Information Fusion 76 (2021) 89–106

Contents lists available at ScienceDirect

Information Fusion

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

Notions of explainability and evaluation approaches for explainable artificial intelligence

Giulia Vilone *, Luca Longo

School of Computer Science, College of Science and Health, Technological University Dublin, Dublin, Republic of Ireland

ARTICLE INFO

Keywords:
Explainable artificial intelligence
Notions of explainability
Evaluation methods

ABSTRACT

Explainable Artificial Intelligence (XAI) has experienced a significant growth over the last few years. This is due to the widespread application of machine learning, particularly deep learning, that has led to the development of highly accurate models that lack explainability and interpretability. A plethora of methods to tackle this problem have been proposed, developed and tested, coupled with several studies attempting to define the concept of explainability and its evaluation. This systematic review contributes to the body of knowledge by clustering all the scientific studies via a hierarchical system that classifies theories and notions related to the concept of explainability and the evaluation approaches for XAI methods. The structure of this hierarchy builds on top of an exhaustive analysis of existing taxonomies and peer-reviewed scientific material. Findings suggest that scholars have identified numerous notions and requirements that an explanation should meet in order to be easily understandable by end-users and to provide actionable information that can inform decision making. They have also suggested various approaches to assess to what degree machine-generated explanations meet these demands. Overall, these approaches can be clustered into human-centred evaluations and evaluations with more objective metrics. However, despite the vast body of knowledge developed around the concept of explainability, there is not a general consensus among scholars on how an explanation should be defined, and how its validity and reliability assessed. Eventually, this review concludes by critically discussing these gaps and limitations, and it defines future research directions with explainability as the starting component of any artificial intelligent system.

1. Introduction

The number of scientific articles, conferences and symposia around the world in eXplainable Artificial Intelligence (XAI) have significantly increased over the last decade [1,2]. This has led to the development of a plethora of domain-dependent and context-specific methods for dealing with the interpretation of machine learning (ML) models and the formation of explanations for humans. This has been closely followed by research devoted to the investigation of approaches to evaluate the quality of automatically-generated explanations. The upturn in the XAI research outputs of the last decade is prominently due to the fast increase in the popularity of ML and in particular of deep learning, with many applications in several business areas, spanning from e-commerce [3] to games [4] and including applications in criminal justice [5,6], healthcare [7], computer vision [6] and battlefield simulations [8], just to mention a few. Unfortunately, most of the models that have been built with ML and deep learning have been labelled 'black-box' by scholars because their underlying structures are complex, non-linear and extremely difficult to be interpreted and explained to laypeople. This opacity has created the need for XAI architectures that is motivated mainly by three reasons, as suggested by [8,9]: (i) the demand to produce more transparent models; (ii) the need of techniques that enable humans to interact with them; (iii) the requirement of trustworthiness of their inferences. Additionally, as many scholars have rightly pointed out [9–12], models induced from data must be liable as liability will likely soon become a legal requirement. Article 22 of the General Data Protection Regulation (GDPR) sets out the rights and obligations of the use of automated decision making. Noticeably, it introduces the *right of explanation* by giving individuals the right to obtain an explanation of the inference(s) automatically produced by a model, confront and challenge an associated recommendation, particularly when it might negatively affect an individual legally, financially, mentally or physically. By approving this GDPR article, the European Parliament attempted to tackle the problem related to the propagation of potentially biased inferences to society, that a computational model might have learnt from biased and unbalanced data.

* Corresponding author.
E-mail address: D18126441@mytudublin.ie (G. Vilone).

https://doi.org/10.1016/j.inffus.2021.05.009
Received 1 June 2020; Received in revised form 29 April 2021; Accepted 15 May 2021
Available online 25 May 2021

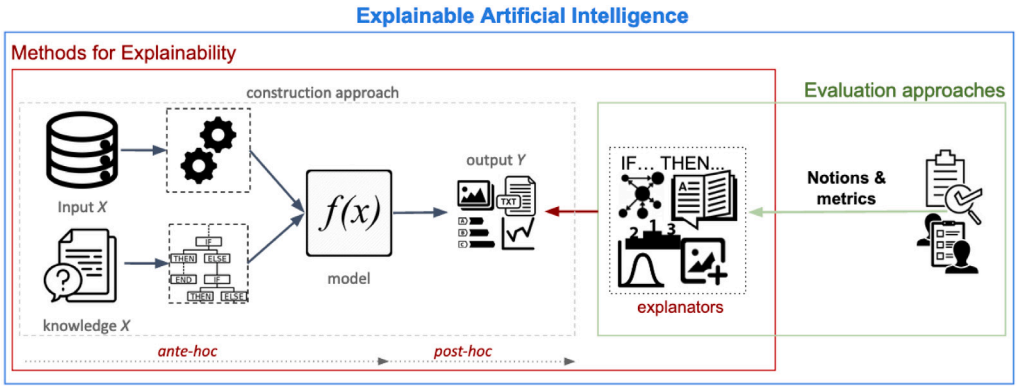
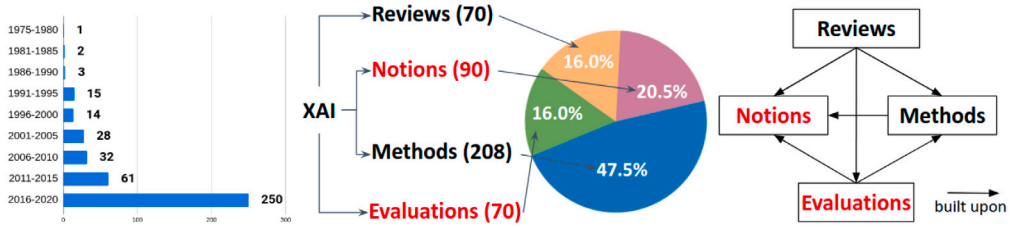


