.11279.

[145] A. Polino, R. Pascanu, D. Alistarh, Model compression via distillation and quantization (2018). arXiv:1802.05668.

[146] W. J. Murdoch, A. Szlam, Automatic rule extraction from long short term memory networks (2017). arXiv:1702.02540.

[147] M. W. Craven, J. W. Shavlik, Using sampling and queries to extract rules from trained neural networks, in: Machine learning proceedings 1994, Elsevier, 1994, pp. 37–45.

[148] A. D. Arbatli, H. L. Akin, Rule extraction from trained neural networks using genetic algorithms, Nonlinear Analysis: Theory, Methods & Applications 30 (3) (1997) 1639–1648.

[149] U. Johansson, L. Niklasson, Evolving decision trees using oracle guides, in: 2009 IEEE Symposium on Computational Intelligence and Data Mining, IEEE, 2009, pp. 238–244.

[150] T. Lei, R. Barzilay, T. Jaakkola, Rationalizing neural predictions (2016). arXiv:1606.04155.

[151] A. Radford, R. Jozefowicz, I. Sutskever, Learning to generate reviews and discovering sentiment (2017). arXiv:1704.01444.

[152] R. R. Selvaraju, A. Das, R. Vedantam, M. Cogswell, D. Parikh, D. Batra, Grad-CAM: Why did you say that? (2016).

[153] R. Shwartz-Ziv, N. Tishby, Opening the black box of deep neural networks via information (2017). arXiv:1703.00810.

[154] J. Yosinski, J. Clune, A. Nguyen, T. Fuchs, H. Lipson, Understanding neural networks through deep visualization (2015). arXiv:1506.06579.

[155] P. E. Pope, S. Kolouri, M. Rostami, C. E. Martin, H. Hoffmann, Explainability methods for graph convolutional neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 10772–10781.

[156] P. Gajane, M. Pechenizkiy, On formalizing fairness in prediction with machine learning (2017). arXiv:1710.03184.

[157] C. Dwork, C. Ilvento, Composition of fairsystems (2018). arXiv:1806.06122.

[158] S. Barocas, M. Hardt, A. Narayanan, Fairness and Machine Learning, fairmlbook.org, 2019, `http://www.fairmlbook.org`.

[159] H.-X. Wang, L. Fratiglioni, G. B. Frisoni, M. Viitanen, B. Winblad, Smoking and the occurence of alzheimer's disease: Cross-sectional and longitudinal data in a population-based study, American journal of epidemiology 149 (7) (1999) 640–644.

[160] P. Rani, C. Liu, N. Sarkar, E. Vanman, An empirical study of machine learning techniques for affect recognition in human–robot interaction, Pattern Analysis and Applications 9 (1) (2006) 58–69.

[161] J. Pearl, Causality, Cambridge university press, 2009.

[162] M. Kuhn, K. Johnson, Applied predictive modeling, Vol. 26, Springer, 2013.

[163] G. James, D. Witten, T. Hastie, R. Tibshirani, An introduction to statistical learning, Vol. 112, Springer, 2013.

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

[165] D. Ruppert, Robust statistics: The approach based on influence functions, Taylor & Francis, 1987.

[166] S. Basu, K. Kumbier, J. B. Brown, B. Yu, Iterative random forests to discover predictive and stable high-order interactions, Proceedings of the National Academy of Sciences 115 (8) (2018) 1943–1948.

[167] B. Yu, et al., Stability, Bernoulli 19 (4) (2013) 1484–1500.

[168] K. Burns, L. A. Hendricks, K. Saenko, T. Darrell, A. Rohrbach, Women also Snowboard: Overcoming Bias in Captioning Models (2018). arXiv:1803.09797.

[169] A. Bennetot, J.-L. Laurent, R. Chatila, N. Díaz-Rodríguez, Towards explainable neural-symbolic visual reasoning, in: NeSy Workshop IJCAI 2019, Macau, China, 2019.

[170] R. Tibshirani, Regression shrinkage and selection via the lasso, Journal of the Royal Statistical Society: Series B (Methodological) 58 (1) (1996) 267–288.

[171] Y. Lou, R. Caruana, J. Gehrke, Intelligible models for classification and regression, in: ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2012, pp. 150–158.

[172] K. Kawaguchi, Deep learning without poor local minima, in: Advances in neural information processing systems, 2016, pp. 586–594.

[173] A. Datta, S. Sen, Y. Zick, Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems, in: 2016 IEEE symposium on security and privacy (SP), IEEE, 2016, pp. 598–617.

[174] Z. Bursac, C. H. Gauss, D. K. Williams, D. W. Hosmer, Purposeful selection of variables in logistic regression, Source code for biology and medicine 3 (1) (2008) 17.

[175] J. Jaccard, Interaction effects in logistic regression: Quantitative applications in the social sciences, Sage Thousand Oaks, CA, 2001.

[176] D. W. Hosmer Jr, S. Lemeshow, R. X. Sturdivant, Applied logistic regression, Vol. 398, John Wiley & Sons, 2013.

[177] C.-Y. J. Peng, K. L. Lee, G. M. Ingersoll, An introduction to logistic regression analysis and reporting, The journal of educational research 96 (1) (2002) 3–14.

[178] U. Hoffrage, G. Gigerenzer, Using natural frequencies to improve diagnostic inferences, Academic medicine 73 (5) (1998) 538–540.

[179] C. Mood, Logistic regression: Why we cannot do what we think we can do, and what we can do about it, European sociological review 26 (1) (2010) 67–82.

[180] H. Laurent, R. L. Rivest, Constructing optimal binary decision trees is Np-complete, Information processing letters 5 (1) (1976) 15–17.

[181] P. E. Utgoff, Incremental induction of decision trees, Machine learning 4 (2) (1989) 161–186.

[182] J. R. Quinlan, Induction of decision trees, Machine learning 1 (1) (1986) 81–106.

[183] L. Rokach, O. Z. Maimon, Data mining with decision trees: theory and applications, Vol. 69, World scientific, 2014.

[184] S. Rovnyak, S. Kretsinger, J. Thorp, D. Brown, Decision trees for real-time transient stability prediction, IEEE Transactions on Power Systems 9 (3) (1994) 1417–1426.

[185] H. Nefeslioglu, E. Sezer, C. Gokceoglu, A. Bozkir, T. Duman, Assessment of landslide susceptibility by decision trees in the metropolitan area of istanbul, turkey, Mathematical Problems in Engineering 2010 (2010) Article ID 901095.

