IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received August 5, 2018, accepted September 4, 2018, date of publication September 17, 2018, date of current version October 12, 2018.

*Digital Object Identifier 10.1109/ACCESS.2018.2870052*

# Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)

**AMINA ADADI AND MOHAMMED BERRADA**
Computer and Interdisciplinary Physics Laboratory, Sidi Mohammed Ben Abdellah University, Fez 30050, Morocco
Corresponding author: Amina Adadi (amina.adadi@gmail.com)

**ABSTRACT** At the dawn of the fourth industrial revolution, we are witnessing a fast and widespread adoption of artificial intelligence (AI) in our daily life, which contributes to accelerating the shift towards a more algorithmic society. However, even with such unprecedented advancements, a key impediment to the use of AI-based systems is that they often lack transparency. Indeed, the black-box nature of these systems allows powerful predictions, but it cannot be directly explained. This issue has triggered a new debate on explainable AI (XAI). A research field holds substantial promise for improving trust and transparency of AI-based systems. It is recognized as the sine qua non for AI to continue making steady progress without disruption. This survey provides an entry point for interested researchers and practitioners to learn key aspects of the young and rapidly growing body of research related to XAI. Through the lens of the literature, we review the existing approaches regarding the topic, discuss trends surrounding its sphere, and present major research trajectories.

**INDEX TERMS** Explainable artificial intelligence, interpretable machine learning, black-box models.

## I. INTRODUCTION

### A. CONTEXT

Nowadays, Artificial Intelligence (AI) is democratized in our everyday life. To put this phenomenon into numbers, International Data Corporation (IDC) forecasts that global investment on AI will grow from 12 billion U.S. dollars in 2017 to 52.2 billion U.S. dollars by 2021 [1]. Meanwhile, the statistics portal Statista, expects that revenues from the AI market worldwide will grow from 480 billion U.S. dollars in 2017 to 2.59 trillion U.S. dollars by 2021 [2]. Gartner identifies AI as an inescapable technology among the Gartner Top 10 Strategic Technology Trends for 2018. Along with immersive experiences, digital twins, event-thinking and continuous adaptive security, they are shaping the next generation of digital business models and ecosystems [3]. Consequently, the proliferation of AI is having a significant impact on society. Indeed, AI has already become ubiquitous and we have become accustomed about AI making decisions for us in our daily life, from product and movie recommendations on Netflix and Amazon to friend suggestions on Facebook and tailored advertisements on Google search result pages. However, in life-changing decisions such as disease diagnosis, it is important to know the reasons behind such a critical decision. Here, the crucial need for explaining AI outcomes becomes fully apparent.

Problematically, though they appear powerful in terms of results and predictions, AI algorithms suffer from opacity, that it is difficult to get insight into their internal mechanism of work, especially Machine Learning (ML) algorithms. Which further compound the problem, because entrusting important decisions to a system that cannot explain itself presents obvious dangers.

To address this issue, Explainable Artificial Intelligence (XAI) proposes to make a shift towards more transparent AI. It aims to create a suite of techniques that produce more explainable models whilst maintaining high performance levels.

### B. XAI'S LANDSCAPE DYNAMIC

XAI has been gaining increasing attention recently. The growing dynamic around this field has been reflected in several scientific events. Examples of annual international conference series dedicated exclusively to the topic include Fairness, Accountability, and Transparency (FAT-ML) workshop at KDD 2014-2018 [4] and ICML Workshop on Human Interpretability in Machine Learning (WHI) 2016-2018 [5]. The topic has also become the key concern in panel discussions at specific sessions in major conferences such as NIPS 2016 Workshop on Interpretable ML for Complex Systems [6], IJCAI 2017 and IJCAI/ECAI 2018 Workshops on

Explainable Artificial Intelligence [7], XCI 2017 on Explainable Computational Intelligence [8] and IJCNN 2017 on Explainability of Learning Machines [9]. This year (2018) is flourishing by a wide range of dedicated workshops to the topic: CD-MAKE 2018 Workshop on Explainable Artificial Intelligence [10], ICAPS 2018 Workshop on EXplainable AI Planning [11], HRI 2018 Workshop on Explainable Robotic Systems [12], ACM Intelligent User Interfaces (IUI) 2018 workshop on Explainable Smart Systems (EXSS 2018) [13], IPMU 2018 on Advances on Explainable Artificial Intelligence [14], and finally ICCBR 2018 organize XCBR: the First Workshop On Case-Based Reasoning For The Explanation Of Intelligent Systems [15].

A high-level analysis of XAI's landscape leads to identify the key players and influencers behind this intense dynamic. Indeed, two of the most prominent actors pursuing XAI research are: (i) a group of academics operating under the acronym FAT* [4] and (ii) civilian and military researchers funded by the Defense Advanced Research Projects Agency (DARPA) [16].

FAT* academics (meaning fairness, accountability, and transparency in multiple artificial intelligence, machine learning, computer science, legal, social science, and policy applications) are primarily focused on promoting and enabling explainability and fairness in algorithmic decision-making systems with social and commercial impact. With over than 500 participants and more than 70 papers, FAT* conference, which held its fifth annual event in February 2018, brings together annually researchers and practitioners interested in fairness, accountability, and transparency in socio-technical systems.

The other group, DARPA, launched its XAI program in 2017 with the aim of developing new techniques capable of making intelligent systems explainable, the program includes 11 projects and will continue running until 2021. DARPA funded researchers seem primarily interested in increasing explainability in sophisticated pattern recognition models needed for security applications. Even though DARPA is funded by the US Department of Defense, the program involves researchers drawn from various academic institutions and diverse corporate teams.

Increasing interest in XAI has also been observed in the industrial community. Companies on the cutting edge of contributing to make AI more explainable include H2O.ai with its driverless AI product [17], Microsoft with its next generation of Azure: Azure ML Workbench.[1] Kyndi with its XAI platform for government, financial services, and healthcare,[2] and FICO with its Credit Risk Models.[3] To push the state of XAI even further, FICO is running the Explainable Machine Learning Challenge (xML challenge) [18]. The goal of this challenge is to identify new approaches for creating machine learning based AI models with both high accuracy and explainability. On the other hand, Cognilytica has examined in its "AI Positioning Matrix" (CAPM) the market of AI products. It proposed a chart where XAI technologies are arguably identified as high-sophisticated implementations beyond the threshold of the actual technology [19].

[1]https://azure.microsoft.com/en-us/services/machine-learning-services/
[2]https://kyndi.com
[3]http://www.fico.com

### C. CONTRIBUTION AND ORGANIZATION

Motivated by the preceding concerns and observations, this article collects and shares findings from a comprehensive and in-depth survey on XAI. In fact, by considering XAI as a field in terms of its research, we propose to step back for a holistic view of the present state-of-the-art advancements in this research field, in order to chart a path toward promising and suitable directions for future research. Unlike studies, that focus on specific dimensions of explainability, this work advocates the multidisciplinary nature of the studied field and introduces the major aspects and domains of explainability from different perspectives. A key claim of this paper is that the issue of explaining AI-powered systems is scientifically interesting and increasingly important, hence the necessity of providing a firm basis from the lens of literature to ground further discussion. The aim is to help interested researchers to quickly and effectively grasp important facets of the topic by having a clear idea about key aspects and related body of research.

In this sense, we make three contributions:

- We propose a comprehensive background regarding the main concepts, motivations, and implications of enabling explainability in intelligent systems.
- Based on a literature analysis of 381 papers, we provide an organized overview of the existing XAI approaches.
- We identify and discuss future research opportunities and potential trends in this field.

Accordingly, the remainder of the survey is organized as follows. Section II presents a preliminary background. Section III surveys the latest developments in the XAI field and organizes surveyed approaches according to four perspectives. Section IV discusses research directions and open problems that we gathered and distilled from the literature survey. Finally, Section V concludes this survey.

## II. BACKGROUND

### A. UNDERSTANDING XAI: A CONTEXTUAL DEFINITION

XAI is a research field that aims to make AI systems results more understandable to humans.

The term was first coined in 2004 by Van Lent *et al.* [20], to describe the ability of their system to explain the behavior of AI-controlled entities in simulation games application.

While the term is relatively new, the problem of explainability has existed since the mid-1970s when researchers studied explanation for expert systems [21]. However, the pace of progress towards resolving such problem has slowed down as AI reached an inflection point with the spectacular advances in ML. Since then the focus of AI research has shifted towards implementing models and algorithms

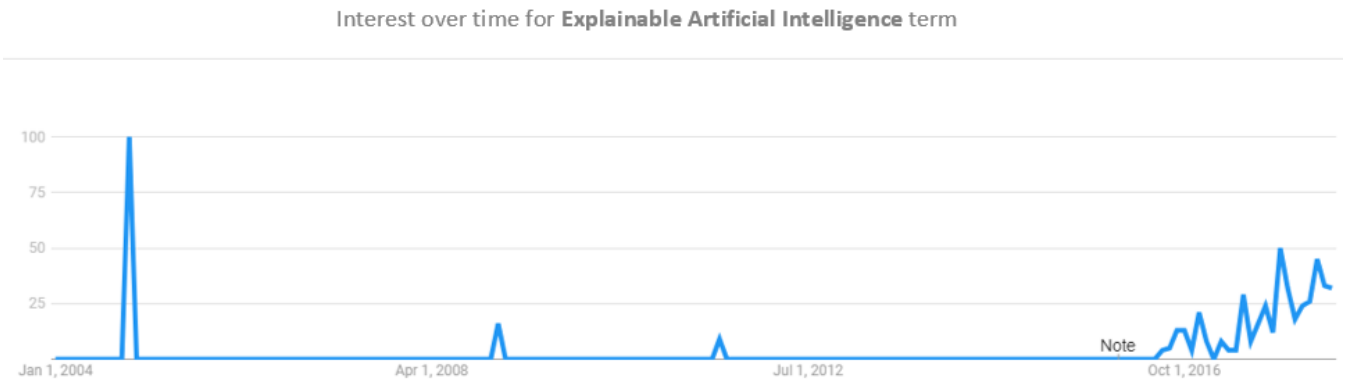


**FIGURE 1.** Google trends result for research interest of ≪ Explainable Artificial Intelligence ≫ term.

that emphases predictive power while the ability to explain decision processes has taken a back seat.

Recently, XAI topic has received renewed attention from academia and practitioners. Figure 1 illustrates the remarkable resurgence of XAI term research interest using google trends. The re-emergence of this research topic is the direct result of the unstoppable penetration of AI/ML across industries and its crucial impact in critical decision-making processes, without being able to provide detailed information about the chain of reasoning that leads to certain decisions, recommendations, predictions or actions made by it. Therefore, the social, ethical and legal pressure calls for new AI techniques that are capable of making decisions explainable and understandable.

Technically, there is no standard and generally accepted definition of explainable AI. Actually, XAI term tends to refer to the movement, initiatives, and efforts made in response to AI transparency and trust concerns, more than to a formal technical concept. Thus, to put some clarification around this trend, we quote some XAI definitions as seen by those who are calling for it. According to DARPA [16], XAI aims to ''produce more explainable models, while maintaining a high level of learning performance (prediction accuracy); and enable human users to understand, appropriately, trust, and effectively manage the emerging generation of artificially intelligent partners''.

The goal of enabling explainability in ML, as stated by FAT* [4], ''is to ensure that algorithmic decisions as well as any data driving those decisions can be explained to end-users and other stakeholders in non-technical terms''.

FICO [19], the organizer of xML Challenge, see XAI as ''an innovation towards opening up the black-box of ML'' and as ''a challenge to create models and techniques that both accurate and provide good trustworthy explanation that will satisfies customers' needs''.

For an interested researcher, we believe that it is very important, to gain a deep understanding of the XAI concept, beyond colloquial definitions, primary goals and shallow facts. For this, we propose to explore the big picture of the key concepts shaping the XAI landscape.

While performing our scan of literature, which will be detailed in the next section, we conducted a linguistic search to identify and record relevant terms across research communities that strongly relate to the concept of XAI. The goal of this analysis is to gain insights into how research communities approach explainability and to detect the main concepts that contribute to define this notion. As a result, the word cloud shown in Figure 2 provides an intuitive grasp of the XAI's scope and allows drawing the big picture of this research field by highlighting the important related concepts. Important terms are ordered according to the frequency of its appearance as keywords in the surveyed papers, and this after filtering technical terms like deep learning, decision tree, sensitive analysis etc.


**FIGURE 2.** XAI word cloud.

In Table 1, we cast insights on a sample of relevant and common related XAI concepts that we believe help to define contextually the studied field.

As detailed in Table 1, XAI is not a monolithic concept, it reflects several distinct related notions. Explainability is closely related to the concept of interpretability: interpretable

**TABLE 1.** Key related concepts of XAI.

| Term | Description |
|---|---|
| Interpretable Machine Learning | An interpretable system is a system where a user cannot only see but also study and understand how inputs are mathematically mapped to outputs [22]. This term is favored over "explainable" in the ML context where it refers to the capability of understanding the work logic in ML algorithms.<br>Researchers often use the two terms "interpretability" and "explainability" synonymously [23], [24]. Even though there is an acknowledgment for the need of a clear taxonomy [25]. Other authors use other terms such as understandability [26] or comprehensibility [27] to refer to the same issue, while some industrials [28] prefer the term intelligible AI. |
| Black-box problem | In science, computing, and engineering, the terms black box, gray box and white box are used with reference to different levels of closure of the component internal essence [29]. In particular, a black box component does not disclose anything about its internal design, structure and implementation, whereas its opposite side, a white box component is completely exposed to its user. In between, there may exist different levels of grey box components depending upon how much details are available. Commercially, The "black box" concept has been exploited by technological enterprises, usually in their efforts to protect intellectual property and maintain competitiveness.<br>In AI, the difficulty for the system to provide a suitable explanation for how it arrived at an answer is referred to as "the black-box problem." |
| Responsible Artificial Intelligence | Responsible AI is an AI that takes into account societal values, moral and ethical considerations.<br>Responsible AI has three main pillars: Accountability, Responsibility, and Transparency. Together, these considerations form the A.R.T. principles for AI [30]:<br>**Accountability** refers to the need to explain and justify one's decisions and actions to its partners, users and others with whom the system interacts.<br>**Responsibility** refers to the role of people themselves and to the capability of AI systems to answer for one's decision and identify errors or unexpected results.<br>**Transparency** refers to the need to describe, inspect and reproduce the mechanisms through which AI systems make decisions and learns to adapt to its environment, and to the governance of the data used created.<br>Other initiatives focus on some additional considerations such as fairness and ethics in defining Responsible AI [4]. |
| Accurate Artificial Intelligence | AI's accuracy is a performance metric that refers to the number of correct predictions made by the model (typically a ML model) over all kinds of predictions made [31]. |
| Data science | AI models usually require getting a training and testing set of Data. Data science is a field that unifies statistics, data analysis, machine learning and their related methods in order to understand and analyze actual phenomena with data [32]. |
| Social science | Explanation is, first and foremost, a form of social interaction. The general discipline of social science is concerned with society and the relationships among individuals within a society [33]. Some interesting social sciences theories include causality, systematic cognitive biases, contrastive explanation, and argumentation. |
| Third-wave AI | Driven by a contextual adaptation, new researches are shaping the so-called third wave AI (also called Artificial Intelligence 3.0) where AI systems construct explanatory models for classes of real-world phenomena (XAI), learn and reason as they encounter new tasks and situations (Continuous learning) and can establish natural communication with human (Interactive AI, Human-machine symbiosis, Brain-Computer Interface) [34]. |
| Artificial General Intelligence (AGI) | AGI was a primary goal of the initial AI field, it is the intelligence of a machine that could successfully perform any intellectual task that a human being can, AGI is also referred to as "strong AI", "full AI" or as the ability of a machine to perform "general intelligent action"[35].<br>AI research that study machines that can perform actions superior to human intellect are known as: Artificial Superintelligence (ASI)[36]. |

systems are explainable if their operations can be understood by human. We note that even though ''Explainable'' is a keyword in the XAI appellation, in ML community the term ''interpretable'' is more used than ''Explainable''. Figure 3 confirms this observation, it shows trends regarding the use of the two terms in both scientific and public settings.