Fig. 1. Diagrammatic view of Explainable Artificial Intelligence with interaction between methods for explanations and their evaluation approaches.



(a) (b) (c)

Fig. 2. Proposed classification of the XAI literature with (a) the distribution of published scientific articles over time, (b) the root of our hierarchical classification system representing the main four categories and the percentage of articles in each , and (c) the salient relations between these categories that have emerged.

Many authors surveyed scientific articles surrounding explainability within Artificial Intelligence (AI) in specific sub-domains, motivating the need for literature organisation. For instance, [13,14] respectively reviewed the methods for explanations with neural and bayesian networks while [15] clustered the scientific contributions devoted to extracting rules from Support Vector Machines (SVMs). The goal was, and in general is, to create rules highly interpretable by humans while maintaining a degree of accuracy offered by trained models. Only a few scholars attempted to make a more comprehensive survey and organisation of the methods for explainability as a whole [1,16,17]. Other scholars have identified a wide set of notions and requirements that an explanation should meet in order to be easily understandable by end-users [18], in particular laypeople, and in order to provide effective actionable information to support decision making processes. However, a thorough analysis of these studies have highlighted the lack of a systematic organisation of the various notions that are related to the concept of explainability, and those approaches devoted to the evaluation of the quality of the explanations. Therefore, this paper aims at filling this gap and it aims at systematically review research studies in the field of XAI by focusing on a subset of scientific peer-reviewed articles that tackled the problem of explainability, from a conceptual and theoretical point of view, and that proposed approaches on how to evaluate XAI methods. The conceptual framework at the basis of the proposed goal is represented in Fig. 1 whereby methods for explainability are built from automatically induced models using explanators, and these can be evaluated by employing notions and metrics.

The remainder of this paper is organised as it follows. Section 2 provides a detailed description of the research methods employed for searching for relevant research articles and a general classification structure of XAI while Sections 3–4 specifically focus on its branches related to notions related to explainability and XAI evaluation approaches. Eventually, Section 5 concludes this systematic review by trying to define the boundaries of the discipline of XAI, as well as suggesting future research work and challenges.

2. Research methods

Organising the literature of explainability within AI in a precise and indisputable way as well as setting clear boundaries is far from being an easy task. This is due to the multidisciplinarity surroundings this new fascinating field of research spanning from computer science to mathematics, from psychology to human factors, from philosophy to ethics. The development of computational models from data belongs mainly to computer science, statistics and mathematics, whereas the study of explainability belongs more to human factors and psychology since humans are involved. Reasoning over the notion of explainability touches ethics and philosophy. Therefore, some constraints had to be defined, and the following article types were excluded:

- scientific studies discussing the notion of explainability in different contexts than AI and computer science, such as philosophy or psychology;
- articles or technical reports that have not gone through a peer-review process.

Taking into account the above constraints, this systematic review was carried out in three phases:

1. articles discussing explainability were searched by using Google Scholar and the following terms: *'explainable artificial intelligence'*, *'explainable machine learning'*, *'interpretable machine learning'*. The queries returned several thousands of results, but it became immediately clear that only the first ten pages, with ten results each, for each query, contained relevant articles. Altogether, these searches provided a basis of almost three hundred peer-reviewed articles for initial screening;
2. the bibliographic section of each of the identified articles was subsequently checked thoroughly. This led to the selection of one hundred further articles whose bibliographic section was recursively analysed. This process was iterated until it converged and no more new articles were found.

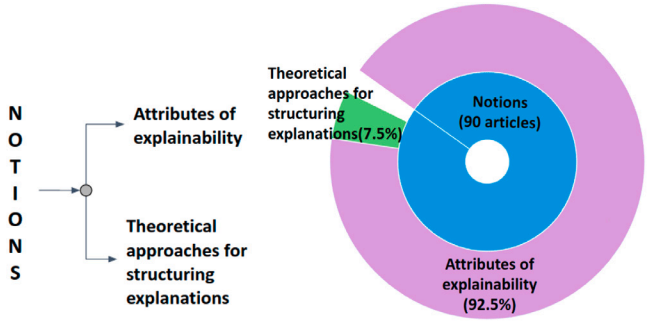


Fig. 3. Classification of the notions related to the concept of explainability (left) and distribution of the relative scientific studies across categories (right).