[186] S. B. Imandoust, M. Bolandraftar, Application of k-nearest neighbor (knn) approach for predicting economic events: Theoretical background, International Journal of Engineering Research and Applications 3 (5) (2013) 605–610.

[187] L. Li, D. M. Umbach, P. Terry, J. A. Taylor, Application of the GA/KNN method to SELDI proteomics data, Bioinformatics 20 (10) (2004) 1638–1640.

[188] G. Guo, H. Wang, D. Bell, Y. Bi, K. Greer, An KNN model-based approach and its application in text categorization, in: International Conference on Intelligent Text Processing and Computational Linguistics, Springer, 2004, pp. 559–570.

[189] S. Jiang, G. Pang, M. Wu, L. Kuang, An improved k-nearest-neighbor algorithm for text categorization, Expert Systems with Applications 39 (1) (2012) 1503–1509.

[190] U. Johansson, R. König, L. Niklasson, The truth is in there-rule extraction from opaque models using genetic programming., in: FLAIRS Conference, Miami Beach, FL, 2004, pp. 658–663.

[191] J. R. Quinlan, Generating production rules from decision trees., in: ijcai, Vol. 87, Citeseer, 1987, pp. 304–307.

[192] P. Langley, H. A. Simon, Applications of machine learning and rule induction, Communications of the ACM 38 (11) (1995) 54–64.

[193] D. Berg, Bankruptcy prediction by generalized additive models, Applied Stochastic Models in Business and Industry 23 (2) (2007) 129–143.

[194] R. Calabrese, et al., Estimating bank loans loss given default by generalized additive models, UCD Geary Institute Discussion Paper Series, WP2012/24 (2012).

[195] P. Taylan, G.-W. Weber, A. Beck, New approaches to regression by generalized additive models and continuous optimization for modern applications in finance, science and technology, Optimization 56 (5-6) (2007) 675–698.

[196] H. Murase, H. Nagashima, S. Yonezaki, R. Matsukura, T. Kitakado, Application of a generalized additive model (GAM) to reveal relationships between environmental factors and distributions of pelagic fish and krill: a case study in sendai bay, Japan, ICES Journal of Marine Science 66 (6) (2009) 1417–1424.

[197] N. Tomić, S. Božić, A modified geosite assessment model (M-GAM) and its application on the lazar canyon area (serbia), International journal of environmental research 8 (4) (2014) 1041–1052.

[198] A. Guisan, T. C. Edwards Jr, T. Hastie, Generalized linear and generalized additive models in studies of species distributions: setting the scene, Ecological Modelling 157 (2-3) (2002) 89–100.

[199] P. Rothery, D. B. Roy, Application of generalized additive models to butterfly transect count data, Journal of Applied Statistics 28 (7) (2001) 897–909.

[200] A. Pierrot, Y. Goude, Short-term electricity load forecasting with generalized additive models, in: 16th Intelligent System Applications to Power Systems Conference, ISAP 2011, IEEE, 2011, pp. 410–415.

[201] T. L. Griffiths, C. Kemp, J. B. Tenenbaum, Bayesian models of cognition. (4 2008). doi:10.1184/R1/6613682.v1.
URL https://kilthub.cmu.edu/articles/Bayesian_models_of_cognition/6613682

[202] B. H. Neelon, A. J. O'Malley, S.-L. T. Normand, A bayesian model for repeated measures zero-inflated count data with application to outpatient psychiatric service use, Statistical modelling 10 (4) (2010) 421–439.

[203] M. McAllister, G. Kirkwood, Bayesian stock assessment: a review and example application using the logistic model, ICES Journal of Marine Science 55 (6) (1998) 1031–1060.

[204] G. Synnaeve, P. Bessiere, A bayesian model for opening prediction in RTS games with application to starcraft, in: Computational Intelligence and Games (CIG), 2011 IEEE Conference on, IEEE, 2011, pp. 281–288.

[205] S.-K. Min, D. Simonis, A. Hense, Probabilistic climate change predictions applying bayesian model averaging, Philosophical transactions of the royal society of london a: mathematical, physical and engineering sciences 365 (1857) (2007) 2103–2116.

[206] G. Koop, D. J. Poirier, J. L. Tobias, Bayesian econometric methods, Cambridge University Press, 2007.

[207] A. R. Cassandra, L. P. Kaelbling, J. A. Kurien, Acting under uncertainty: Discrete bayesian models for mobile-robot navigation, in: Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems. IROS'96, Vol. 2, IEEE, 1996, pp. 963–972.

[208] H. A. Chipman, E. I. George, R. E. McCulloch, Bayesian cart model search, Journal of the American Statistical Association 93 (443) (1998) 935–948.

[209] B. Kim, C. Rudin, J. A. Shah, The bayesian case model: A generative approach for case-based reasoning and prototype classification, in: Advances in Neural Information Processing Systems, 2014, pp. 1952–1960.

[210] B. Kim, R. Khanna, O. O. Koyejo, Examples are not enough, learn to criticize! criticism for interpretability, in: Advances in Neural Information Processing Systems, 2016, pp. 2280–2288.

[211] U. Johansson, L. Niklasson, R. König, Accuracy vs. comprehensibility in data mining models, in: Proceedings of the seventh international conference on information fusion, Vol. 1, 2004, pp. 295–300.

[212] R. Konig, U. Johansson, L. Niklasson, G-rex: A versatile framework for evolutionary data mining, in: 2008 IEEE International Conference on Data Mining Workshops, IEEE, 2008, pp. 971–974.

[213] H. Lakkaraju, E. Kamar, R. Caruana, J. Leskovec, Interpretable & explorable approximations of black box models (2017). arXiv:1707.01154.

[214] S. Mishra, B. L. Sturm, S. Dixon, Local interpretable model-agnostic explanations for music content analysis., in: ISMIR, 2017, pp. 537–543.

[215] G. Su, D. Wei, K. R. Varshney, D. M. Malioutov, Interpretable two-level boolean rule learning for classification (2015). arXiv:1511.07361.

[216] M. T. Ribeiro, S. Singh, C. Guestrin, Nothing else matters: Model-agnostic explanations by identifying prediction invariance (2016). arXiv:1611.05817.

[217] M. W. Craven, Extracting comprehensible models from trained neural networks, Ph.D. thesis, aAI9700774 (1996).

[218] O. Bastani, C. Kim, H. Bastani, Interpretability via model extraction (2017). arXiv:1706.09773.

[219] G. Hooker, Discovering additive structure in black box functions, in: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, 2004, pp. 575–580.