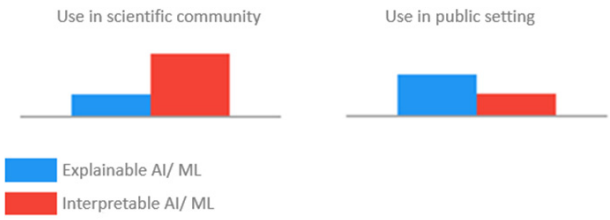


**FIGURE 3.** Google trends result for comparing the use of "Explainable" and "Interpretable" according to the context.

Furthermore, it should be noted that none of the aforementioned variation terms (understandable, comprehensible, intelligible. . .) is enough specific to enable formalization. They implicitly depend on the user's expertise, preferences and other contextual variables.

XAI is centered on the challenge of demystifying the black boxes, it also implies Responsible AI as it can helps to produce transparent models. This should happen without affecting the AI models accuracy, thus in AI in general and in ML specifically, often a tradeoff must be made between accuracy and interpretability. An obvious link with data science field arises as accuracy is closely tied to the quality and the quantity of the training data.

Rarely in literature, we come across the term "social science" or it derivative, yet explanation is a form of social interaction and clearly, it has psychological, cognitive and philosophical projections. Based on the conducted analysis, ideas from social science and human behavior are not sufficiently visible in this field.

Finally, XAI is a part of a new generation of AI technologies called the third wave AI, one of the objectives of this ambition "wave" is to precisely generate algorithms than can explain themselves. Ultimately, all this culminates in the quest for reaching human intelligence level, a goal known as AGI.

Based on this terms analysis, we built a unified and structured view of the main concepts related to XAI field (illustrated in Figure 4). We believe that aiming holism in approaching XAI concept, helps researchers to quickly be initiated about the topic and its context. Moreover, knowing the main keywords used in the field and the variation of

terms that are relatively referring to the same concepts, represent a helpful prerequisite to conduct a relevant and fruitful research.

### B. USING XAI: THE NEED AND THE APPLICATION OPPORTUNITIES

#### 1) THE NEED FOR XAI

For commercial benefits, for ethics concerns or for regulatory considerations, XAI is essential if users are to understand, appropriately trust, and effectively manage AI results. Based on the explored literature, the need for explaining AI systems may stem from (at least) four reasons, although it may appear that there is an overlap between these four reasons, from our standpoint, they capture the different motivations for explainability.

##### *a: EXPLAIN TO JUSTIFY*

The past several years have seen multiple controversies over AI/ML enabled systems yielding biased or discriminatory results [37], [38]. That implies an increasing need for explanations to ensure that AI based decisions were not made erroneously. When we talk about an explanation for a decision, we generally mean the need for reasons or justifications for that particular outcome, rather than a description of the inner workings or the logic of reasoning behind the decision-making process in general.

Using XAI systems provides the required information to justify results, particularly when unexpected decisions are made. It also ensures that there is an auditable and provable way to defend algorithmic decisions as being fair and ethical, which leads to building trust.

Furthermore, henceforth AI needs to provide justifications in order to be in compliance with legislation, for instance the "right to explanation", which is a regulation included in the

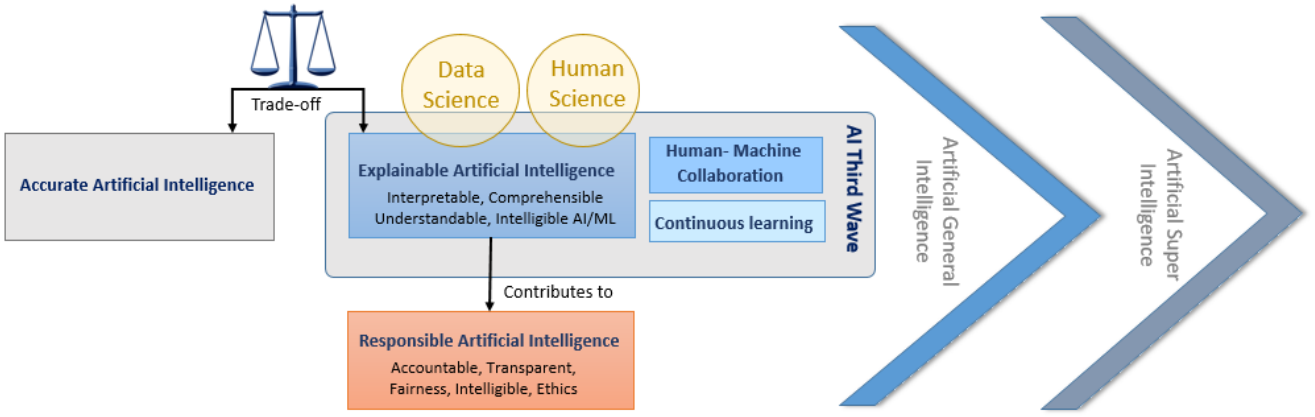


**FIGURE 4.** A schematic view of XAI related concepts.

General Data Protection Regulation (GDPR) that comes into effect across the EU on May 25, 2018 [39].

#### b: EXPLAIN TO CONTROL

Explainability is not just important for justifying decisions. It can also help prevent things from going wrong. Indeed, understanding more about system behavior provides greater visibility over unknown vulnerabilities and flaws, and helps to rapidly identify and correct errors in low criticality situations (debugging). Thus enabling an enhanced control.

#### c: EXPLAIN TO IMPROVE

Another reason for building explainable models is the need to continuously improve them. A model that can be explained and understood is one that can be more easily improved. Because users know why the system produced specific outputs, they will also know how to make it smarter. Thus, XAI could be the foundation for ongoing iteration and improvement between human and machine.

#### d: EXPLAIN TO DISCOVER

Asking for explanations is a helpful tool to learn new facts, to gather information and thus to gain knowledge. Only explainable systems can be useful for that. For example, given that AlphaGo Zero [40] can excel at the game of Go much better than human players, it would be desirable that the machine can explain its learned strategy (knowledge) to us. So It will come as no surprise if, in future, XAI models taught us about new and hidden laws in biology, chemistry and physics.

We conclude that explainability is a powerful tool for justifying AI based decisions. It can help to verify predictions, for improving models, and for gaining new insights into the problem at hand. Which leads towards more trustworthy AI systems (Figure 5).

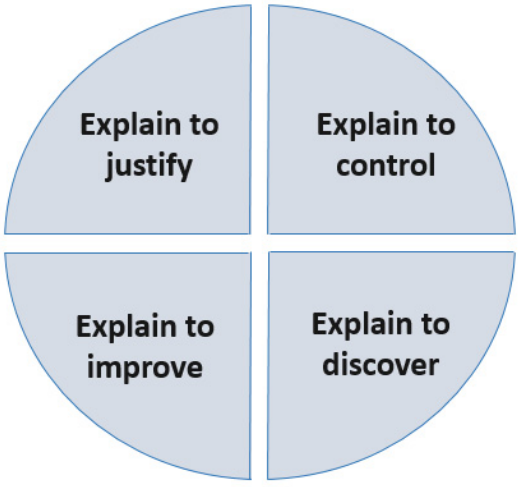


**FIGURE 5.** Reasons for XAI.

Even though academics and practitioners approve on the importance of XAI, not everyone agrees that there is a pressing need for greater interpretability in AI systems. At this regard, the value of XAI was called into question recently by Google research director Norvig [41], who noted that humans are not very good at explaining their decisions either, and claimed that the credibility of an AI system results could be gauged simply by observing its outputs over time. The AI powerhouse company researcher has stressed, indeed, an important point. Certainly, explainability is an essential property; however, it is not always a necessity. In fact, requiring every AI system to explain every decision could result in less efficient systems, forced design choices, and a bias towards explainable, but less capable and versatile outcomes. Furthermore, making AI systems explainable is undoubtedly expensive; they require considerable resources both in the development of the AI system and in the way it is interrogated in practice. Thus, it is important to think about why and when explanations are useful. The need for explainability depends on: (a) The degree of functional opacity caused by the complexity of AI algorithms: if it is low, no high level of interpretability is required. (b) The degree of resistance of the application domain to errors. If it has high resistance, unexpected errors are acceptable. For an AI system for targeted advertising, for example, a relatively low level of interpretability could suffice, as the consequences of it going wrong are negligible. On the other hand, the interpretability for an AI-based diagnosis system would be significantly higher. Any errors could not only harm the patient but also deter adoption of such systems. Therefore, any domain where the cost of making a wrong prediction is very high present a potential application domain of XAI approaches.

### 2) XAI APPLICATION DOMAINS

Interestingly, XAI can bring significant benefit to a large range of domains relying on AI systems. Herein we explore some potential domains where there is a need for a research work on explainable models.

#### a: TRANSPORTATION

Automated vehicles hold the promise for decreasing traffic deaths and providing enhanced mobility but also pose challenges in addressing the explainability of AI decisions. Autonomous vehicles have to make split-second decisions based on how they classify the objects in the scene in front of them. If a self-driving car suddenly acts abnormally because of some misclassification problem. The consequence can be dangerous. This is not a possibility this is already happening, recently, a self-driving Uber killed a woman in Arizona. It was the first known fatality involving a fully autonomous vehicle. The information reported anonymous sources who claimed the car's software registered an object in front of the vehicle, but treated it in the same way it would a plastic bag or tumbleweed carried on the wind [42]. Only an explainable system can clarify the ambiguous circumstances of such situation and eventually prevent it from happening.

Transportation is a potential application domain of the XAI. Works towards explaining self- driving vehicle behavior

has been already started [43], [44], but there is a long way to go.

#### b: HEALTHCARE

Medical diagnosis model is responsible for human life. How can we be confident enough to treat a patient as instructed by a black-box model?

In the mid-1990s, an artificial neural network (ANN) was trained to predict which pneumonia patients should be admitted to hospitals and which treated as outpatients. Initial findings indicated neural nets were far more accurate than classical statistical methods. However, after an extensive test, it turned out that the neural net had inferred that pneumonia patients with asthma have a lower risk of dying, and should not be admitted. Medically, this is counterintuitive however, it reflected a real pattern in the training data—asthma patients with pneumonia usually were admitted not only to the hospital but directly to the ICU, treated aggressively, and survived [37]. It was then decided to abandon the AI system because it was too dangerous to use it clinically. Only by interpreting the model, we can discover such a crucial problem and avoid it. Recently, researchers have conducted preliminary work aiming to make clinical AI-based systems explainable [37], [45]–[47]. The increasing number of these works confirms the challenge and the interest of applying XAI approaches on the healthcare domain.

#### c: LEGAL

In criminal justice, AI has the potential to better assess risks for recidivism and reduce costs associated with both crime and incarceration. However, when using a criminal decision model to predict the risk of recidivism at the court, we have to make sure the model behaves in an equitable, honest and nondiscriminatory manner. In Loomis v. Wisconsin [48], the case challenged the use of proprietary, closed source risk assessment software in sentencing Mr. Loomis to prison. The case alleged that the software ''Correctional Offender Management Profiling for Alternative Sanctions: COMPAS'' [49] violates the process rights by taking gender and race into account. The algorithms used were considered trade secrets and the causal audit process was not clearly known to the Judge.

Transparency of how a decision is made is a necessity in this critical domain, yet very few works are made towards making automated decision making in legal system explainable [49]–[51].

#### d: FINANCE

In financial services, benefits of using AI tools include improvements related to wealth-management activities, access to investment advice, and customer service. However, these tools also pose questions around data security and fair lending. Indeed, the financial industry is highly regulated and loan issuers are required by law to make fair decisions. Thus, one significant challenge of using AI-based systems, in credit scores and models, is that it's harder to provide the needed ''reason code'' to borrowers — the explanation of why they were denied credit. Especially when the basis for denial is the output from an opaque ML algorithm. Some credit bureaus agency such as Equifax and Experian are working on promising research projects to generate automated reason codes and make AI credit-based score decisions more explainable and auditor friendly [52].

#### e: MILITARY

Originally, the current famous XAI's initiative is made by military researchers [16], and the growing visibility of XAI topic is due largely to the call for research and the solicitation of DAPRA Projects.

Unsurprisingly, AI in the military arena also suffers from the AI explainability problem. In a report from MIT Technology review, Knight [53] delves into the challenges of relying on autonomous systems for military operations. As in the healthcare domain, this often involves life and death decisions, which again leads to similar types of ethical and legal dilemmas. The academic AI research community is well represented in this application domain with the DAPRA Ambitious XAI program, along with some research initiatives that studies explainability in this domain [54].

The line of work made in each of the discussed domains confirms the need of XAI. However, such works are only in their infancy, hearty research effort is yet to be done.

Moreover, XAI can find an interesting application in others domains like cybersecurity, education, entertainment, government, image recognition etc. An interesting chart of potential harms from automated decision-making was presented by Future of Privacy Forum [55], it depicts the various spheres of life where automated decision-making can cause injury and where providing automated explanation can turn them to a trustful processes, this includes employment, insurance and social benefit, housing and differential pricing of goods and services.

### C. ENABLING XAI: THE TECHNICAL CHALLENGE

Clearly, the awareness and demand for explainability are growing in various domains, hence the question as ''why the use of XAI is not systematic?'' or more simply ''why is not everyone using XAI?''.