[220] P. Adler, C. Falk, S. A. Friedler, T. Nix, G. Rybeck, C. Scheidegger, B. Smith, S. Venkatasubramanian, Auditing black-box models for indirect influence, Knowledge and Information Systems 54 (1) (2018) 95–122.

[221] P. W. Koh, P. Liang, Understanding black-box predictions via influence functions, in: Proceedings of the 34th International Conference on Machine Learning-Volume 70, JMLR. org, 2017, pp. 1885–1894.

[222] P. Cortez, M. J. Embrechts, Opening black box data mining models using sensitivity analysis, in: 2011 IEEE Symposium on Computational Intelligence and Data Mining (CIDM), IEEE, 2011, pp. 341–348.

[223] P. Cortez, M. J. Embrechts, Using sensitivity analysis and visualization techniques to open black box data mining models, Information Sciences 225 (2013) 1–17.

[224] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, in: Advances in Neural Information Processing Systems, 2017, pp. 4765–4774.

[225] I. Kononenko, et al., An efficient explanation of individual classifications using game theory, Journal of Machine Learning Research 11 (Jan) (2010) 1–18.

[226] H. Chen, S. Lundberg, S.-I. Lee, Explaining models by propagating shapley values of local components (2019). arXiv:arXiv:1911.11888.

[227] P. Dabkowski, Y. Gal, Real time image saliency for black box classifiers, in: Advances in Neural Information Processing Systems, 2017, pp. 6967–6976.

[228] A. Henelius, K. Puolamäki, H. Boström, L. Asker, P. Papapetrou, A peek into the black box: exploring classifiers by randomization, Data mining and knowledge discovery 28 (5-6) (2014) 1503–1529.

[229] J. Moeyersoms, B. d'Alessandro, F. Provost, D. Martens, Explaining classification models built on high-dimensional sparse data (2016). arXiv:1607.06280.

[230] D. Baehrens, T. Schroeter, S. Harmeling, M. Kawanabe, K. Hansen, K.-R. MÃžller, How to explain individual classification decisions, Journal of Machine Learning Research 11 (Jun) (2010) 1803–1831.

[231] J. Adebayo, L. Kagal, Iterative orthogonal feature projection for diagnosing bias in black-box models (2016). arXiv:1611.04967.

[232] R. Guidotti, A. Monreale, S. Ruggieri, D. Pedreschi, F. Turini, F. Giannotti, Local rule-based explanations of black box decision systems (2018). arXiv:1805.10820.

[233] S. Krishnan, E. Wu, Palm: Machine learning explanations for iterative debugging, in: Proceedings of the 2nd Workshop on Human-In-the-Loop Data Analytics, ACM, 2017, p. 4.

[234] M. Robnik-Šikonja, I. Kononenko, Explaining classifications for individual instances, IEEE Transactions on Knowledge and Data Engineering 20 (5) (2008) 589–600.

[235] M. T. Ribeiro, S. Singh, C. Guestrin, Anchors: High-precision model-agnostic explanations, in: AAAI Conference on Artificial Intelligence, 2018, pp. 1527–1535.

[236] D. Martens, F. Provost, Explaining data-driven document classifications, MIS Quarterly 38 (1) (2014) 73–100.

[237] D. Chen, S. P. Fraiberger, R. Moakler, F. Provost, Enhancing transparency and control when drawing data-driven inferences about individuals, Big data 5 (3) (2017) 197–212.

[238] A. Goldstein, A. Kapelner, J. Bleich, E. Pitkin, Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation, Journal of Computational and Graphical Statistics 24 (1) (2015) 44–65.

[239] G. Casalicchio, C. Molnar, B. Bischl, Visualizing the feature importance for black box models, in: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer, 2018, pp. 655–670.

[240] G. Tolomei, F. Silvestri, A. Haines, M. Lalmas, Interpretable predictions of tree-based ensembles via actionable feature tweaking, in: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2017, pp. 465–474.

[241] L. Auret, C. Aldrich, Interpretation of nonlinear relationships between process variables by use of random forests, Minerals Engineering 35 (2012) 27–42.

[242] N. F. Rajani, R. Mooney, Stacking with auxiliary features for visual question answering, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), 2018, pp. 2217–2226.

[243] N. F. Rajani, R. J. Mooney, Ensembling visual explanations, in: Explainable and Interpretable Models in Computer Vision and Machine Learning, Springer, 2018, pp. 155–172.

[244] H. Núñez, C. Angulo, A. Català, Rule-based learning systems for support vector machines, Neural Processing Letters 24 (1) (2006) 1–18.

[245] Z. Chen, J. Li, L. Wei, A multiple kernel support vector machine scheme for feature selection and rule extraction from gene expression data of cancer tissue, Artificial Intelligence in Medicine 41 (2) (2007) 161–175.

[246] H. Núñez, C. Angulo, A. Català, Support vector machines with symbolic interpretation, in: VII Brazilian Symposium on Neural Networks, 2002. SBRN 2002. Proceedings., IEEE, 2002, pp. 142–147.

[247] P. Sollich, Bayesian methods for support vector machines: Evidence and predictive class probabilities, Machine learning 46 (1-3) (2002) 21–52.

[248] P. Sollich, Probabilistic methods for support vector machines, in: Advances in neural information processing systems, 2000, pp. 349–355.

[249] W. Landecker, M. D. Thomure, L. M. Bettencourt, M. Mitchell, G. T. Kenyon, S. P. Brumby, Interpreting individual classifications of hierarchical networks, in: 2013 IEEE Symposium on Computational Intelligence and Data Mining (CIDM), IEEE, 2013, pp. 32–38.

[250] A. Jakulin, M. Možina, J. Demšar, I. Bratko, B. Zupan, Nomograms for visualizing support vector machines, in: Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining, ACM, 2005, pp. 108–117.

[251] L. Fu, Rule generation from neural networks, IEEE Transactions on Systems, Man, and Cybernetics 24 (8) (1994) 1114–1124.

[252] G. G. Towell, J. W. Shavlik, Extracting refined rules from knowledge-based neural networks, Machine Learning 13 (1) (1993) 71–101.

[253] S. Thrun, Extracting rules from artificial neural networks with distributed representations, in: Proceedings of the 7th International Conference on Neural Information Processing Systems, NIPS'94, 1994, pp. 505–512.

[254] R. Setiono, W. K. Leow, FERNN: An algorithm for fast extraction of rules from neural networks, Applied Intelligence 12 (1) (2000) 15–25.

[255] I. A. Taha, J. Ghosh, Symbolic interpretation of artificial neural networks, IEEE Transactions on Knowledge and Data Engineering 11 (3) (1999) 448–463.

[256] H. Tsukimoto, Extracting rules from trained neural networks, IEEE Transactions on Neural Networks 11 (2) (2000) 377–389.

[257] J. R. Zilke, E. L. Mencía, F. Janssen, Deepred–rule extraction from deep neural networks, in: International Conference on Discovery Science, Springer, 2016, pp. 457–473.

[258] G. P. J. Schmitz, C. Aldrich, F. S. Gouws, ANN-DT: an algorithm for extraction of decision trees from artificial neural networks, IEEE Transactions on Neural Networks 10 (6) (1999) 1392–1401.

[259] M. Sato, H. Tsukimoto, Rule extraction from neural networks via decision tree induction, in: IJCNN'01. International Joint Conference on Neural Networks. Proceedings (Cat. No. 01CH37222), Vol. 3, IEEE, 2001, pp. 1870–1875.

[260] R. Féraud, F. Clérot, A methodology to explain neural network classification, Neural networks 15 (2) (2002) 237–246.

[261] A. Shrikumar, P. Greenside, A. Kundaje, Learning Important Features Through Propagating Activation Differences (2017). arXiv:1704.02685.

[262] M. Sundararajan, A. Taly, Q. Yan, Axiomatic attribution for deep networks, in: International Conference on Machine Learning, Vol. 70, JMLR. org, 2017, pp. 3319–3328.

[263] J. Adebayo, J. Gilmer, I. Goodfellow, B. Kim, Local explanation methods for deep neural networks lack sensitivity to parameter values (2018). arXiv:1810.03307.

[264] N. Papernot, P. McDaniel, Deep k-nearest neighbors: Towards confident, interpretable and robust deep learning (2018). arXiv:1803.04765.

[265] J. Li, X. Chen, E. Hovy, D. Jurafsky, Visualizing and understanding neural models in NLP (2015). arXiv:1506.01066.

[266] S. Tan, K. C. Sim, M. Gales, Improving the interpretability of deep neural networks with stimulated learning, in: 2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), IEEE, 2015, pp. 617–623.

[267] L. Rieger, C. Singh, W. J. Murdoch, B. Yu, Interpretations are useful: penalizing explanations to align neural networks with prior knowledge (2019). arXiv:arXiv:1909.13584.

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

[269] Y. Li, J. Yosinski, J. Clune, H. Lipson, J. E. Hopcroft, Convergent learning: Do different neural networks learn the same representations?, in: ICLR, 2016.

[270] M. Liu, J. Shi, Z. Li, C. Li, J. Zhu, S. Liu, Towards better analysis of deep convolutional neural networks, IEEE transactions on visualization and computer graphics 23 (1) (2016) 91–100.

[271] Y. Goyal, A. Mohapatra, D. Parikh, D. Batra, Towards transparent AI systems: Interpreting visual question answering models (2016). arXiv:1608.08974.

[272] K. Simonyan, A. Vedaldi, A. Zisserman, Deep inside convolutional networks: Visualising image classification models and saliency maps (2013). arXiv:1312.6034.

[273] A. Nguyen, J. Yosinski, J. Clune, Deep neural networks are easily fooled: High confidence predictions for unrecognizable images, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 427–436.

[274] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, T. Darrell, Long-term recurrent convolutional networks for visual recognition and description, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 2625–2634.

[275] M. Lin, Q. Chen, S. Yan, Network in network (2013). arXiv:1312.4400.

[276] L. A. Hendricks, Z. Akata, M. Rohrbach, J. Donahue, B. Schiele, T. Darrell, Generating Visual Explanations (2016). arXiv:1603.08507.

[277] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, X. Tang, Residual attention network for image classification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 3156–3164.

[278] T. Xiao, Y. Xu, K. Yang, J. Zhang, Y. Peng, Z. Zhang, The application of two-level attention models in deep convolutional neural network for fine-grained image classification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 842–850.

[279] Q. Zhang, R. Cao, Y. Nian Wu, S.-C. Zhu, Growing Interpretable Part Graphs on ConvNets via Multi-Shot Learning (2016). arXiv:1611.04246.

[280] L. Arras, G. Montavon, K.-R. Müller, W. Samek, Explaining recurrent neural network predictions in sentiment analysis (2017). arXiv:1706.07206.

[281] A. Karpathy, J. Johnson, L. Fei-Fei, Visualizing and understanding recurrent networks (2015). arXiv:1506.02078.

[282] J. Clos, N. Wiratunga, S. Massie, Towards explainable text classification by jointly learning lexicon and modifier terms, in: IJCAI-17 Workshop on Explainable AI (XAI), 2017, p. 19.

[283] S. Wisdom, T. Powers, J. Pitton, L. Atlas, Interpretable recurrent neural networks using sequential sparse recovery (2016). arXiv:1611.07252.

[284] V. Krakovna, F. Doshi-Velez, Increasing the interpretability of recurrent neural networks using hidden markov models (2016). arXiv:1606.05320.

[285] E. Choi, M. T. Bahadori, J. Sun, J. Kulas, A. Schuetz, W. Stewart, Retain: An interpretable predictive model for healthcare using reverse time attention mechanism, in: Advances in Neural Information Processing Systems, 2016, pp. 3504–3512.

[286] L. Breiman, Classification and regression trees, Routledge, 2017.

[287] A. Lucic, H. Haned, M. de Rijke, Explaining predictions from tree-based boosting ensembles (2019). arXiv:arXiv:1907.02582.

[288] S. M. Lundberg, G. G. Erion, S.-I. Lee, Consistent individualized feature attribution for tree ensembles (2018). arXiv:arXiv:1802.03888.

[289] C. Buciluǎ, R. Caruana, A. Niculescu-Mizil, Model compression, in: ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2006, pp. 535–541.

[290] R. Traoré, H. Caselles-Dupré, T. Lesort, T. Sun, G. Cai, N. D. Rodríguez, D. Filliat, DisCoRL: Continual reinforcement learning via policy distillation (2019). arXiv:1907.05855.

[291] M. D. Zeiler, G. W. Taylor, R. Fergus, et al., Adaptive deconvolutional networks for mid and high level feature learning., in: ICCV, Vol. 1, 2011, p. 6.

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

[293] C. Olah, A. Mordvintsev, L. Schubert, Feature visualization., DistillHttps://distill.pub/2017/feature-visualization (2017). doi:10.23915/distill.00007.