In fact, bringing interpretability to AI systems is a very challenging technical issue. Explainability of intelligent systems has run the gamut from traditional expert systems, which are totally explainable but inflexible and hard to use, to Deep Neural Networks (DNN), which are effective but virtually impossible to see inside.

If we look back at the expert systems of the 80's, we had what we would consider a scrutable system: an inference engine leveraged a knowledge base to make assertions that it could explain using the chain of reasoning that led to the assertion [56]. Explanation capabilities are frequently the most significant benefit provided by an expert system, but these systems were completely built on subject matter

expertise and while powerful, were somewhat inflexible. Moreover, though significant progress was made on explainability during this period, with solid principles established, however the ''explainability'' problem was not considered to have been completely solved [57].

Modern machine learning algorithms go to the opposite end of the spectrum, yielding systems capable of working purely from observations and creating their own representations of the world on which to base their predictions. Nevertheless, the complexity that bestows the extraordinary predictive abilities on ML algorithms also makes the results the algorithms produce hard to understand. Indeed, ML algorithms are difficult to interpret, because of their structure and the way they are working. ML algorithms intrinsically consider high-degree interactions between input features, which make disaggregating such functions into human understandable form difficult. We take as an example the DNN, the most successful contemporary ML model. DNN has a generic multi-layer nonlinear structure consisting of many hidden layers and numerous number of neurons per layer, such architecture helps to produce high-level prediction through multiple levels of linear transformations and non-linear activations. While a single linear transformation may be interpreted by looking at the weights from the input features to each of the output classes, multiple layers with non-linear interactions at every layer imply disentangling a super complicated nested structure which is a difficult task and potentially even a questionable one [58].

Another perception of ML interpretability's technical challenge is explained by Hall and Gill [59] in their introduction to Machine Learning Interpretability book. They present the mathematical problems in interpretable ML so-called ''the multiplicity of good models'' [60]. As mentioned before, given the complicated structure of ML models, for the same set of input variables and prediction targets, complex machine learning algorithms can produce multiple accurate models by taking very similar but not the same internal pathway in the network, so details of explanations can also change across multiple accurate models. This systematic instability makes automated generated explanations difficult.

Arguing that AI/ML interpretability is a challenging issue, does not mean that all AI/ML techniques have the same level of opacity. Indeed, there are algorithms that are more interpretable than others are, and there is often a tradeoff between accuracy and interpretability: the most accurate AI/ML models usually are not very explainable (for example, deep neural nets, boosted trees, random forests, and support vector machines), and the most interpretable models usually are less accurate (for example, linear or logistic regression). Rather than being a static tradeoff, accuracy along with interpretability is a dynamic target that ongoing researches try to reach, as it will be discussed in Section III.B.

Furthermore, beyond the technical challenge, given the goal, the nature and the implications of XAI, progress towards overcoming this challenge, can only be achieved through interdisciplinary collaboration, where expertise and theories from different research fields are combined and methods and techniques are developed from multiple perspectives to move research forward. From our standpoint, enabling technologies and methods for XAI potentially belong to four basic research areas: **(i) Data science:** AI/ML algorithms are data hungry, they need more data to produce better predictions and decisions. The backward path that targets to produce better explanation and justification eventually also depends on that data. Data science is then a core element in the explainability process. **(ii) Artificial Intelligence/Machine Learning:** to generate explanation we need a computational process, we claim that using AI/ML as a computational process to explain AI/ML is an interesting work trail. **(iii) Human science:**to produce artificial explanations, it is worth first to models how humans explain decisions and behavior to each other [33]. Therefore, approaching theories from human science can lead to innovative explainable models. **(IV) Human Computer Interaction (HCI):** the user' understanding and trust of the system partly depends on the way he interacts with the machine. Given HCI's [61] core interest in technology that entails understanding and better empowering users, techniques from this research field can help in developing transparent systems.

Focusing on the What, the Why, the Where and the How, we tried to propose an extensive background regarding XAI by defining the concept of XAI, exposing the motivation behind its reemergence, identifying the segments of the market where the results are promising, and finally presenting some potential research areas that could potentially contribute to overcome the technical challenge related to XAI systems. The next section aims to capture researchers' attention on the growing research body of XAI through a literature survey.

## III. REVIEW

### A. RELATED SURVEYS

Despite the fact that the volume of research in interpretable and explainable AI is quickly expanding, a holistic survey and a systematic classification of these research works are missing. Indeed, according to the literature, there are few review papers in this field.

Two inescapable position papers are [25] and [62] that try to formalize the concept of explainability. The former attempted to provide a taxonomy of both the desiderata and methods in interpretability research. Lipton's work is not a survey in itself but it provides a solid discussion about what might constitute interpretability through the lens of the literature.

The survey of Doshi-Velez and Kim tried to define taxonomies and best practices for interpretability as a ''rigorous science''. The main contribution of this paper is a taxonomy of interpretability evaluation. In doing so, the authors shifted the focus on only one dimension of expandability: its measurement.

In their survey, Abdul *et al.* [61] analyzed a sizable literature of explainable research based on 289 core papers and

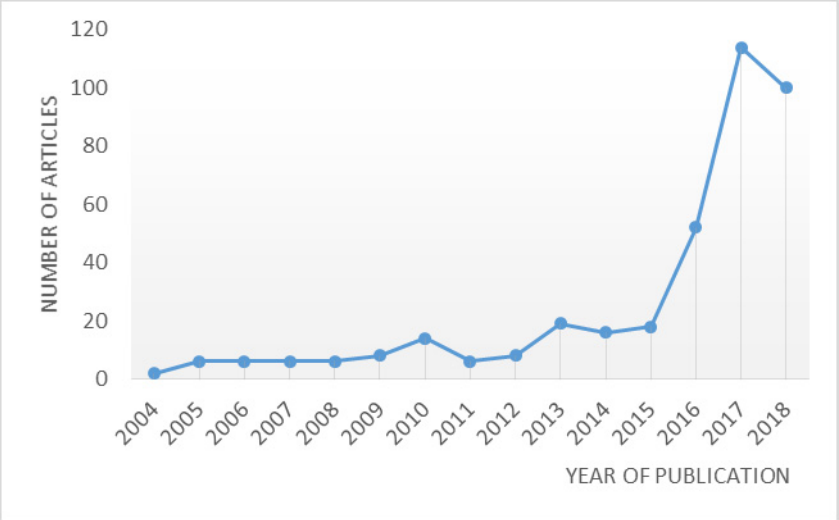


**FIGURE 6.** Surveyed articles by year (2004–2018).

12 412 citing papers and built a citation network. However, this work focus mainly on setting an HCI research agenda in explainability.

A recent survey by Guidotti *et al.* [63] reviewed methods for explaining black-box models at a large scale including data mining as well as machine learning. They presented a detailed taxonomy of explainability methods according to the type of problem faced. Even though the survey considered holism in terms of models (it discusses all black-box models), it emphasized only interpretability's mechanisms, ignoring by this other explainability dimensions such as evaluation. Hence, the detailed technical overview of surveyed methods makes it hard to get a quick understanding of the explanation methods space.

Finally, Dosilovic *et al.* [64] proposed a general overview of the topic in their conference paper. They presented the advances on explainability in machine learning models under the supervised learning paradigm, with a particular focus on DNN.

In contrast of existing surveys that focus on particular aspects of explainability, our survey provides a comprehensive and an organized overview of XAI research contributions from different perspectives. We target holism and clarity in exploring and exposing explainable approaches space.

### B. A HOLISTIC SURVEY

We conducted an extensive literature review by examining relevant papers from six major academic databases: SCOPUS, IEEExplore, ACM Digital Library, Google Scholar, Citeseer Library and ScienceDirect, in addition of preprints posted on arXiv. A keyword based search was used to select papers, it consists of searching for index keywords based on common variations in literature of the terms ''intelligible'', ''interpretable'', ''transparency'', ''black box'',''understandable'',''comprehensible'' and ''explainable'' and its union with a set of terms related to AI including ''Artificial Intelligence'', ''Intelligent system'' and ''Machine learning'', or terms referring to ML algorithms such as: ''deep learning'', ''classifier'', ''decision tree''. As we are mainly interested in recent advances in this field, the research was restricted to articles published between 2004 and 2018. The gathered papers were then scanned based on the titles, abstracts and keywords to determine relevant articles for further analysis. The list of selected papers was largely completed afterwards, by using a backward and forward snowballing strategy that consists of using the reference list of the selected papers and the citations to these papers to identify additional papers [65]. The final list of papers includes 381 papers. The publication timeline of these papers is shown in Figure 6, it illustrates the recent exponential increase of papers in this field.

Next, we will present a brief pointer to the relevant works by approaching them from different perspectives.

In a broad stroke, we describe the XAI space along four main axes, each one spanning its own spectrum and together they shape holistically XAI research landscape. Indeed, the aim is to present a comprehensive and holistic analysis of the state of art of this field by projecting works on four complementary axes. We do not intend to enumerate all the surveyed papers. However, we try to meet two criteria. The papers included in our discussion (a) are deemed to be a significant work (received high citation level) and (b) have good coverage of the corresponding axis.

## AXIS I. XAI METHODS TAXONOMY: EXPLAINABILITY STRATEGIES

In the quest to make AI system explainable, several explanation methods and strategies have been proposed in relatively short period, especially for ML algorithms. In this axis, we propose an overview of existing interpretability methods.

The majority of works discuss explainability in ML algorithms and thus, it is ''interpretability'' term that will usually be used.

Based on the conducted survey of the literature, we arrive to classify the methods according to three criteria: (i) the complexity of interpretability, (ii) the scoop of interpretability, and (iii) the level of dependency from the used ML model.

We note that, since explainability in AI is still an emerging field, the classes of methods belonging to the proposed taxonomy are neither mutually exclusive nor exhaustive. However, this can be a good yardstick to compare and contrast across multiple methods.

In the following subsections, we will describe the main features of each class and give examples from current research.

### A. COMPLEXITY RELATED METHODS

The complexity of a machine-learning model is directly related to its interpretability. Generally, the more complex the model, the more difficult it is to interpret and explain. Thus, the most straightforward way to get to interpretable AI/ML would be to design an algorithm that is inherently and intrinsically interpretable. Many papers support this classic approach, to name a few:

Letham *et al.* [66] proposed a model called Bayesian Rule Lists (BRL) based on decision tree, the authors claimed that preliminary interpretable models provide a concise and convincing capabilities to gain domain experts trust.

Caruana *et al.* [37] described an application of a learning method based on generalized additive models to the pneumonia problem. They proved the intelligibility of their model through case studies on real medical data.

Xu *et al.* [67] introduced an attention based model that automatically learns to describe the content of images. They showed through visualization how the model is able to interpret the results.

Ustun and Rudin [68] presented a sparse linear models for creating a data-driven scoring systems called SLIM, the results of this work highlight the interpretability capability of the proposed system to provide users with qualitative understanding due to their high level of sparsity and small integer coefficients.

A common challenge, which hinders the usability of this class of methods, is the tradeoff between interpretability and accuracy [69]. As noted by Breiman [70] ''accuracy generally requires more complex prediction methods . . .[and] simple and interpretable functions do not make the most accurate predictors''. In a sense, intrinsic interpretable models come at a cost of accuracy.

An alternative approach to interpretability in machine learning is to construct a high complex uninterpretable black-box model with high accuracy and subsequently use a separate set of techniques to perform, what we could define as a reverse engineering process to provide the needed explanations without altering or even knowing the inner works of the original model. This class of methods offers then a post-hoc explanation [25]. Though it could be significantly complex and costly, most recent works done in XAI field belong to post-hoc class, it includes natural language explanations [71], visualizations of learned models [72], and explanations by example [73]. More approaches will be presented in detail in the next subsections.

In light of this, we can conclude that the overall utility value of interpretability depends on the nature of the prediction task. As long as the model is accurate for the task, and uses a reasonably restricted number of internal components, intrinsic interpretable models are sufficient. If otherwise the prediction target involved complex and highly accurate models, considering post-hoc interpretation models is necessary.

It should also be noted that in literature there is a group of intrinsic methods for complex uninterpretable models. These methods aim to modify the internal structure of a complex black-box model that are not primarily interpretable - typically a DNN- to mitigated their opacity and thus improve their interpretability [74]. The used methods may either be components that add additional capabilities, components that belong to the model architecture [75], [76], e.g. as part of the loss function [77], or as part of the architecture structure, in terms of operations between layers [78], [79].

### B. SCOOP RELATED METHODS

Interpretability implies understanding an automated model, this supports two variations according to the scoop of interpretability: understanding the entire model behavior or understanding a single prediction. In the studied literature, contributions are made in both directions. Accordingly, we distinguish between two subclasses: (i) Global interpretability and (ii) Local interpretability.

#### 1) GLOBAL INTERPRETABILITY

Global interpretability facilitates the understanding of the whole logic of a model and follows the entire reasoning leading to all the different possible outcomes. This class of methods is helpful when ML models are crucial to inform population level decisions, such as drugs consumption trends or a climatic change [80]. In such cases, a global effect estimate would be more helpful than many explanations for all the possible idiosyncrasies.

Works that propose globally interpretable models include the aforementioned additive models for predicting pneumonia risk [37] and rule sets generated from sparse Bayesian generative model [66]. However, these models are usually specifically structured thus limited in predictability to preserve interpretability.

Yang *et al.* [80] proposed a global model interpretation via recursive partitioning called (GIRP) to build a global interpretation tree for a wide range of ML models based on their local explanations. In their experiments, the authors highlighted that their method can discover whether a particular ML model is behaving in a reasonable way or overfit to some unreasonable pattern.

Valenzuela-Escárcega *et al.* [81] proposed a supervised approach for information extraction, which provides a global,

deterministic interpretation. This work supports the idea that representation learning can be successfully combined with traditional, pattern-based bootstrapping yielding models that are interpretable.

Nguyen *et al.* [82] proposed an approach based on activation maximization—synthesizing the preferred inputs for neurons in neural networks—via a learned prior in the form of a deep generator network to produce a global interpretable model for image recognition. Activation maximization technique was previously used by Erhan *et al.* [83].

Even though a multitude of techniques is used in literature to enable global interpretability. Arguably, global model interpretability is hard to achieve in practice, especially for models that exceed a handful of parameters. Analogically to human, who focus effort on only part of the model in order to comprehend the whole of it, local interpretability can be more readily applicable.

#### 2) LOCAL INTERPRETABILITY

Explaining the reasons for a specific decision or single prediction means that interpretability is occurring locally. This scoop of interpretability is used to generate an individual explanation, generally, to justify why the model made a specific decision for an instance. Several explored papers propose local explanation methods. We provide next an overview of the explanation methods examined in reviewed papers.