[294] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, B. Kim, Sanity checks for saliency maps, in: Advances in Neural Information Processing Systems, 2018, pp. 9505–9515.

[295] C. Olah, A. Satyanarayan, I. Johnson, S. Carter, L. Schubert, K. Ye, A. Mordvintsev, The building blocks of interpretability, Distill (2018).
URL `https://distill.pub/2018/building-blocks/`

[296] Z. Che, S. Purushotham, R. Khemani, Y. Liu, Distilling knowledge from deep networks with applications to healthcare domain (2015). arXiv:1512.03542.

[297] I. Donadello, L. Serafini, A. D. Garcez, Logic tensor networks for semantic image interpretation, Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI (2017) 1596–1602.

[298] I. Donadello, Semantic image interpretation-integration of numerical data and logical knowledge for cognitive vision, Ph.D. thesis, University of Trento (2018).

[299] A. S. d'Avila Garcez, M. Gori, L. C. Lamb, L. Serafini, M. Spranger, S. N. Tran, Neural-symbolic computing: An effective methodology for principled integration of machine learning and reasoning (2019). arXiv:1905.06088.

[300] R. Manhaeve, S. Dumancic, A. Kimmig, T. Demeester, L. De Raedt, DeepProbLog: Neural probabilistic logic programming, in: Advances in Neural Information Processing Systems 31, 2018, pp. 3749–3759.

[301] I. Donadello, M. Dragoni, C. Eccher, Persuasive explanation of reasoning inferences on dietary data, in: First Workshop on Semantic Explainability @ ISWC 2019, 2019.

[302] R. G. Krishnan, U. Shalit, D. Sontag, Deep Kalman Filters (2015). arXiv:1511.05121.

[303] M. Karl, M. Soelch, J. Bayer, P. van der Smagt, Deep Variational Bayes Filters: Unsupervised Learning of State Space Models from Raw Data (2016). arXiv:1605.06432.

[304] M. J. Johnson, D. K. Duvenaud, A. Wiltschko, R. P. Adams, S. R. Datta, Composing graphical models with neural networks for structured representations and fast inference, in: Advances in Neural Information Processing Systems 29, 2016, pp. 2946–2954.

[305] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, P. H. Torr, Conditional random fields as recurrent neural networks, in: Proceedings of the IEEE international conference on computer vision, 2015, pp. 1529–1537.

[306] N. Narodytska, A. Ignatiev, F. Pereira, J. Marques-Silva, Learning optimal decision trees with SAT, in: Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18, 2018, pp. 1362–1368.

[307] O. Loyola-González, Black-box vs. white-box: Understanding their advantages and weaknesses from a practical point of view, IEEE Access 7 (2019) 154096–154113.

[308] F. Petroni, T. Rocktäschel, P. Lewis, A. Bakhtin, Y. Wu, A. H. Miller, S. Riedel, Language models as knowledge bases? (2019). arXiv:1909.01066.

[309] K. Bollacker, N. Díaz-Rodríguez, X. Li, Extending knowledge graphs with subjective influence networks for personalized fashion, in: E. Portmann, M. E. Tabacchi, R. Seising, A. Habenstein (Eds.), Designing Cognitive Cities, Springer International Publishing, 2019, pp. 203–233.

[310] W. Shang, A. Trott, S. Zheng, C. Xiong, R. Socher, Learning world graphs to accelerate hierarchical reinforcement learning (2019). arXiv:1907.00664.

[311] M. Zolotas, Y. Demiris, Towards explainable shared control using augmented reality, 2019.

[312] M. Garnelo, K. Arulkumaran, M. Shanahan, Towards deep symbolic reinforcement learning (2016). arXiv:1609.05518.

[313] V. Bellini, A. Schiavone, T. Di Noia, A. Ragone, E. Di Sciascio, Knowledge-aware autoencoders for explainable recommender systems, in: Proceedings of the 3rd Workshop on Deep Learning for Recommender Systems, DLRS 2018, 2018, pp. 24–31.

[314] C.-Z. A. Huang, A. Vaswani, J. Uszkoreit, N. Shazeer, C. Hawthorne, A. M. Dai, M. D. Hoffman, D. Eck, Music transformer: Generating music with long-term structure (2018). arXiv:1809.04281.

[315] M. Cornia, L. Baraldi, R. Cucchiara, Smart: Training shallow memory-aware transformers for robotic explainability (2019). arXiv:1910.02974.

[316] A. Aamodt, E. Plaza, Case-based reasoning: Foundational issues, Methodological Variations, and System Approaches 7 (1) (1994) 39–59.

[317] R. Caruana, Case-based explanation for artificial neural nets, in: Artificial Neural Networks in Medicine and Biology, Proceedings of the ANNIMAB-1 Conference, 2000, pp. 303–308.

[318] M. T. Keane, E. M. Kenny, The Twin-System Approach as One Generic Solution for XAI: An Overview of ANN-CBR Twins for Explaining Deep Learning (2019). arXiv:1905.08069.

[319] T. Hailesilassie, Rule extraction algorithm for deep neural networks: A review (2016). arXiv:1610.05267.

[320] J. M. Benitez, J. L. Castro, I. Requena, Are artificial neural networks black boxes?, IEEE Trans. Neural Networks 8 (5) (1997) 1156–1164.

[321] U. Johansson, R. König, L. Niklasson, Automatically balancing accuracy and comprehensibility in predictive modeling, in: Proceedings of the 8th International Conference on Information Fusion, Vol. 2, 2005, p. 7 pp.

[322] D. Smilkov, N. Thorat, B. Kim, F. Viégas, M. Wattenberg, SmoothGrad: removing noise by adding noise (2017). arXiv:1706.03825.

[323] M. Ancona, E. Ceolini, C. Öztireli, M. Gross, Towards better understanding of gradient-based attribution methods for Deep Neural Networks (2017). arXiv:1711.06104.

[324] J. Yosinski, J. Clune, Y. Bengio, H. Lipson, How transferable are features in deep neural networks? (2014). arXiv:1411.1792.

[325] A. Sharif Razavian, H. Azizpour, J. Sullivan, S. Carlsson, CNN Features off-the-shelf: an Astounding Baseline for Recognition (2014). arXiv:1403.6382.

[326] S. Du, H. Guo, A. Simpson, Self-driving car steering angle prediction based on image recognition, Tech. rep., Technical Report, Stanford University (2017).

[327] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, A. Torralba, Object Detectors Emerge in Deep Scene CNNs (2014). arXiv:1412.6856.

[328] Y. Zhang, X. Chen, Explainable Recommendation: A Survey and New Perspectives (2018). arXiv:1804.11192.

[329] J. Frankle, M. Carbin, The Lottery Ticket Hypothesis: Finding Sparse, Trainable Neural Networks (2018). arXiv:1803.03635.

[330] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention Is All You Need (2017). arXiv:1706.03762.

[331] J. Lu, J. Yang, D. Batra, D. Parikh, Hierarchical question-image co-attention for visual question answering, in: Proceedings of the 30th International Conference on Neural Information Processing Systems, NIPS'16, 2016, pp. 289–297.

[332] A. Das, H. Agrawal, C. L. Zitnick, D. Parikh, D. Batra, Human Attention in Visual Question Answering: Do Humans and Deep Networks Look at the Same Regions? (2016). arXiv:1606.03556.

[333] D. Huk Park, L. A. Hendricks, Z. Akata, A. Rohrbach, B. Schiele, T. Darrell, M. Rohrbach, Multimodal Explanations: Justifying Decisions and Pointing to the Evidence (2018). arXiv:1802.08129.

[334] A. Slavin Ross, M. C. Hughes, F. Doshi-Velez, Right for the Right Reasons: Training Differentiable Models by Constraining their Explanations (2017). arXiv:1703.03717.

[335] I. T. Jolliffe, Principal Component Analysis and Factor Analysis, Springer New York, 1986, pp. 115–128.

[336] A. Hyvärinen, E. Oja, Oja, e.: Independent component analysis: Algorithms and applications. neural networks 13(4-5), 411-430, Neural networks 13 (2000) 411–430.

[337] M. W. Berry, M. Browne, A. N. Langville, V. P. Pauca, R. J. Plemmons, Algorithms and applications for approximate nonnegative matrix factorization, Computational Statistics & Data Analysis 52 (2007) 155–173.

[338] D. P. Kingma, M. Welling, Auto-Encoding Variational Bayes (2013). arXiv:1312.6114.

[339] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. M. Botvinick, S. Mohamed, A. Lerchner, beta-vae: Learning basic visual concepts with a constrained variational framework, in: ICLR, 2017.

[340] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, P. Abbeel, InfoGAN: Interpretable Representation Learning by Information Maximizing Generative Adversarial Nets (2016). arXiv:1606.03657.

[341] Q. Zhang, Y. Yang, Y. Liu, Y. Nian Wu, S.-C. Zhu, Unsupervised Learning of Neural Networks to Explain Neural Networks (2018). arXiv:1805.07468.

[342] S. Sabour, N. Frosst, G. E Hinton, Dynamic Routing Between Capsules (2017). arXiv:1710.09829.

[343] A. Agrawal, J. Lu, S. Antol, M. Mitchell, C. L. Zitnick, D. Batra, D. Parikh, VQA: Visual Question Answering (2015). arXiv:1505.00468.

[344] A. Fukui, D. Huk Park, D. Yang, A. Rohrbach, T. Darrell, M. Rohrbach, Multimodal Compact Bilinear Pooling for Visual Question Answering and Visual Grounding (2016). arXiv:1606.01847.

[345] D. Bouchacourt, L. Denoyer, EDUCE: explaining model decisions through unsupervised concepts extraction (2019). arXiv:1905.11852.

[346] C. Hofer, M. Denker, S. Ducasse, Design and Implementation of a Backward-In-Time Debugger, in: NODe 2006, Vol. P-88 of Lecture Notes in Informatics, 2006, pp. 17–32.

[347] C. Rudin, Please stop explaining black box models for high stakes decisions (2018). arXiv:1811.10154.

[348] A. Diez-Olivan, J. Del Ser, D. Galar, B. Sierra, Data fusion and machine learning for industrial prognosis: Trends and perspectives towards Industry 4.0, Information Fusion 50 (2019) 92–111.

[349] R. R. Hoffman, S. T. Mueller, G. Klein, J. Litman, Metrics for explainable ai: Challenges and prospects (2018). arXiv:arXiv:1812.04608.

[350] S. Mohseni, N. Zarei, E. D. Ragan, A multidisciplinary survey and framework for design and evaluation of explainable ai systems (2018). arXiv:arXiv:1811.11839.

[351] R. M. J. Byrne, Counterfactuals in explainable artificial intelligence (XAI): Evidence from human reasoning, in: Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19, 2019, pp. 6276–6282.

[352] M. Garnelo, M. Shanahan, Reconciling deep learning with symbolic artificial intelligence: representing objects and relations, Current Opinion in Behavioral Sciences 29 (2019) 17–23.

[353] G. Marra, F. Giannini, M. Diligenti, M. Gori, Integrating learning and reasoning with deep logic models (2019). arXiv:1901.04195.

[354] K. Kelley, B. Clark, V. Brown, J. Sitzia, Good practice in the conduct and reporting of survey research, International Journal for Quality in Health Care 15 (3) (2003) 261–266.

[355] S. Wachter, B. Mittelstadt, L. Floridi, Why a right to explanation of automated decision-making does not exist in the general data protection regulation, International Data Privacy Law 7 (2) (2017) 76–99.

[356] T. Orekondy, B. Schiele, M. Fritz, Knockoff nets: Stealing functionality of black-box models (2018). arXiv:1812.02766.

[357] S. J. Oh, B. Schiele, M. Fritz, Towards reverse-engineering black-box neural networks, in: Explainable AI: Interpreting, Explaining and Visualizing Deep Learning, Springer, 2019, pp. 121–144.

[358] I. J. Goodfellow, J. Shlens, C. Szegedy, Explaining and harnessing adversarial examples (2014). arXiv:1412.6572.

[359] K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, C. Xiao, A. Prakash, T. Kohno, D. Song, Robust physical-world attacks on deep learning models (2017). arXiv:1707.08945.

[360] I. J. Goodfellow, N. Papernot, P. D. McDaniel, cleverhans v0.1: an adversarial machine learning library (2016). arXiv:1610.00768.

[361] H. Xiao, B. Biggio, B. Nelson, H. Xiao, C. Eckert, F. Roli, Support vector machines under adversarial label contamination, Neurocomputing 160 (C) (2015) 53–62.

[362] B. Biggio, I. Corona, D. Maiorca, B. Nelson, N. Šrndić, P. Laskov, G. Giacinto, F. Roli, Evasion attacks against machine learning at test time, in: Proceedings of the 2013th European Conference on Machine Learning and Knowledge Discovery in Databases - Volume Part III, ECMLPKDD'13, 2013, pp. 387–402.