Ribeiro *et al.* [84] proposed LIME for Local Interpretable Model-Agnostic Explanation. This model can approximate a black-box model locally in the neighborhood of any prediction of interest. Newer, related, and highly anticipated work from the creators of LIME, called anchors [85], extends LIME using decision rules. In the same vein, leave-one-covariate-out (LOCO) [86] is another popular technique for generating local explanation models that offer local variable importance measures.

Another attempt to produce local explanations is made by Baehrens *et al.* [87]. In this work, the authors presented a method capable of explaining the local decision taken by arbitrary nonlinear classification algorithms, using the local gradients that characterize how a data point has to be moved to change its predicted label. By following this line of work, we find a set of works using similar methods for image classification models [88]–[91]. Actually, it is a common approach to understanding the decisions of image classification systems by finding regions of an image that were particularly influential to the final classification. Also called sensitivity maps, saliency maps, or pixel attribution maps [92], these approaches use occlusion techniques or calculations with gradients to assign an ''importance'' value to individual pixels which are meant to reflect their influence on the final classification.

Based on the decomposition of a model's predictions on individual contributions of each feature, Robnik-Ŝikonja and Kononenko [93] proposed to explain the model prediction for one instance by measuring the difference between the original prediction and the one made with omitting a set of features. Recent works that use decomposition to explain locally include [94] and [95].

While there are several different techniques for obtaining local-explanations [96]–[100], recent work by Lundberg and Lee [101] have shown that there are equivalences among these techniques. They introduced a promising newer technique with solid theoretical support called Shapely Explanations that unifies local approaches.

An interesting and promising line of work is focusing on combining the strength and the benefits of both local and global interpretability. Examples include [102]–[104]. The four possible combinations are: (i) The standard global model interpretability answers how does the model make predictions. (ii) Global model interpretability on a modular level identifies how do parts of the model influence predictions. (iii) Local interpretability for a group of predictions indicates why did the model make specific decisions for a group of instances. (iv) And finally, the usual local interpretability for a single prediction used to justify why did the model make a specific decision for an instance [105].

Another observation that should be noted is that in the reviewed literature, local explanations is the most used methods to generate explanations in DNNs. However, even though these approaches are developed to explain neural networks, authors usually underline that their approaches can be potentially adopted to explain any kind of model, which means they are agnostic models. Another way to classify explanations method that will be explored next.

### C. MODEL RELATED METHODS

Another important way to classify model interpretability techniques is whether they are model agnostic, meaning they can be applied to any types of ML algorithms, or model specific, meaning techniques that are applicable only for a single type or class of algorithm.

#### 1) MODEL-SPECIFIC INTERPRETABILITY

Model-specific interpretability methods are limited to specific model classes. Intrinsic methods are by definition model-specific. The drawback of this practice is that when we require a particular type of interpretation, we are limited in terms of choice to models that provide it, potentially at the expense of using a more predictive and representative model. Therefore, there has been a recent surge in interest in model-agnostic interpretability methods as they are model-free.

#### 2) MODEL-AGNOSTIC INTERPRETABILITY

Model-agnostic methods are not tied to a particular type of ML model. In other words, this class of methods separates prediction from explanation. Model-agnostic interpretations are usually post-hoc, they are generally used to interpret ANN and could be local or global interpretable models. In the interest of improving interpretability AI models, a large amount of model-agnostic methods have been developed recently using range techniques from statistics, machine learning and data

science. Since the reviewed papers lie mostly in this class, we present herein an overview of the studied works grouped by techniques. These broadly fall into four technique types: (i) Visualization, (ii) Knowledge extraction, (iii) Influence methods and (iv) Example-based explanation.

#### a: VISUALIZATION

A natural idea to understand a ML model, especially DNN, is to visualize its representations to explore the pattern hidden inside a neural unit. Unsurprisingly, a consistent body of research investigates this way with the help of diverse visualization techniques in order to see inside these black boxes. Visualization techniques are essentially applied to supervised learning models. Amongst the reviewed literature, the popular visualization techniques are: (i) Surrogate models, (ii) Partial Dependence Plot (PDP) and (iii) Individual Conditional Expectation (ICE).

##### i) SURROGATE MODELS

A surrogate model is a simple model used to explain a complex model. More specifically, it is an interpretable model (like a linear model or decision tree) which is trained on the predictions of the original black-box model in order to interpret the latter. However, there are almost no theoretical guarantees that the simple surrogate model is highly representative of the more complex model. The aforementioned LIME [84] approach is a prescribed method for building local surrogate models around single observations. Bastani *et al.* [106] used a surrogate model approach where they extract a decision tree that represents model behavior. Another remarkable work by Thiagarajan *et al.* [107] proposed an approach for building TreeView visualizations using a surrogate model.

##### ii) PARTIAL DEPENDENCE PLOT (PDP)

PDP is a graphical representation that helps visualizing the average partial relationship between one or more input variables and the predictions of a black-box model. Works that use PDP to understand supervised learning model include: Green and Kern [108] who used PDPs to understand the relationship between predictors and the conditional average treatment effect for a voter mobilization experiment, with the predictions being made by Bayesian Additive Regression Trees (Chipman *et al.* [109]). In the ecological literature, Elith *et al.* [110], who rely on stochastic gradient boosting, used PDPs to understand how different environmental factors influence the distribution of a particular freshwater. Berk and Bleich [51] demonstrated the advantage of using Random Forests and the associated PDPs to accurately model predictor-response relationships under asymmetric classification costs that often arise in criminal justice settings. Recently Welling *et al.* [111] proposed a methodology called Forest Floor, to visualize and interpret random forest models, the proposed techniques rely on the feature contributions method rather than PDP. As argued by the authors the advantages of Forest Floor over PDP is that interactions are not masked by averaging. Thus, it is possible to locate interactions, which are not visualized in a given projection.

##### iii) INDIVIDUAL CONDITIONAL EXPECTATION (ICE)

ICE plots extend PDP, whereas PD plots provide a coarse view of a model's workings, ICE plots reveal interactions and individual differences by disaggregating the PDP output. Recent works use ICE rather than the classical PDP. For instance, Goldstein *et al.* [112] introduced ICE techniques and proved it advantage over PDP. Later Casalicchio *et al.* [113] proposed a local feature importance based approach that uses both partial importance (PI) and individual conditional importance (ICI) plots as visual tools.

#### b: KNOWLEDGE EXTRACTION

It is difficult to explain how ML models work, especially when the models are based on ANN. indeed, as cited before, multilayer feedforward networks are universal approximators. However, since learning algorithms modify cells in the hidden layer, this may constitute interesting internal representations. The task of extracting explanations from the network is therefore to extract, in a comprehensible form, the knowledge acquired by an ANN during training and encoded as an internal representation.

In the explored literature, several works propose methods to extract the knowledge embedded in the ANN that mainly rely on two techniques: (i) Rule Extraction and (ii) Model Distillation.

##### i) RULE EXTRACTION

One effort to gain insight into highly complex models is the use of rule extraction [114]–[116]. Works supporting this technique propose approaches that provide a symbolic and comprehensible description of the knowledge learned by the network during its training by extracting rules that approximate the decision-making process in ANN by utilizing the input and output of the ANN. Which is, by the way, the kind of knowledge used in traditional artificial intelligence expert systems. The survey by Ras [74] had taken on the classification of rule extraction strategies proposed earlier in [117] and [118] and proposed three modes to extract rules: (a) pedagogical rule extraction, (b) decompositional rule extraction and (c) eclectic Rule-Extraction.

Decompositional approaches focus on extracting rules at the level of individual units within the trained ANN, i.e. the view of the underlying ANN is one of transparency (e.g. [93]–[95]). While pedagogical approaches treat the trained ANN as a black-box i.e. the view of the underlying ANN is opaque, the Orthogonal Search-based Rule Extraction algorithm (OSRE) from [119] is a successful pedagogical methodology often applied in biomedicine. The third type (eclectic) is a hybrid approach for rule extraction that incorporates elements of both the decompositionnel and pedagogical rule-extraction techniques [120].

#### ii) MODEL DISTILLATION

Another technique that falls in the knowledge extraction category is model distillation. Distillation is a model compression to transfer information (dark knowledge) from deep networks (the ''teacher'') to shallow networks (the ''student'') [121], [122]. Model compression was originally proposed to reduce the computational cost of a model at runtime but has later been applied for interpretability.

Tan *et al.* [49] investigated how model distillation can be used to distill complex models into transparent models. Che *et al.* [123] introduced in their paper a knowledge-distillation approach called Interpretable Mimic Learning, to learn interpretable phenotype features for making robust prediction while mimicking the performance of deep learning models. A recent work by Xu *et al.* [124] presented DarkSight, a visualization method for interpreting the predictions of a black-box classifier on a data set in a way inspired by the notion of dark knowledge. The proposed method combines ideas from knowledge distillation, dimension reduction, and visualization of DNN. Interested researchers can also consult these [125]–[127] for further details about this technique.

### c: INFLUENCE METHODS

This type of techniques estimates the importance or the relevance of a feature by changing the input or internal components and recording how much the changes affect model performance. Influence techniques are often visualized. In the reviewed literature, there are three alternative methods to obtain input variable's relevance: (i) Sensitivity analysis, (ii) Layer-wise relevance propagation and (iii) Feature Importance.

#### i) SENSITIVITY ANALYSIS

Sensitivity refers to how an ANN output is influenced by its input and/or weight perturbations [128]. It is used to verify whether model behavior and outputs remain stable when data is intentionally perturbed or other changes are simulated in data. Visualizing the results of sensitivity analysis (SA) is considered an agnostic explanation technique, since displaying models stability as data change over time enhance trust in machine learning results. SA has been increasingly used in explaining ANN in general and DNN classification of images in particular [129] and [130]. However, it is important to note that SA does not produce an explanation of the function value itself, but rather a variation of it. The purpose of performing a SA is thus usually not to actually explain the relationship found. Instead, SA is generally used to test models for stability and trustworthiness, either as a tool to find and remove unimportant input attributes or as a starting point for some more powerful explanation technique (e.g. decomposition).

#### ii) LAYER-WISE RELEVANCE PROPAGATION (LRP)

Another technique to compute relevances was proposed in [131] as the Layer-wise Relevance Propagation algorithm. LRP redistributes prediction function backwards, starting from the output layer of the network and backpropagating up to the input layer. The key property of this redistribution process is referred to as relevance conservation. In contrast to SA, this method explains predictions relative to the state of maximum uncertainty, i.e. it identifies properties which are pivotal for the prediction ''rooster''.

#### iii) FEATURE IMPORTANCE

Variable importance quantifies the contribution of each input variable (feature) to the predictions of a complex ML model. The increase of the model's prediction error is calculated after permuting the feature in order to measure a feature's importance. Permuting the values of important features increases the model error. While permuting the values of unimportant features are ignored by the model and thus keeps model error unchanged. Based on this technique, Fisher *et al.* [132] proposed a model-agnostic version of the feature importance called Model Class Reliance (MCR). While the aforementioned work [113] proposed a local version of the feature importance called SFIMP for permutation-based shapley feature importance. LOCO [86] use a local feature importance as well.

### d: EXAMPLE-BASED EXPLANATION

Example-based explanation techniques select particular instances of the dataset to explain the behavior of machine learning models. Example-based explanations are mostly model-agnostic because they make any ML model more interpretable. The slight difference with model-agnostic methods is that the example-based explanation methods interpret a model by selecting instances of the dataset and not by acting on features or transforming the model.

Based on the conducted review we identified two promising example-based interpretability techniques: (i) Prototypes and criticisms and (ii) Counterfactuals explanations.

#### i) PROTOTYPES AND CRITICISMS

Prototypes are a selection of representative instances from the data [133]–[135], thus item membership is determined by its similarity to the prototypes which leads to overgeneralization. To avoid this, advantage exceptions have to be shown, also called criticisms: instances that are not well represented by those prototypes. Kim [136] developed an unsupervised algorithm for automatically finding prototypes and critics for a dataset, called MMD-critic. When applied to unlabeled data, it finds prototypes and critics that characterize the dataset as a whole.

#### ii) COUNTERFACTUALS EXPLANATIONS

Wachter *et al.* [137] presented the concept of ''unconditional counterfactual explanations'' as a novel type of explanation of automated decisions. Counterfactual explanations describe the minimum conditions that would have led to an alternative decision (e.g. a bank loan being approved), without the need to describe the full logic of the algorithm. The focus here is

on explaining a single prediction in contrast to adversarial examples where the emphasis is on reversing the prediction and not explaining it [138].

The proposed classification of techniques is based on our study of the actual literature. As the research contributions in this class of methods are actively growing, new model-agnostic techniques are regularly proposed.

Finally, it is worth to note that the main advantage of model-agnostic methods is "flexibility" at the model, the explanation and representation level. Nevertheless, although model-agnostic interpretability techniques are convenient, they often rely on surrogate models or other approximations that can degrade the accuracy of the explanations they provide. While model-specific interpretation techniques tend to use the model to be interpreted directly, leading to potentially more accurate explanations.

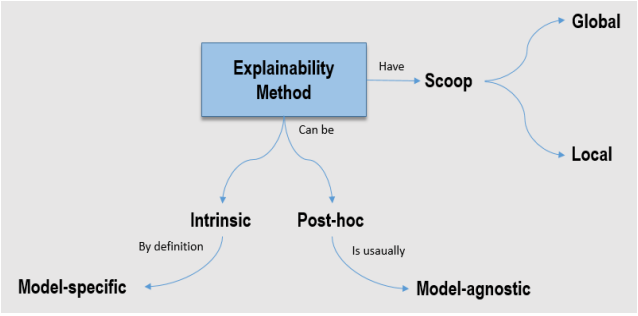


**FIGURE 7.** A pseudo ontology of XAI methods taxonomy.

To sum up (Figure 7), a key distinction is drawn in current thinking in terms of explaining the ML based AI system between true transparency (interpretable models such as decision tree, rules and linear models) and post-hoc interpretations, additional techniques used to lighten up the darkness of complex black-box models such as DNN, and that either by generating local explanations for particular inputs or by globally explaining the entire model. Local explanations focus on data and provide individual explanations, they provide trust to model outcomes. While global explanations focus on the model and provide an understanding of the decision process, it connotes some sense of understanding the mechanism by which the model works. Thus in term trustworthiness, local explanations are more faithful than global explanations.

Unarguably, the most popular class of explainability methods is model-agnostic class, this type of methods is usually used for ANN models. Because they are model independent, consequently, model-agnostic techniques are comparable, that is possible to compare the behavior of the same model with different types of agnostic model techniques.

Table 2 summarizes the various explainability techniques listed so far. Together with some good references and their projection on the detailed methods' taxonomy, form a useful reference for the reader to gain knowledge about recent XAI techniques.

## AXIS 2. XAI MEASUREMENT: EVALUATING EXPLANATIONS

"Are all models in all defined-to-be-interpretable model classes equally interpretable?" This is how Doshi-Velez and Kim [62] questioned interoperability measurement and evaluation issue. Indeed, despite the growing body of research that produces interpretable ML methods, there have been few works on evaluating these methods and quantifying their relevance (only 5% of the studied papers focus on this issue). This probably due to the subjective nature of explainability. However, given the number of existing interpretability methods, the need for comparing, validating, quantifying and thus evaluating these methods arises.

Doshi-Velez and Kim established a baseline of evaluation approaches and proposed three major types of interpretability evaluation: (i) application-grounded: put the explanation into the application and let the end user (typically a domain expert) test it. This type evaluates the quality of an explanation in the context of its end-task, (ii) human-grounded: is about conducting simplified application-grounded evaluation where experiments are run with lay humans rather than domain experts. This type is most appropriate when the goal is to test more general notions of the quality of an explanation, and (iii) functionally-grounded: this type does not involve humans, it is most appropriate once we have a class of models or regularizers that have already been validated, e.g. via human-grounded experiments.

Based on Doshi-Velez's evaluation classification, Mohseni and Ragan [148] presented a human-grounded evaluation benchmark for evaluating instance explanations of images and textual data. They demonstrated that by comparing the explanation results from classification models to the benchmark's annotation meta-data, it is possible to evaluate the quality and appropriateness of local explanations. Thus, they showed how human-grounded evaluation could be used as a measure to qualify local machine-learning explanations.

Earlier, Huysmans *et al.* [149] investigated decision trees, decision tables, propositional rules, and oblique rules in order to understand which is the most interpretable. To this end, they performed an end-user experiment to compare them. They found that overall decision trees and decision tables were the most Interpretable, but that different tasks made the tree or table more desirable.

Backhaus and Seiffert [150] suggested quantitative measures to compare ML methods in their capability to offer interpretation. A number of machine learning methods learned on real-world spectral data was considered for testing.

Poursabzi-Sangdeh *et al.* [151] argued that quantifying interpretability implies defining it in terms of alignment with a set of human-interpretable concepts and proposed a general framework called Network Dissection for quantifying the interpretability of latent representations of ANN by identifying hidden units' semantics for any given neural net, then aligning them with human-interpretable concepts.

**TABLE 2.** Summary of explainability techniques.

| Techniques | References | Intrinsic/Post-hoc | Global/Local | Model-specific/ Model-agnostic |
|---|---|---|---|---|
| *Decision trees* | [139], [140], [141], [142], [143] | I | G | SP |
| *Rule lists* | [66], [143], [144], [145], [146] | I | G | SP |
| *LIME* | [84], [85], [102], [147] | H | L | AG |
| *Shapely explanations* | [101] | H | L | AG |
| *Saliency map* | [87], [88], [89], [90], [91], [96], [97] | H | L | AG |
| *Activation maximization* | [82], [83] | H | G | AG |
| *Surrogate models* | [106], [107], [84] | H | G/L | AG |
| *Partial Dependence Plot (PDP)* | [108], [51], [110] | H | G/L | AG |
| *Individual Conditional Expectation (ACI)* | [112], [113] | H | L | AG |
| *Rule extraction* | [74], [114], [115], [116], [117], [118] | H | G/L | AG |
| *Decomposition* | [93], [94], [95] | H | L | AG |
| *Model distillation* | [49], [123], [124], [125], [126], [127] | H | G | AG |
| *Sensitive analysis* | [129], [130] | H | G/L | AG |
| *Layer-wise Relevance Propagation (LRP)* | [131] | H | G/L | AG |
| *Feature importance* | [113], [132], [86] | H | G/L | AG |
| *Prototype and criticism* | [133], [134], [135], [136] | H | G/L | AG |
| *Counterfactuals explanations* | [137] | H | L | AG |

I: Intrinsic, H: Post-hoc, G: Global, L: Local, SP: Model-specific, AG: Model-agnostic

Bau *et al.* [152] have a different perception of explainability, they see it as a latent property that can be influenced by different manipulable factors (such as the number of inputs, the complexity of the model, or even the user interface) and that affects different measurable outcomes (such as an end user's ability to trust or debug the model). They ran in their work related to manipulation and measurement of model interpretability, an interesting experiment which consists of changing factors that are thought to make models more or less interpretable and measuring how these changes affect people's decision making, they focused on two factors: the number of input and whether the model is transparent or black-box. The finding of this experiment stipulates that participants who are presented with a transparent and minimum inputs model are better able to simulate the model's predictions. However, they do not find significant differences in participants' trust or prediction error.

Paul 's claim [153] is based on the fact that a considerable number of methods have been proposed for improving and evaluating the interpretability of topic models and discusses how ideas from topic modeling such as human feedback and automated metrics could be applied to evaluating ML interpretability.

A recent work by Gilpin *et al.* [154] proposed a methodological approach for evaluating interpretability of ML models according to a taxonomy that distinguishes three types of explainability: emulate the processing, explain the representation and explanation-producing networks.

A common factor that directly impacts the quality of explainability and which is approached from different viewpoints in the above studies is: Human. In the next axis, we propose to discuss in detail this factor by highlighting the works focusing on its impact on AI explainability.

## AXIS 3. XAI PERCEPTION: HUMAN IN THE LOOP

Explain and understand are two different actions, explaining depends mainly on what is explained (i.e. the original model) and how explanation is made (i.e. the interpretability method), while understanding depends in addition of these elements on who is receiving the explanation (i.e. explainee,

in other words human). To be explainable, a ML model has to be human-understandable. This represents a challenge for designing XAI as it implies communicating a complex computational process to human which requires in addition of ML expertise, HCI skills as well.

Furthermore, since explanation as a human action has long been studied in philosophy and psychology. Thus, these fields should be consulted in order to simulate the human explanation process and take inspiration from developed models in these fields.

Keeping human in the loop is then a determinant factor of the overall explainability value. However, the conducted literature review has identified the dearth of works focusing on the human factor impact in XAI.

In this axis, we survey works that discussed the role of human from two perspectives: (i) the first one focuses on how to produce explanations that simulate the human cognitive process, while (ii) the second one focuses on how to produce human-centered explanations.

### A. HUMAN-LIKE EXPLANATIONS

The work of Miller [33] is perhaps the most significant attempt at articulating the link between human science and XAI. In his paper, Miller [33] provided an in-depth survey on research in philosophy, psychology, and cognitive science which study the explanation topic. The author noted that the latter could be a valuable resource for the progress of the field of XAI. He highlighted three major findings: (i) Explanations are contrastive: people do not ask why event E happened, but rather why event E happened instead of some event F. (ii) Explanations are selective and focus on one or two possible causes and not all causes for the recommendation. (iii) Explanations are social conversation and interaction for transfer of knowledge, implying that the explainer must be able to leverage the mental model of the explainee while engaging in the explanation process. He asserted that it is imperative to take into account these three points if the goal is to build a useful XAI.

A call for using social science models in XAI was made in [155]. The authors of this paper argue that most of the existing literature on XAI methods are based on the developer's intuitions rather than to be focused on the intended users. Based on a light literature survey, they demonstrate that social science aspects are rarely undertaken in current XAI research and present some key results from human science field that are relevant to XAI.

By going back to the interpretability taxonomy of methods, it is worth to note at this point that post-hoc interpretability techniques are analogous to the human way of explaining decisions. As noted by Lipton [25], ''To the extent that we might consider humans to be interpretable, it is [post-hoc] interpretability that applies''. Furthermore, the example based explanations agnostic method [136] is explicitly inspired by the cognitive science of human reasoning. Specifically, human reasoning is often prototype-based, using representative examples as a basis for categorization and decision-making. Similarly, Kim's method use representative examples to explain and cluster data.

### B. HUMAN-FRIENDLY EXPLANATIONS

An early work by Bauer and Baldes [156] proposed an ontology-based interface that allows (non-expert) user to gain a deeper insight into the knowledge represented by ML models, towards intelligible and transparent ML models.

Recently, Zhu *et al.* [157] noted that most existing works focus on new explaining methods, and not on usability, practical interpretability and efficacy on real users. They introduced a derived research area called eXplainable AI for Designers (XAID) and proposed a human-centered approach for facilitating game designers to co-create with AI/ML techniques through XAID.

Tamagnini *et al.* [158] proposed Rivelo, a pedagogical visual analytics interface that enables expert user of binary classifiers by interactively exploring a set of instance-level explanations.

Abdul *et al.* [61] investigated how HCI research can help to develop practical explainable systems with efficacy for the end users. The authors performed a sizable data-driven literature analysis through which they set an HCI research agenda in explainability. They also pointed the most relevant works that attempt to make explanation human-understandable through interfaces, in textual form or through visual explanations.

Amongst agnostic methods, visualization is the most human-centered technique, indeed this method produce better explanations to see through the black-box, but unfortunately, some techniques belonging to this method produce visualizations that, while visually interesting, are not fully understandable by their human viewers. In a recent work by Hohman *et al.* [159] where a survey of the role of visual analytics in deep learning is presented, the authors acknowledge the importance of producing visualizations and interpretations for DNN that are human understandable and expose works that attempt to produce such visualizations.

## AXIS 4. XAI ANTITHESIS: EXPLAIN OR PREDICT

So far, we have presented works that support XAI from different perspectives. Off the beaten path, and before presenting our synthesized ideas, we propose in this axis to expose works that challenge typical approaches, adjust intuitive beliefs and conjecture previous findings regarding XAI. Seeking to be holistic, the aim here is to propose a structured proposal that respects the triad thesis, antithesis and synthesis.

In their work ''Are Explanations Always Important?'' Bunt *et al.* [160] raised questions as to the importance of, and consequently anticipated usage of, explanation techniques within systems that support users in making low-cost decisions. Based on their studies, they found that generally these opaque intelligent systems are positively perceived despite the lack of meaningful or accessible explanation. They noted: ''While some users were interested in accessing more information, the dominant responses were that the applications

were sufficiently transparent, or that the cost of viewing an explanation would outweigh the benefit''. XAI is thus not yet ready to penetrate the market of this kind of intelligent systems.

''Too Much, Too Little, or Just Right?'' is a work proposed by Kulesza *et al.* [161] where they presented their findings regarding how explanations impact end users' mental models. Interestingly, they suggest that completeness is more important than soundness in explanation: increasing completeness via certain information types helped user's mental models and, surprisingly, their perception of the cost/benefit tradeoff of attending to the explanations. They also found that contrary to what the human-friendly explanation would have us believe, oversimplification can be a problem: ''when soundness was very low, user experienced more mental demand and lost trust in the explanations, thereby reducing the likelihood that users will pay attention to such explanations at all''.

Even though the work by Holliday *et al.* [162] entitled ''User Trust in Intelligent Systems: A Journey Over Time is work'', confirmed based on experimental studies that explanation impact trust. This work challenges the typical approach that considered trust in intelligent systems is only captured as a single quantitative measure at the conclusion of a task.

Most research works on the ML interpretability agreed and contribute towards more rigorous notion of interpretability [62]. In contrast to this wave of thoughts, Offert [163] suggested in his work ''I know it when I see it'' that a better understanding of the deficiencies of the intuitive notion of interpretability is needed as well. That is we have to ''consider interpretability precisely in terms of what it is not'' in order to identify where it is impaired by intuitive considerations.

Wang *et al.* [164] proposed in their work ''Trading Interpretability for Accuracy'' an Oblique Treed Sparse Additive Models that sacrifices a certain degree of interpretability for accuracy in order to achieve entirely sufficient accuracy.

From a statistical standpoint, Shmueli [165] debated ''To Explain or to Predict?'' dilemma by giving a special emphasis on machine learning field.

Finally, by taking machine learning as a model where prediction is more important than explanation, Yarkoni and Westfall [166] argued that in psychology ''an increased focus on prediction, rather than explanation, can ultimately lead us to a greater understanding of behavior''. They named their work: ''Choosing Prediction Over Explanation in Psychology: Lessons From Machine Learning''.

## IV. DISCUSSION

Due to the broad spectrum of XAI approaches, it is almost impossible to perform an exhaustive survey of all XAI works. It is also inconceivable and unthinkable to include all the 381 studied papers in this work, thus for synthesis and relevance concerns, only a subset of works was detailed in this survey. As mentioned before, the selection criteria were mostly based on the popularity and impact of proposals. As a supplement, at each axis, we made sure to include fresh works in order to give interested researchers an idea about recent trends.

The proposed review was underpinned by a solid background that covers all aspects related to the XAI topic. In the background section, we deliberately include non-academic venues with significant attention. Indeed, due to the youth of the studied domain and its rapid growth, it turns out that these non-traditional sources are also important to review, as they are highly influential and impactful to the field.

We conclude our survey with a compilation of the main findings as well as interesting facts from previous studies. In parallel, we discuss some research directions and open problems distilled from the surveyed works.

### A. TOWARDS MORE FORMALISM

XAI is a multifaceted objective that cannot be addressed by singular disciplinary efforts. However, synergetic use of methods from different research horizons must be done in a soundly integrated way. In other words, for the field to progress, it should be supported by a standalone research community who, at this stage of advancement, should mainly be engaged towards more formalism in terms of:

*(i) Systematic definitions:* depending on their background, researchers use synonymously concepts that are semantically different [25], and refers to the same notions by different names (e.g. [113] and [132]). A consensus on definitions must be done in order to enable easier transfer of results and information.

*(ii) Abstraction:* given the number of research proposals, there is a sufficient material for efforts consolidation in form of a generic explainable framework that would guide the production of end-to-end explainable approaches. Instead of isolated interpretability methods that though their technical relevance remain only fragments of the whole solution, which is larger than a technical operation on ML algorithm outcomes.

In this vein, abstracted explanation generation is another potential venue. Dosilovic *et al.* [64] discussed in their work, the utility of such abstraction in finding properties and generating hypotheses about data-generating processes, which is important for future Artificial General Intelligence (AGI) systems.

*(iii) Formalizing and quantifying:* Guidotti *et al.* [63], Dhurandhar *et al.* [167], Puri *et al.* [168], and Varshney *et al.* [169] tend to base their proposal on a detailed problem formulation that becomes invalid as soon as the method of interpretability or the explained model change. To further progress in this field, it is imperative to generalize the expansibility problem formulation in a rigorous way, irrespective of changing factors and variables. As a direct effect, this will advance the state of art of explainability classifying, qualifying and evaluating sub-issues.

Indeed, with the amount of the existing explainability methods in the literature, the first area for future work is developing formalized rigorous evaluation metrics

and methods. Otherwise, we risk to be forced to explain explanation. As observed before, in literature, there is no clear way to quantify explainability, the related line of work is just in its infancy, which represents an opportunity and a challenge at the same time.