[363] B. Biggio, I. Pillai, S. R. Bulò, D. Ariu, M. Pelillo, F. Roli, Is data clustering in adversarial settings secure? (2018). arXiv:1811.09982.

[364] Z. Pan, W. Yu, X. Yi, A. Khan, F. Yuan, Y. Zheng, Recent progress on generative adversarial networks (gans): A survey, IEEE Access 7 (2019) 36322–36333.

[365] D. Charte, F. Charte, S. García, M. J. del Jesus, F. Herrera, A practical tutorial on autoencoders for nonlinear feature fusion: Taxonomy, models, software and guidelines, Information Fusion 44 (2018) 78–96.

[366] C. F. Baumgartner, L. M. Koch, K. Can Tezcan, J. Xi Ang, E. Konukoglu, Visual feature attribution using wasserstein gans, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 8309–8319.

[367] C. Biffi, O. Oktay, G. Tarroni, W. Bai, A. De Marvao, G. Doumou, M. Rajchl, R. Bedair, S. Prasad, S. Cook, et al., Learning interpretable anatomical features through deep generative models: Application to cardiac remodeling, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2018, pp. 464–471.

[368] S. Liu, B. Kailkhura, D. Loveland, Y. Han, Generative counterfactual introspection for explainable deep learning (2019). arXiv:arXiv:1907.03077.

[369] K. R. Varshney, H. Alemzadeh, On the safety of machine learning: Cyber-physical systems, decision sciences, and data products, Big data 5 (3) (2017) 246–255.

[370] G. M. Weiss, Mining with rarity: a unifying framework, ACM Sigkdd Explorations Newsletter 6 (1) (2004) 7–19.

[371] J. Attenberg, P. Ipeirotis, F. Provost, Beat the machine: Challenging humans to find a predictive model's "unknown unknowns", Journal of Data and Information Quality (JDIQ) 6 (1) (2015) 1.

[372] G. Neff, A. Tanweer, B. Fiore-Gartland, L. Osburn, Critique and contribute: A practice-based framework for improving critical data studies and data science, Big data 5 (2) (2017) 85–97.

[373] A. Iliadis, F. Russo, Critical data studies: An introduction, Big Data & Society 3 (2) (2016) 2053951716674238.

[374] A. Karpatne, G. Atluri, J. H. Faghmous, M. Steinbach, A. Banerjee, A. Ganguly, S. Shekhar, N. Samatova, V. Kumar, Theory-guided data science: A new paradigm for scientific discovery from data, IEEE Transactions on Knowledge and Data Engineering 29 (10) (2017) 2318–2331.

[375] G. Hautier, C. C. Fischer, A. Jain, T. Mueller, G. Ceder, Finding nature's missing ternary oxide compounds using machine learning and density functional theory, Chemistry of Materials 22 (12) (2010) 3762–3767.

[376] C. C. Fischer, K. J. Tibbetts, D. Morgan, G. Ceder, Predicting crystal structure by merging data mining with quantum mechanics, Nature materials 5 (8) (2006) 641.

[377] S. Curtarolo, G. L. Hart, M. B. Nardelli, N. Mingo, S. Sanvito, O. Levy, The high-throughput highway to computational materials design, Nature materials 12 (3) (2013) 191.

[378] K. C. Wong, L. Wang, P. Shi, Active model with orthotropic hyperelastic material for cardiac image analysis, in: International Conference on Functional Imaging and Modeling of the Heart, Springer, 2009, pp. 229–238.

[379] J. Xu, J. L. Sapp, A. R. Dehaghani, F. Gao, M. Horacek, L. Wang, Robust transmural electrophysiological imaging: Integrating sparse and dynamic physiological models into ecg-based inference, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2015, pp. 519–527.

[380] T. Lesort, M. Seurin, X. Li, N. Díaz-Rodríguez, D. Filliat, Unsupervised state representation learning with robotic priors: a robustness benchmark (2017). arXiv:arXiv:1709.05185.

[381] J. Z. Leibo, Q. Liao, F. Anselmi, W. A. Freiwald, T. Poggio, View-tolerant face recognition and hebbian learning imply mirror-symmetric neural tuning to head orientation, Current Biology 27 (1) (2017) 62–67.

[382] F. Schrodt, J. Kattge, H. Shan, F. Fazayeli, J. Joswig, A. Banerjee, M. Reichstein, G. Bönisch, S. Díaz, J. Dickie, et al., Bhpmf–a hierarchical bayesian approach to gap-filling and trait prediction for macroecology and functional biogeography, Global Ecology and Biogeography 24 (12) (2015) 1510–1521.

[383] D. Leslie, Understanding artificial intelligence ethics and safety (2019). arXiv:arXiv:1906.05684, doi:10.5281/zenodo.3240529.

[384] C. Rudin, Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead (2018). arXiv:arXiv:1811.10154.

[385] J. Fjeld, H. Hilligoss, N. Achten, M. L. Daniel, J. Feldman, S. Kagay, Principled artificial intelligence: A map of ethical and rights-based approaches (2019).
URL `https://ai-hr.cyber.harvard.edu/images/primp-viz.pdf`

[386] R. Benjamins, A. Barbado, D. Sierra, Responsible AI by design (2019). arXiv:1909.12838.

[387] United-Nations, Transforming our world: the 2030 agenda for sustainable development, Tech. rep., eSocialSciences (2015).
URL `https://EconPapers.repec.org/RePEc:ess:wpaper:id:7559`

[388] G. D. Hager, A. Drobnis, F. Fang, R. Ghani, A. Greenwald, T. Lyons, D. C. Parkes, J. Schultz, S. Saria, S. F. Smith, M. Tambe, Artificial intelligence for social good (2019). arXiv:arXiv:1901.05406.

[389] B. C. Stahl, D. Wright, Ethics and privacy in ai and big data: Implementing responsible research and innovation, IEEE Security & Privacy 16 (3) (2018) 26–33.

[390] High Level Expert Group on Artificial Intelligence, Ethics guidelines for trustworthy ai, Tech. rep., European Commission (2019).

[391] B. d'Alessandro, C. O'Neil, T. LaGatta, Conscientious classification: A data scientist's guide to discrimination-aware classification, Big data 5 (2) (2017) 120–134.

[392] S. Barocas, A. D. Selbst, Big data's disparate impact, Calif. L. Rev. 104 (2016) 671.