### B. HUMAN-MACHINE TEAMING

It is not enough to just explain the model, the user has to understand it. However, even with an accurate explanation, developing such an understanding could require supplementary answers for questions that users would likely have. Thus, explainability can only happen throw interaction between human and machine. Envisioning interactive explanation systems that support many different follow-up and drill-down actions after presenting an initial explanation to the user, is a potential research path to pursue in order to advance the XAI field.

This two-way partnership motivates naturally the use of HCI and human sciences disciplines. Nevertheless as discussed before, there is a lack of literature around explainable systems that take into account these two dimensions. Two keen observations made respectively by Miller [33] and Abdul *et al.* [61] attest to this lack: (i) ''social sciences and human behavioral studies are not having enough impact in explainable AI'' and (ii) ''the streams of research in explainable systems and in the HCI community tend to be relatively isolated''. The challenge is then to link the results of HCI empirical studies with human science theories in order to drive from both of them added value into explainability approaches and hopefully contribute to more human-centric explainable models. Consequently, adaptive explainable models would make their appearance, by offering context-aware explanations that would adapt according to their environment changes such as: the user profile (level of expertise, domain knowledge, cultural background, interests and preferences and other contextual variables) and the explanation request setting (justification, teaching, audit . . .).

If machine human teaming is expected to spark significant research in AI explanability. In the era of Internet of Things (IoT) we should also be waiting for the emerging of another research body focusing on the machine-to-machine explanation. Conceiving explanation for machine consumption will drive some considerations that worth further research. Ultimately, however, it is likely that future explainable approaches, especially adaptive one, will need to provide both kinds of explanation.

### C. EXPLAINABILITY METHODS COMPOSITION

Work on explainability tends to advance quantitatively interpretability methods reflected in a huge proliferation of interpretability techniques (which by the way makes defining a taxonomy of interpretability methods a challenging task). Comparatively, little attention is given to approaches that discussed the potential of combining different interpretability methods to achieve a more powerful explanation. Indeed, in literature we have seen how some techniques can be used in complementary to others (e.g. sensitive analysis and visualization), but not how to treat disparate interpretability methods as elementary and composable building blocks that could synergistically create new added value techniques. We believe this is a rich, under-explored area for future research. Hence, enabling composability in XAI can potentially contribute to effectively solve optimization issue in this field and making explainability and accuracy move in the same direction.

Furthermore, an eventual combination could also concern actual interpretability methods focusing on ML models and classical solutions of explainability related to expert systems. How exactly to combine elements from both classical explainable expert systems and present interpretable ML methods is a topic of debate, for instance, Preece [57] argues that elements of that earlier work on expert systems offer routes to making progress towards XAI today.

### D. OTHERS EXPLAINABLE INTELLIGENT SYSTEMS

Most of the existing works in literature focus on explainability in machine learning, which is just one type of AI. However, the same issues also confront other intelligent systems. Particularly (i) explainable AI planning and (ii) explainable agent are beginning to gain recognition as a promising derived field of XAI.

Planning is an important area of AI, it is used in domains where learning is not an option. Where the planner is mostly concerned with establishing the correctness and quality of a given plan with respect to its own model, adding expandability implies to translate the produced plan steps (e.g., PDDL plans) in a human understandable form. Thus, intuitively, explainable AI planning (XAIP) is mostly algorithm dependent and serve more as a debugging system for an expert user. The explainability opportunities that arise in AI planning was recently explored by Fox *et al.* [170]. They described some initial analysis of the issue and proposed a roadmap for achieving an effective explainability. They based their idea on the fact that, in contrast of ML, AI Planning is potentially more favorably disposed to be explainable: (i) planners can eventually be trusted, (ii) planners can allow an easy interaction with humans, and (iii) planners are relatively transparent. In fact, the main issue in enabling explainability in planners is mainly related to the gap between planning algorithms and human problem solving.

Other introductory works include [171] and [172], in addition of a dozen papers that was discussed very recently in a dedicated session of the ICAPS Workshop [11].

These initial works open up a number of future directions for XAIP, begin by a need of a full formulation of the explainable planning problem, other open problems include the feasibility explainability in case of planning under uncertainty and explainability of task and motion planning in robotics. A revision of the rich AI planning literature to identify works that could contribute to XAIP can also serve as a useful starting point for progress in this subfield.

In a related thread of work, researchers have looked at the idea of explainable agents. As agents are supposed to represent human behavior, works in this area mainly focus on behavior explanations generation so that agents could explain the reasons behind their actions. Significant earlier works proposed approaches for self-explaining virtual agents in scenario-based training systems [173], [174]. Recent works have investigated the possibilities of explainability of advanced autonomous agents like socially interactive robots [175]–[178].

### E. ECONOMIC PERSPECTIVE

There are significant benefits to gain from being on the front foot and investing in explainability today. Indeed, due to the social and ethical pressure, XAI could turn into a competitive differentiator and drives a real business value. Moreover, the technical challenge of explainability involving the tradeoff between accuracy and interpretability, affects significantly the cost of XAI products. Curiously, in literature economical perspective of XAI is an area of less research yet no less important. Encouraging economic interpretations is essential to address several issues such as explainability cost estimation and variation, algorithm propriety, revealing trade secrets and predicting XAI market evolution.

Amongst few works found in literature, Akyol *et al.* [179] proposed a first attempt to quantitatively analyze the cost of transparency (PoT) in ML algorithms. The work of Igami [180] about the connections between machine learning and econometrics, proposed the perspective of ''Structural Econometrics for Explainable AI''. He noted that ''relaxing the implicit econometric assumptions would make the results economically interpretable''.

The discussed outlooks are by no means exhaustive, but give a few leads for further exploration from different perspectives based on our compilation of existent studies in literature. We hope that the proposed directions will inspire new research that can improve the current state of the art in XAI.

## V. CONCLUSION

Matter-of-factly, XAI is a vital interdisciplinary research field in the AI ecosystem. In the spirit of holism, we presented in this paper a comprehensive background regarding this field. Taking inspiration from how we assimilate and familiarize ourselves with new topics, we focused on the Five W's and How (What, Who, When, Why, Where, and How) to cover all aspects related to XAI. Moreover, in the interest of mapping the broad landscape around XAI research, this survey has thoroughly reviewed a portfolio of explainability approaches and organized them from different perspectives.

Findings showed that XAI is not just a labcoat research field, its impact is spanning in a large range of application domains. However, we have seen evidence throughout this work for the lack of formalism in terms of problem formulation and clear unambiguous definitions. Furthermore, it has been noted that the human's role is not sufficiently studied in existing explainability approaches. In essence, attention is devoted to interpreting ML models letting other promising AI system explainability under-explored. It has then been concluded that considerable effort will be required in the future to tackle the challenges and open issues with XAI.

## REFERENCES

[1] International Data Corporation IDC. (2018). *Worldwide Semiannual Cognitive Artificial Intelligence Systems Spending Guide*. Accessed: Jun. 6, 2018. [Online]. Available: https://www.idc.com/getdoc.jsp?containerId=prUS43662418
[2] Statista. (2018). *Revenues From the Artificial Intelligence (AI) Market Worldwide From 2016 to 2025*. Accessed: Jun. 6, 2018. [Online]. Available: https://www.statista.com/statistics/607716/worldwide-artificial-intelligence-market-revenues/
[3] Gartner. (2017). *Top 10 Strategic Technology Trends for 2018*. Accessed: Jun. 6, 2018. [Online]. Available: https://www.gartner.com/doc/3811368?srcId=1-6595640781
[4] S. Barocas, S. Friedler, M. Hardt, J. Kroll, S. Venka-Tasubramanian, and H. Wallach. *The FAT-ML Workshop Series on Fairness, Accountability, and Transparency in Machine Learning*. Accessed: Jun. 6, 2018. [Online]. Available: http://www.fatml.org/
[5] B. Kim, K. R. Varshney, and A. Weller. *2018 Workshop on Human Interpretability in Machine Learning (WHI)*. [Online]. Available: https://sites.google.com/view/whi2018/
[6] A. G. Wilson, B. Kim, and W. Herlands. (2016). *Proceedings of NIPS 2016 Workshop on Interpretable Machine Learning for Complex Systems*. [Online]. Available: https://arxiv.org/abs/1611.09139
[7] D. W. Aha, T. Darrell, M. Pazzani, D. Reid, C. Sammut, and P. Stone, in *Proc. Workshop Explainable AI (XAI) IJCAI*, 2017.
[8] M. P. Farina and C. Reed, in *Proc. XCI, Explainable Comput. Intell. Workshop*, 2017.
[9] I. Guyon *et al.*, in *Proc. IJCNN Explainability Learn. Mach.*, 2017.
[10] A. Chander *et al.*, in *Proc. MAKE-Explainable AI*, 2018.
[11] S. Biundo, P. Langley, D. Magazzeni, and D. Smith, in *Proc. ICAPS Workshop, EXplainable AI Planning*, 2018.
[12] M. Graaf, B. Malle, A. Dragan, and T. Ziemke, in *Proc. HRI Workshop, Explainable Robot. Syst.*, 2018.
[13] T. Komatsu and A. Said, in *Proc. ACM Intell. Interfaces (IUI) Workshop, Explainable Smart Syst. (EXSS)*, 2018.
[14] J. M. Alonso, C. Castiello, C. Mencar, and L. Magdalena, in *Proc. IPMU, Adv. Explainable Artif. Intell.*, 2018.
[15] B. D. Agudo, D. Aha, and J. R. Garcia, in *Proc. ICCBR, 1st Workshop Case-Based Reasoning Explanation Intell. Syst. (XCBR)*, 2018.
[16] D. Gunning. Explainable artificial intelligence (XAI), Defense Advanced Research Projects Agency (DARPA). Accessed: Jun. 6, 2018. [Online]. Available: http://www.darpa.mil/program/explainable-artificial-intelligence
[17] P. Hall, M. Kurka, and A. Bartz. (2018). *Using $H_2O$ Driverless AI, $H_2O.AI$*. Accessed: Jun. 6, 2018. [Online]. Available: https://www.h2o.ai/wp-content/uploads/2018/01/DriverlessAIBooklet.pdf
[18] Cognilytica. (2018). *Cognilytica's AI Positioning Matrix (CAPM)*. Accessed: Jun. 6, 2018. [Online]. Available: https://www.cognilytica.com/2018/01/09/cognilyticas-ai-positioning-matrix-capm/
[19] FICO. (2018). *Explainable Machine Learning Challenge*. Accessed: Jun. 6, 2018. [Online]. Available: https://community.fico.com/s/explainable-machine-learning-challenge
[20] M. van Lent, W. Fisher, and M. Mancuso, ''An explainable artificial intelligence system for small-unit tactical behavior,'' in *Proc. 16th Conf. Innov. Appl. Artif. Intell.*, 2004, pp. 900–907.
[21] W. R. Swartout and J. D. Moore, ''Explanation in expert systems: A survey,'' Univ. Southern California, Los Angeles, CA, USA, Tech. Rep. ISI/RR-88-228, 1988.
[22] D. Doran, S. Schulz, and T. R. Besold. (2017). ''What does explainable AI really mean? A new conceptualization of perspectives.'' [Online]. Available: https://arxiv.org/abs/1710.00794
[23] P. W. Koh and P. Liang. (2017). ''Understanding black-box predictions via influence functions.'' [Online]. Available: https://arxiv.org/abs/1703.04730
[24] M. Bojarski *et al.* (2017). ''Explaining how a deep neural network trained with end-to-end learning steers a car.'' [Online]. Available: https://arxiv.org/abs/1704.07911

[25] Z. C. Lipton, "The mythos of model interpretability," in *Proc. ICML Workshop Hum. Interpretability Mach. Learn.*, 2016, pp. 96–100.

[26] A. Andrzejak, F. Langner, and S. Zabala, "Interpretable models from distributed data via merging of decision trees," in *Proc. CIDM*, Singapore, Apr. 2013, pp. 1–9.

[27] R. Piltaver, M. Luštrek, M. Gams, and S. Martinčić-Ipšić, "Comprehensibility of classification trees—Survey design validation," in *Proc. ITIS*, Šmarješke toplice, Slovenia, 2014, pp. 5–7.

[28] D. S. Weld and G. Bansal. (2018). "The challenge of crafting intelligible intelligence." [Online]. Available: https://arxiv.org/abs/1803.04263

[29] R. R. Suman, R. Mall, S. Sukumaran, and M. Satpathy, "Extracting state models for black-box software components," *J. Object Technol.*, vol. 9, no. 3, pp. 79–103, 2010.

[30] V. Dignum, "Responsible artificial intelligence: Designing AI for human values," *ITU J., ICT Discoveries*, vol. 1, pp. 1–8, Sep. 2017.

[31] K. J. Danjuma. (2015). "Performance evaluation of machine learning algorithms in post-operative life expectancy in the lung cancer patients." [Online]. Available: https://arxiv.org/abs/1504.04646

[32] L. M. Chen, "Machine learning for data science: Mathematical or computational," in *Mathematical Problems in Data Science*. 2015, pp. 63–74.

[33] T. Miller. (2017). "Explanation in artificial intelligence: Insights from the social sciences." [Online]. Available: https://arxiv.org/abs/1706.07269

[34] A. Prabhakar, "Powerful but limited: A DARPA perspective on AI," in *Proc. DARPA*, 2017. Accessed: Jun. 6, 2018. [Online]. Available: https://sites.nationalacademies.org/cs/groups/pgasite/documents/webpage/pga_177035.pdf

[35] S. Baum, "A survey of artificial general intelligence projects for ethics, risk, and policy," Global Catastrophic Risk Inst., Working Paper 17-1, 2017. [Online]. Available: https://gcrinstitute.org/about/

[36] K. S. Gill, "Artificial super intelligence: Beyond rhetoric," *AI Soc.*, vol. 31, no. 2, pp. 137–143, 2016.

[37] R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, and N. Elhadad, "Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission," in *Proc. 21th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2015, pp. 1721–1730.

[38] A. Howard, C. Zhang, and E. Horvitz, "Addressing bias in machine learning algorithms: A pilot study on emotion recognition for intelligent systems," in *Proc. Adv. Robot. Social Impacts (ARSO)*, Mar. 2017, pp. 1–7.

[39] (2016). *European Union General Data Protection Regulation (GDPR)*. Accessed: Jun. 6, 2018. [Online]. Available: http://www.eugdpr.org/

[40] D. Silver *et al.*, "Mastering the game of go without human knowledge," *Nature*, vol. 550, no. 7676, pp. 354–359, 2017.

[41] P. Norvig. Google's approach to artificial intelligence and machine learning, UNSW, Sydney, NSW, Australia. Accessed: Jun. 6, 2018. [Online]. Available: https://www.engineering.unsw.edu.au/video/googles-approach-to-artificial-intelligence-and-machine-learning

[42] M. McFarland. (2018). *Uber Shuts Down Self-Driving Operations in Arizona, CNN*. Accessed Jun. 6, 2018. [Online]. Available: http://money.cnn.com/2018/05/23/technology/uber-arizona-self-driving/index.html

[43] M. Bojarski *et al.* (2016). "End to end learning for self-driving cars." [Online]. Available: https://arxiv.org/abs/1604.07316

[44] J. Haspiel *et al.* (2018). *Explanations and Expectations: Trust Building in Automated Vehicles, deepblue.lib.umich.edu*. [Online]. Available: https://doi.org/10.1145/3173386.3177057

[45] A. Holzinger, C. Biemann, C. S. Pattichis, and D. B. Kell. (2017). "What do we need to build explainable AI systems for the medical domain?" [Online]. Available: https://arxiv.org/abs/1712.09923

[46] G. J. Katuwal and R. Chen. (2016). *Machine Learning Model Interpretability for Precision Medicine*. [Online]. Available: https://arxiv.org/abs/1610.09045

[47] Z. Che , S. Purushotham, R. Khemani, and Y. Liu, "Interpretable deep models for ICU outcome prediction," in *Proc. AMIA Annu. Symp.*, 2017, pp. 371–380.

[48] J. Lightbourne, "Damned lies & criminal sentencing using evidence-based tools," 15 Duke Law & Technol. Rev., Tech. Rep., 2017, pp. 327–343. Accessed: Jun. 6, 2018. https://scholarship.law.duke.edu/dltr/vol15/iss1/16

[49] S. Tan, R. Caruana, G. Hooker, and Y. Lou. (2018). "Detecting bias in black-box models using transparent model distillation." [Online]. Available: https://arxiv.org/abs/1710.06169

[50] C. Howell, "A framework for addressing fairness in consequential machine learning," in *Proc. FAT Conf., Tuts.*, 2018, pp. 1–2.

[51] R. Berk and J. Bleich, "Statistical procedures for forecasting criminal behavior: A comparative assessment," *Criminol. Public Policy*, vol. 12, no. 3, pp. 513–544, 2013.

[52] Equifax. (2018). *Equifax Launches NeuroDecision Technology*. Accessed: Jun. 6, 2018. [Online]. Available: https://investor.equifax.com/news-and-events/news/2018/03-26-2018-143044126

[53] W. Knight. (2017). The U.S. military wants its autonomous machines to explain themselves, MIT Technology Review. Accessed: Jun. 6, 2018. [Online]. Available: https://www.technologyreview.com/s/603795/the-us-military-wants-its-autonomous-machines-to-explain-themselves

[54] A. Henelius, K. Puolamäki, and A. Ukkonen. (2017). "Interpreting classifiers through attribute interactions in datasets." [Online]. Available: https://arxiv.org/abs/1707.07576

[55] Future of Privacy Forum. (2017). *Unfairness by Algorithm: Distilling the Harms of Automated Decision-Making*. Accessed: Jun. 6, 2018. [Online]. Available: https://fpf.org/wp-content/uploads/2017/12/FPF-Automated-Decision-Making-Harms-and-Mitigation-Charts.pdf

[56] R. Kass, T.Finin , "The need for user models in generating expert system explanations," PENN library, Tech. Rep., 1988.

[57] A. Preece, "Asking 'Why' in AI: Explainability of intelligent systems—Perspectives and challenges," *Intell. Syst. Accounting, Finance Manage.*, vol. 25, no. 1, pp. 63–72, 2018

[58] S. Tan, K. C. Sim, and M. Gales, "Improving the interpretability of deep neural networks with stimulated learning," in *Proc. IEEE Workshop Autom. Speech Recognit. Understand. (ASRU)*, Dec. 2015, pp. 617–623.

[59] P. Hall and N. Gill, *An Introduction to Machine Learning Interpretability*. Newton, MA, USA: O'Reilly Media, 2018.

[60] L. Breiman. (2001). *Statistical Modeling: The Two Cultures, Statistical Science*. Accessed: Jun. 6, 2018. [Online]. Available: http://bit.ly/2pwz6m5

[61] A. Abdul, J. Vermeulen, D. Wang, B. Y. Lim, and M. Kankanhalli, "Trends and trajectories for explainable, accountable and intelligible systems: An HCI research agenda," in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst. (CHI)*, 2018, p. 582.

[62] F. Doshi-Velez and B. Kim. (2018). "Towards a rigorous science of interpretable machine learning." [Online]. Available: https://arxiv.org/abs/1702.08608

[63] R. Guidotti, A. Monreale, F. Turini, D. Pedreschi, and F. Giannotti. (2018). "A survey of methods for explaining black box models." [Online]. Available: https://arxiv.org/abs/1802.01933

[64] F. K. Došilovič, M. Brčić, and N. Hlupić, "Explainable artificial intelligence: A survey," in *Proc. 41st Int. Conv. Inf. Commun. Technol., Electron. Microelectron. (MIPRO)*, May 2018, pp. 0210–0215.

[65] C. Wohlin, "Guidelines for snowballing in systematic literature studies and a replication in software engineering," in *Proc. 18th Int. Conf. Eval. Assessment Softw. Eng. (EASE)*, 2014, Art. no. 38.

[66] B. Letham, C. Rudin, T. H. McCormick, and D. Madigan, "Interpretable classifiers using rules and Bayesian analysis: Building a better stroke prediction model," *Ann. Appl. Statist.*, vol. 9, no. 3, pp. 1350–1371, 2015.

[67] K. Xu *et al.*, "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2015, pp. 1–10.

[68] B. Ustun and C. Rudin, "Supersparse linear integer models for optimized medical scoring systems," *Mach. Learn.*, vol. 102, no. 3, pp. 349–391, 2015.

[69] S. Sarkar, "Accuracy and interpretability trade-offs in machine learning applied to safer gambling," in *Proc. CEUR Workshop*, 2016, pp. 79–87.

[70] L. Breiman, "Statistical modeling: The two cultures (with comments and a rejoinder by the author)," *Stat. Sci.*, vol. 16, no. 3, pp. 199–231, 2001.

[71] S. Krening, B. Harrison, K. M. Feigh, C. L. Isbell, M. Riedl, and A. Thomaz, "Learning from explanations using sentiment and advice in RL," *IEEE Trans. Cogn. Develop. Syst.*, vol. 9, no. 1, pp. 44–55, Mar. 2016.

[72] A. Mahendran and A. Vedaldi, "Understanding deep image representations by inverting them," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 5188–5196.

[73] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2013, pp. 3111–3119.

[74] G. Ras, M. van Gerven, and P. Haselager. (2018). "Explanation methods in deep learning: Users, values, concerns and challenges." [Online]. Available: https://arxiv.org/abs/1803.07517

[75] A. Santoro *et al.* (2017). "A simple neural network module for relational reasoning." [Online]. Available: https://arxiv.org/abs/1706.01427
[76] R. B. Palm, U. Paquet, and O. Winther. (2017). "Recurrent relational networks for complex relational reasoning." [Online]. Available: https://arxiv.org/abs/1711.08028
[77] Y. Dong, H. Su, J. Zhu, and B. Zhang, "Improving interpretability of deep neural networks with semantic information," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Mar. 2017, pp. 4306–4314.
[78] C. Louizos, U. Shalit, J. M. Mooij, D. Sontag, R. Zemel, and M. Welling, "Causal effect inference with deep latent-variable models," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2017, pp. 6446–6456.
[79] O. Goudet *et al.* (2017). "Learning functional causal models with generative neural networks." [Online]. Available: https://arxiv.org/abs/1709.05321
[80] C. Yang, A. Rangarajan, and S. Ranka. (2018). "Global model interpretation via recursive partitioning." [Online]. Available: https://arxiv.org/abs/1802.04253
[81] M. A. Valenzuela-Escárcega, A. Nagesh, and M. Surdeanu. (2018). "Lightly-supervised representation learning with global interpretability." [Online]. Available: https://arxiv.org/abs/1805.11545
[82] A. Nguyen, A. Dosovitskiy, J. Yosinski, T. Brox, and J. Clune, "Synthesizing the preferred inputs for neurons in neural networks via deep generator networks," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2016, pp. 3387–3395.
[83] D. Erhan, A. Courville, and Y. Bengio, "Understanding representations learned in deep architectures," Dept. d'Informatique Recherche Operationnelle, Univ. Montreal, Montreal, QC, Canada, Tech. Rep. 1355, 2010.
[84] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should i trust you?': Explaining the predictions of any classifier," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2016, pp. 1135–1144.
[85] M. T. Ribeiro, S. Singh, and C. Guestrin, "Anchors: High-precision model-agnostic explanations," in *Proc. AAAI Conf. Artif. Intell.*, 2018, pp. 1–9.
[86] J. Lei, M. G'Sell, A. Rinaldo, R. J. Tibshirani, and L. Wasserman, "Distribution-free predictive inference for regression," *J. Amer. Stat. Assoc.*, to be published. [Online]. Available: http://www.stat.cmu.edu/~ryantibs/papers/conformal.pdf
[87] D. Baehrens, T. Schroeter, S. Harmeling, M. Kawanabe, K. Hansen, and K.-R. Müller, "How to explain individual classification decisions," *J. Mach. Learn. Res.*, vol. 11, no. 6, pp. 1803–1831, 2010.
[88] K. Simonyan, A. Vedaldi, and A. Zisserman. (2013). "Deep inside convolutional networks: Visualising image classification models and saliency maps." [Online]. Available: https://arxiv.org/abs/1312.6034
[89] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proc. Eur. Conf. Comput. Vis.* Zurich, Switzerland: Springer, 2014, pp. 818–833.
[90] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and O. Torralba, "Learning deep features for discriminative localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 2921–2929.
[91] M. Sundararajan, A. Taly, and Q. Yan. (2017). "Axiomatic attribution for deep networks." [Online]. Available: https://arxiv.org/abs/1703.01365
[92] D. Smilkov, N. Thorat, B. Kim, F. Viégas, and M. Wattenberg. (2017). "SmoothGrad: Removing noise by adding noise." [Online]. Available: https://arxiv.org/abs/1706.03825
[93] M. Robnik-Šikonja and I. Kononenko, "Explaining classifications for individual instances," *IEEE Trans. Knowl. Data Eng.*, vol. 20, no. 5, pp. 589–600, May 2008.
[94] G. Montavon, S. Lapuschkin, A. Binder, W. Samek, and K.-R. Müller, "Explaining nonlinear classification decisions with deep taylor decomposition," *Pattern Recognit.*, vol. 65, pp. 211–222, May 2017.
[95] S. Bach, A. Binder, K.-R. Müller, and W. Samek, "Controlling explanatory heatmap resolution and semantics via decomposition depth," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2016, pp. 2271–2275.
[96] R. Fong and A. Vedaldi. (2017). "Interpretable explanations of black boxes by meaningful perturbation." [Online]. Available: https://arxiv.org/abs/1704.03296
[97] P. Dabkowski and Y. Gal, "Real time image saliency for black box classifiers," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 6970–6979.
[98] P.-J. Kindermans *et al.*, "Learning how to explain neural networks: PatternNet and patternAttribution," in *Proc. Int. Conf. Learn. Represent.*, 2018, pp. 1–16. Accessed: Jun. 6, 2018. [Online]. Available: https://openreview.net/forum?id=Hkn7CBaTW
[99] A. Shrikumar, P. Greenside, A. Shcherbina, and A. Kundaje. (2016). "Not just a black box: Interpretable deep learning by propagating activation differences." [Online]. Available: http://arxiv.org/abs/1605.01713
[100] A. Ross, M. C. Hughes, and F. Doshi-Velez, "Right for the right reasons: Training differentiable models by constraining their explanations," in *Proc. Int. Joint Conf. Artif. Intell.*, 2017, pp. 2662–2670.
[101] S. M. Lundberg and S. I. Lee, "A unified approach to interpreting model predictions," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 4768–4777.
[102] R. Guidotti, A. Monreale, S. Ruggieri, D, Pedreschi, F. Turini, and F. Giannotti. (2018). "Local rule-based explanations of black box decision systems." [Online]. Available: https://arxiv.org/abs/1805.10820
[103] D. Linsley, D. Scheibler, S. Eberhardt, and T. Serre. (2018). "Global-and-local attention networks for visual recognition." [Online]. Available: https://arxiv.org/abs/1805.08819
[104] S. Seo, J. Huang, H. Yang, and Y. Liu, "Interpretable convolutional neural networks with dual local and global attention for review rating prediction," in *Proc. 11th ACM Conf. Recommender Syst. (RecSys)*, 2017, pp. 297–305.
[105] C. Molnar. (2018). *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. Accessed: Jun. 6, 2018. [Online]. Available: https://christophm.github.io/interpretable-ml-book/
[106] O. Bastani, C. Kim, and H. Bastani. (2017). "Interpretability via model extraction." [Online]. Available: https://arxiv.org/abs/1706.09773
[107] J. J. Thiagarajan, B. Kailkhura, P. Sattigeri, and K. N. Ramamurthy. (2016). "TreeView: Peeking into deep neural networks via feature-space partitioning." [Online]. Available: https://arxiv.org/abs/1611.07429
[108] D. P. Green and H. L. Kern, "Modeling heterogeneous treatment effects in large-scale experiments using Bayesian additive regression trees," in *Proc. Annu. Summer Meeting Soc. Political Methodol.*, 2010, pp. 1–40.
[109] H. A. Chipman, E. I. George, and R. E. McCulloch, "BART: Bayesian additive regression trees," *Appl. Statist.*, vol. 4, no. 1, pp. 266–298, 2010.
[110] J. Elith, J. Leathwick, and T. Hastie, "A working guide to boosted regression trees," *J. Animal Ecol.*, vol. 77, no. 4, pp. 802–813, 2008.
[111] S. H. Welling, H. H. F. Refsgaard, P. B. Brockhoff, and L. H. Clemmensen. (2016). "Forest floor visualizations of random forests." [Online]. Available: https://arxiv.org/abs/1605.09196
[112] A. Goldstein, A. Kapelner, J. Bleich, and E. Pitkin, "Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation," *J. Comput. Graph. Statist.*, vol. 24, no. 1, pp. 44–65, 2015, doi: 10.1080/10618600.2014.907095.
[113] G. Casalicchio, C. Molnar, and B. Bischl. (2018). "Visualizing the feature importance for black box models." [Online]. Available: https://arxiv.org/abs/1804.06620
[114] U. Johansson, R. König, and I. Niklasson, "The truth is in there—Rule extraction from opaque models using genetic programming," in *Proc. FLAIRS Conf.*, 2004, pp. 658–663.
[115] M. H. Aung *et al.*, "Comparing analytical decision support models through Boolean rule extraction: A case study of ovarian tumour malignancy," in *Proc. Int. Symp. Neural Netw.* Berlin, Germany: Springer, 2007, pp. 1177–1186.
[116] T. Hailesilassie. (2017). "Rule extraction algorithm for deep neural networks: A review." [Online]. Available: https://arxiv.org/abs/1610.05267
[117] R. Andrews, J. Diederich, and A. B. Tickle, "Survey and critique of techniques for extracting rules from trained artificial neural networks," *Knowl.-Based Syst.*, vol. 8, no. 6, pp. 373–389, 1995.
[118] T. GopiKrishna, "Evaluation of rule extraction algorithms," *Int. J. Data Mining Knowl. Manage. Process*, vol. 4, no. 3, pp. 9–19, 2014.
[119] T. A. Etchells and P. J. G. Lisboa, "Orthogonal search-based rule extraction (OSRE) for trained neural networks: A practical and efficient approach," *IEEE Trans. Neural Netw.*, vol. 17, no. 2, pp. 374–384, Mar. 2006.
[120] N. Barakat and J. Diederich, "Eclectic rule-extraction from support vector machines," *Int. J. Comput. Intell.*, vol. 2, no. 1, pp. 59–62, 2005.
[121] P. Sadowski, J. Collado, D. Whiteson, and P. Baldi, "Deep learning, dark knowledge, and dark matter," in *Proc. NIPS Workshop High-Energy Phys. Mach. Learn. (PMLR)*, vol. 42, 2015, pp. 81–87.
[122] G. Hinton, O. Vinyals, and J. Dean. (2015). "Distilling the knowledge in a neural network." [Online]. Available: https://arxiv.org/abs/1503.02531
[123] Z. Che, S. Purushotham, R. Khemani, and Y. Liu. (2015). "Distilling knowledge from deep networks with applications to healthcare domain." [Online]. Available: https://arxiv.org/abs/1512.03542