[393] M. Hardt, E. Price, N. Srebro, et al., Equality of opportunity in supervised learning, in: Advances in neural information processing systems, 2016, pp. 3315–3323.

[394] T. Speicher, H. Heidari, N. Grgic-Hlaca, K. P. Gummadi, A. Singla, A. Weller, M. B. Zafar, A unified approach to quantifying algorithmic unfairness: Measuring individual group unfairness via inequality indices, in: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery Data Mining, ACM, 2018, pp. 2239–2248.

[395] F. Kamiran, T. Calders, Data preprocessing techniques for classification without discrimination, Knowledge and Information Systems 33 (1) (2012) 1–33.

[396] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, C. Dwork, Learning fair representations, in: International Conference on Machine Learning, 2013, pp. 325–333.

[397] B. H. Zhang, B. Lemoine, M. Mitchell, Mitigating unwanted biases with adversarial learning, in: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, ACM, 2018, pp. 335–340.

[398] Y. Ahn, Y.-R. Lin, Fairsight: Visual analytics for fairness in decision making, IEEE transactions on visualization and computer graphics (2019).

[399] E. Soares, P. Angelov, Fair-by-design explainable models for prediction of recidivism, arXiv preprint arXiv:1910.02043 (2019).

[400] J. Dressel, H. Farid, The accuracy, fairness, and limits of predicting recidivism, Science advances 4 (1) (2018) eaao5580.

[401] U. Aivodji, H. Arai, O. Fortineau, S. Gambs, S. Hara, A. Tapp, Fairwashing: the risk of rationalization, in: International Conference on Machine Learning, 2019, pp. 161–170.

[402] S. Sharma, J. Henderson, J. Ghosh, Certifai: Counterfactual explanations for robustness, transparency, interpretability, and fairness of artificial intelligence models, arXiv preprint arXiv:1905.07857 (2019).

[403] M. Drosou, H. Jagadish, E. Pitoura, J. Stoyanovich, Diversity in big data: A review, Big data 5 (2) (2017) 73–84.

[404] J. Lerman, Big data and its exclusions, Stan. L. Rev. Online 66 (2013) 55.

[405] R. Agrawal, S. Gollapudi, A. Halverson, S. Ieong, Diversifying search results, in: Proceedings of the second ACM international conference on web search and data mining, ACM, 2009, pp. 5–14.

[406] B. Smyth, P. McClave, Similarity vs. diversity, in: International conference on case-based reasoning, Springer, 2001, pp. 347–361.

[407] P. Wang, L. T. Yang, J. Li, J. Chen, S. Hu, Data fusion in cyber-physical-social systems: State-of-the-art and perspectives, Information Fusion 51 (2019) 42–57.

[408] W. Ding, X. Jing, Z. Yan, L. T. Yang, A survey on data fusion in internet of things: Towards secure and privacy-preserving fusion, Information Fusion 51 (2019) 129–144.

[409] A. Smirnov, T. Levashova, Knowledge fusion patterns: A survey, Information Fusion 52 (2019) 31–40.

[410] W. Ding, X. Jing, Z. Yan, L. T. Yang, A survey on data fusion in internet of things: Towards secure and privacy-preserving fusion, Information Fusion 51 (2019) 129–144.

[411] P. Wang, L. T. Yang, J. Li, J. Chen, S. Hu, Data fusion in cyber-physical-social systems: State-of-the-art and perspectives, Information Fusion 51 (2019) 42–57.

[412] B. P. L. Lau, S. H. Marakkalage, Y. Zhou, N. U. Hassan, C. Yuen, M. Zhang, U.-X. Tan, A survey of data fusion in smart city applications, Information Fusion 52 (2019) 357–374.

[413] S. Ramírez-Gallego, A. Fernández, S. García, M. Chen, F. Herrera, Big data: Tutorial and guidelines on information and process fusion for analytics algorithms with mapreduce, Information Fusion 42 (2018) 51–61.

[414] J. Konečný, H. B. McMahan, D. Ramage, P. Richtárik, Federated optimization: Distributed machine learning for on-device intelligence (2016). arXiv:1610.02527.

[415] B. McMahan, E. Moore, D. Ramage, S. Hampson, B. A. y Arcas, Communication-efficient learning of deep networks from decentralized data, in: Artificial Intelligence and Statistics, 2017, pp. 1273–1282.

[416] J. Konečnỳ, H. B. McMahan, F. X. Yu, P. Richtárik, A. T. Suresh, D. Bacon, Federated learning: Strategies for improving communication efficiency (2016). arXiv:1610.05492.

[417] S. Sun, A survey of multi-view machine learning, Neural computing and applications 23 (7-8) (2013) 2031–2038.

[418] R. Zhang, F. Nie, X. Li, X. Wei, Feature selection with multi-view data: A survey, Information Fusion 50 (2019) 158–167.

[419] J. Zhao, X. Xie, X. Xu, S. Sun, Multi-view learning overview: Recent progress and new challenges, Information Fusion 38 (2017) 43–54.

[420] S. J. Oh, R. Benenson, M. Fritz, B. Schiele, Faceless person recognition: Privacy implications in social media, in: Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, Proceedings, Part III, 2016, pp. 19–35.

[421] P. Aditya, R. Sen, P. Druschel, S. Joon Oh, R. Benenson, M. Fritz, B. Schiele, B. Bhattacharjee, T. T. Wu, I-pic: A platform for privacy-compliant image capture, in: Proceedings of the 14th annual international conference on mobile systems, applications, and services, ACM, 2016, pp. 235–248.

[422] Q. Sun, A. Tewari, W. Xu, M. Fritz, C. Theobalt, B. Schiele, A hybrid model for identity obfuscation by face replacement, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 553–569.

[423] X. L. Dong, D. Srivastava, Big data integration, in: 2013 IEEE 29th international conference on data engineering (ICDE), IEEE, 2013, pp. 1245–1248.

[424] D. Zhang, J. Zhao, F. Zhang, T. He, comobile: Real-time human mobility modeling at urban scale using multi-view learning, in: Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems, ACM, 2015, p. 40.

[425] S. J. Pan, Q. Yang, A survey on transfer learning, IEEE Transactions on knowledge and data engineering 22 (10) (2009) 1345–1359.

[426] M. Mitchell, S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, I. D. Raji, T. Gebru, Model cards for model reporting, in: Proceedings of the Conference on Fairness, Accountability, and Transparency, ACM, 2019, pp. 220–229.