[124] K. Xu, D. H. Park, D. H. Yi, and C. Sutton. (2018). "Interpreting deep classifier by visual distillation of dark knowledge." [Online]. Available: https://arxiv.org/abs/1803.04042
[125] S. Tan, "Interpretable approaches to detect bias in black-box models," in *Proc. AAAI/ACM Conf. AI Ethics Soc.*, 2017, pp. 1–2.
[126] S. Tan, R. Caruana, G. Hooker, and Y. Lou. (2018). "Auditing black-box models using transparent model distillation with side information." [Online]. Available: https://arxiv.org/abs/1710.06169
[127] S. Tan, R. Caruana, G. Hooker, and A. Gordo. (2018). "Transparent model distillation." [Online]. Available: https://arxiv.org/abs/1801.08640
[128] Y. Zhang and B. Wallace. (2016). "A sensitivity analysis of (and practitioners' Guide to) convolutional neural networks for sentence classification." [Online]. Available: https://arxiv.org/abs/1510.03820
[129] P. Cortez and M. J. Embrechts, "Using sensitivity analysis and visualization techniques to open black box data mining models," *Inf. Sci.*, vol. 225, pp. 1–17, Mar. 2013.
[130] P. Cortez and M. J. Embrechts, "Opening black box data mining models using sensitivity analysis," in *Proc. IEEE Symp. Comput. Intell. Data Mining (CIDM)*, Apr. 2011, pp. 341–348.
[131] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PLoS ONE*, vol. 10, no. 7, p. e0130140, 2015.
[132] A. Fisher, C. Rudin, and F. Dominici. (2018). "Model class reliance: Variable importance measures for any machine learning model class, from the 'rashomon' perspective." [Online]. Available: https://arxiv.org/abs/1801.01489
[133] J. Bien and R. Tibshirani, "Prototype selection for interpretable classification," *Ann. Appl. Statist.*, vol. 5, no. 4, pp. 2403–2424, 2011.
[134] B. Kim, C. Rudin, and J. A. Shah, "The Bayesian case model: A generative approach for case-based reasoning and prototype classification," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 1952–1960.
[135] K. S. Gurumoorthy, A. Dhurandhar, and G. Cecchi. (2017). "ProtoDash: Fast interpretable prototype selection." [Online]. Available: https://arxiv.org/abs/1707.01212
[136] B. Kim, R. Khanna, and O. O. Koyejo, "Examples are not enough, learn to criticize! criticism for interpretability," in *Proc. 29th Conf. Neural Inf. Process. Syst. (NIPS)*, 2016, pp. 2280–2288.
[137] S. Wachter, B. Mittelstadt, and C. Russell. (2017). "Counterfactual explanations without opening the black box: Automated decisions and the GDPR." [Online]. Available: https://arxiv.org/abs/1711.00399
[138] X. Yuan, P. He, Q. Zhu, and X. Li. (2017). "Adversarial examples: Attacks and defenses for deep learning." [Online]. Available: https://arxiv.org/abs/1712.07107
[139] V. Schetinin *et al.*, "Confident interpretation of Bayesian decision tree ensembles for clinical applications," *IEEE Trans. Inf. Technol. Biomed.*, vol. 11, no. 3, pp. 312–319, May 2007.
[140] S. Hara and K. Hayashi. (2016). "Making tree ensembles interpretable." [Online]. Available: https://arxiv.org/abs/1606.05390
[141] H. F. Tan, G. Hooker, and M. T. Wells. (2016). "Tree space prototypes: Another look at making tree ensembles interpretable." [Online]. Available: https://arxiv.org/abs/1611.07115
[142] R. D. Gibbons *et al.*, "The CAD-MDD: A computerized adaptive diagnostic screening tool for depression," *J. Clin. Psychiatry*, vol. 74, no. 7, pp. 669–674, 2013.
[143] S. García, A. Fernández, and F. Herrera, "Enhancing the effectiveness and interpretability of decision tree and rule induction classifiers with evolutionary training set selection over imbalanced problems," *Appl. Soft Comput.*, vol. 9, no. 4, pp. 1304–1314, 2009.
[144] F. Wang and C. Rudin, "Falling rule lists," in *Proc. 18th Int. Conf. Artif. Intell. Statist. (AISTATS)*. San Diego, CA, USA: JMLR W&CP, 2015, pp. 1013–1022.
[145] G. Su, D. Wei, K. R. Varshney, and D. M. Malioutov. (2015). "Interpretable two-level Boolean rule learning for classification." [Online]. Available: https://arxiv.org/abs/1511.07361
[146] D. M. Malioutov, K. R. Varshney, A. Emad, and S. Dash, "Learning interpretable classification rules with boolean compressed sensing," in *Transparent Data Mining for Big and Small Data*. Springer, 2017, pp. 95–121.
[147] S. Mishra, B. L. Sturm, and S. Dixon, "Local interpretable model-agnostic explanations for music content analysis," in *Proc. ISMIR*, 2017, pp. 537–543.
[148] S. Mohseni and E. D. Ragan. (2018). "A human-grounded evaluation benchmark for local explanations of machine learning." [Online]. Available: https://arxiv.org/abs/1801.05075
[149] J. Huysmans, K. Dejaeger, C. Mues, J. Vanthienen, and B. Baesens, "An empirical evaluation of the comprehensibility of decision table, tree and rule based predictive models," *Decis. Support Syst.*, vol. 51, no. 1, pp. 141–154, 2011.
[150] A. Backhaus and U. Seiffert, "Quantitative measurements of model interpretability for the analysis of spectral data," in *Proc. IEEE Symp. Comput. Intell. Data Mining (CIDM)*, 2013, pp. 18–25.
[151] F. Poursabzi-Sangdeh, D. G. Goldstein, J. M. Hofman, J. W. Vaughan, and H. Wallach. (2018). "Manipulating and measuring model interpretability." [Online]. Available: https://arxiv.org/abs/1802.07810
[152] D. Bau, B. Zhou, A. Khosla, A. Oliva, and A. Torralba. (2017). "Network dissection: Quantifying interpretability of deep visual representations." [Online]. Available: https://arxiv.org/abs/1704.05796
[153] M. J. Paul, "Interpretable machine learning: Lessons from topic modeling," in *Proc. CHI Workshop Hum.-Centered Mach. Learn.*, 2016, pp. 1–16.
[154] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal. (2018). "Explaining explanations: An approach to evaluating interpretability of machine learning." [Online]. Available: https://arxiv.org/abs/1806.00069
[155] T. Miller, P. Howe, and L. Sonenberg, "Explainable AI: Beware of inmates running the asylum," in *Proc. IJCAI Workshop Explainable AI (XAI)*, 2017, pp. 36–42.
[156] M. Bauer and S. Baldes, "An ontology-based interface for machine learning," in *Proc. 10th Int. Conf. Intell. Interfaces*, 2005, pp. 314–316.
[157] J. Zhu, A. Liapis, S. Risi, R. Bidarra, and G. M. Youngblood, "Explainable AI for designers: A human-centered perspective on mixed-initiative co-creation," in *Proc. IEEE Conf. Comput. Intell. Games (CIG)*, 2018, pp. 458–465.
[158] P. Tamagnini, J. Krause, A. Dasgupta, and E. Bertini, "Interpreting black-box classifiers using instance-level visual explanations," in *Proc. 2nd Workshop Hum.-Loop Data Anal.*, 2017, Art. no. 6.
[159] F. M. Hohman, M. Kahng, R. Pienta, and D. H. Chau. (2018). "Visual analytics in deep learning: An interrogative survey for the next frontiers." [Online]. Available: https://arxiv.org/abs/1801.06889
[160] A. Bunt, M. Lount, and C. Lauzon, "Are explanations always important?: A study of deployed, low-cost intelligent interactive systems," in *Proc. ACM Int. Conf. Intell. Interfaces*, 2012, pp. 169–178.
[161] T. Kulesza, S. Stumpf, M. Burnett, S. Yang, I. Kwan, and W.-K. Wong, "Too much, too little, or just right? Ways explanations impact end users' mental models," in *Proc. IEEE Symp. Vis. Lang. Hum.-Centric Comput. (VL/HCC)*, Sep. 2013, pp. 3–10.
[162] D. Holliday, S. Wilson, and S. Stumpf, "User trust in intelligent systems: A journey over time," in *Proc. 21st Int. Conf. Intell. User Interfaces*, 2016, pp. 164–168.
[163] F. Offert. (2017). "'I know it when I see it'. Visualization and intuitive interpretability." [Online]. Available: https://arxiv.org/abs/1711.08042
[164] J. Wang, R. Fujimaki, and Y. Motohashi, "Trading interpretability for accuracy: Oblique treed sparse additive models," in *Proc. 21th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2015, pp. 1245–1254.
[165] G. Shmueli, "To explain or to predict?" *Stat. Sci.*, vol. 25, no. 3, pp. 289–310, 2010.
[166] T. Yarkoni and J. Westfall, "Choosing prediction over explanation in psychology: Lessons from machine learning," *Perspect. Psychol. Sci.*, vol. 12, no. 6, pp. 1100–1122, 2017.
[167] A. Dhurandhar, V. Iyengar, R. Luss, and K. Shanmugam. (2017). "TIP: Typifying the interpretability of procedures." [Online]. Available: https://arxiv.org/abs/1706.02952
[168] N. Puri, P. Gupta, P. Agarwal, S. Verma, and B. Krishnamurthy. (2017). "MAGIX: Model agnostic globally interpretable explanations." [Online]. Available: https://arxiv.org/abs/1706.07160
[169] K. R. Varshney, P. Khanduri, P. Sharma, S. Zhang, and P. K. Varshney. (2018). "Why interpretability in machine learning? An answer using distributed detection and data fusion theory." [Online]. Available: https://arxiv.org/abs/1806.09710
[170] M. Fox, D. Long, and D. Magazzeni, "Explainable planning," in *Proc. IJCAI Workshop XAI*, 2017, pp. 24–30.
[171] T. Chakraborti, S. Sreedharan, Y. Zhang, and S. Kambhampati, "Plan explanations as model reconciliation: Moving beyond *Explanation as Soliloquy*," in *Proc. IJCAI*, 2017, pp. 156–163.
[172] Y. Zhang, S. Sreedharan, A. Kulkarni, T. Chakraborti, H. Zhuo, and S. Kambhampati, "Plan explicability and predictability for robot task planning," in *Proc. ICRA*, May 2017, pp. 1313–1320.

[173] M. Harbers, K. van den Bosch, and J.-J. Meyer, "Design and evaluation of explainable BDI agents," in *Proc. IEEE/WIC/ACM Int. Conf. Web Intell., Intell. Agent Technol.*, Aug. 2010, pp. 125–132.
[174] M. Harbers, K. van den Bosch, and J.-J. Meyer, "Self-explaining agents in virtual training," presented EC-TEL PROLEARN Doctoral Consortium, 2009.
[175] P. Langley, B. Meadows, M. Sridharan, and D. Choi, "Explainable agency for intelligent autonomous systems," in *Proc. AAAI*, 2017, pp. 4762–4763.
[176] F. Kaptein, J. Broekens, K. Hindriks, and M. Neerincx, "The role of emotion in self-explanations by cognitive agents," in *Proc. 7th Int. Conf. Affect. Comput. Intell. Interact. Workshops Demos (ACIIW)*, Oct. 2017, pp. 88–93.
[177] M. A. Neerincx, J. van der Waa, F. Kaptein, and J. van Diggelen, "Using perceptual and cognitive explanations for enhanced human-agent team performance," in *Proc. Int. Conf. Eng. Psychol. Cogn. Ergonom. (EPCE)*, 2018, pp. 204–214.
[178] F. J. C. Garcia, D. A. Robb, X. Liu, A. Laskov, P. Patron, and H. Hastie, "Explain yourself: A natural language interface for scrutable autonomous robots," in *Proc. Explainable Robot. Syst. Workshop HRI*, 2018.
[179] E. Akyol, C. Langbort, and T. Basar. (2016). "Price of transparency in strategic machine learning." [Online]. Available: https://arxiv.org/abs/1610.08210
[180] M. Igami. (2017). "Artificial intelligence as structural estimation: Economic interpretations of deep blue, bonanza, and AlphaGo." [Online]. Available: https://arxiv.org/abs/1710.10967

**AMINA ADADI** received the Degree in computer engineering from the National School of Applied Sciences of Fez in 2012 and the Ph.D. degree in computer sciences from Sidi Mohammed Ben Abdellah University, Fez, Morocco, in 2017. Her current research interests include artificial intelligence, machine learning, and semantic Web services.

**MOHAMMED BERRADA** received the Ph.D. degree in computer sciences from the Faculty of Sciences, Sidi Mohammed Ben Abdellah University, Fez, Morocco, in 2008. He is currently a Professor of computer sciences and the Manager of the IT Department, National School of Applied Sciences of Fez. His current research interests include artificial intelligence, e-learning, enterprise architecture, and Web services.