3. overall, 406 articles from 1975 to 2020 have been found (Fig. 2, a) and organised in four main categories (Fig. 2, b). Note that some of the articles can belong to multiple categories. These were further systematically reviewed and the goal was to identify only those matching the following criteria:

    - `notions of explainability` - it includes studies focused on the definition of the notions related to the concept of explainability, its attributes and main characteristics, as well as the requirements of an effective explanation;
    - `evaluation of methods for explainability` - it includes articles aimed at evaluating methods for explainability.

Fig. 2, part c, depicts the dependencies of the main four categories. In general, scholars would not be able to carry out reviews of the XAI literature without the existence and consideration of relevant notions and methods for explainability as well as the approaches for evaluating the performances of these methods. Evaluation approaches naturally followed the creation of methods for explainability which have been engineered to meet as many requirements of an effective explanation as possible.

3. Notions related to the concept of explainability

Explaining a model induced from data by employing a specific learning technique is not a trivial goal. A body of literature focused on achieving such a goal by investigating and attempting to define the concept of *explainability*, leading to many types of explanation and the formation of several attributes and structures. To organise these, the specific following clusters are proposed (see also Fig. 3):

- `attributes of explainability` - it contains criteria and characteristics used by scholars to try to define the construct of 'explainability';
- `theoretical approaches for structuring explanations` - it includes the different ways scholars reported explanations for their ad-hoc applications, what pieces of information are included or left out, and the various components an explanation can be constructed upon, such as causes, context, and consequences of a model's prediction as well as their ordering.

3.1. Attributes of explainability

One of the principal reasons to produce an explanation is to gain the trust of users [19]. Trust is the main way to increase users' confidence with a system [20] and to make them feel comfortable while controlling and using it [21]. Besides trust, researchers determined other positive effects brought by explainability. According to [22], it is part of human nature to assign causal attribution of events. A system that provides a causal explanation on its inferential process is perceived more human-like by end-users as a consequence of the innate tendency of human psychology to anthropomorphism. Thus, several scholars spoke at length about causality which is considered a fundamental attribute of explainability [8,21,23–25]. Explanations must make the causal relationships between the inputs and the model's predictions explicit, especially when these relationships are not evident to end-users. Data-driven models are designed to discover and exploit associations in the data, but they cannot guarantee that there is a causal relationship in these associations. As pointed out in [21], the task of inferring causal relationships strongly depends on prior knowledge, but some associations might be completely unexpected and not explainable yet. Scientists can use these associations to generate hypotheses to be tested in scientific experiments; however, this is outside the scope of the methods for explainability. Other four reasons supporting the necessity to explain the logic of an inferential system or a learning algorithm were suggested in [1]:

- *explain to justify* - the decisions made by utilising an underlying model should be explained in order to increase their justifiability;
- *explain to control* - explanations should enhance the transparency of a model and its functioning, allowing its debugging and the identification of potential flaws;
- *explain to improve* - explanations should help scholars improve the accuracy and efficiency of their models;
- *explain to discover* - explanations should support the extraction of novel knowledge and the learning of relationships and patterns.

Despite the widely recognised importance of explainability, researchers are striving to determine universal, objective criteria on how to build and validate explanations [18]. Numerous notions underlying the effectiveness of explanations were proposed in the literature (as summarised in Table 1). [18] surveyed 250 articles from the fields of philosophy, psychology and cognitive science to analyse in depth how people define, generate, select, evaluate and present explanations. The author also presented an interesting definition of XAI as a human–agent interaction problem where the agent reveals the underlying causes to its or another agent's decision process. In other words, XAI is believed to be a subset of the human–agent interaction field that can be defined as the intersection of AI, social science and HCI.

Two studies on *explainability* demonstrated that this concept is utilised in several fields, spanning from mathematics, physics, computer science to engineering, psychology, medicine and social sciences [32, 46]. Explainability is often replaced with the notion of *interpretability*, considered as synonyms within the general AI community, and in particular by those scholars in automated learning and reasoning, whereas it seems that the software engineering community prefers the term *understandability* [32]. Generally speaking, interpretability is often defined as the capacity to provide or bring out the meaning of an abstract concept and understandability as the capacity to make the model understandable by end-users (see Table 1). However, other definitions are proposed in the literature. Explainability or interpretability is defined in [26] as "the degree to which a human observer can understand the reason behind a decision (or a prediction) made by the model". An interesting distinction between the concepts of *interpretation* and *explanation* was proposed in [57]. On one hand, an interpretation is the mapping of an abstract concept (as a predicted class) into a domain that the human can make sense of, such as, for instance, images or texts that can be inspected and classified by people. On the other hand, an explanation is the collection of features of an interpretable domain that contributed to produce a prediction for a given item. The authors of [57] did not specify how to determine this collection of features. The selection criteria are to be decided by researchers according to several factors like the type of input data and the degree of refinement in the explanation demanded by end-users. An expansion of the definition of interpretability through the determination of its main characteristics was presented in [18,45,60]. In detail, [45] suggested

Table 1
Definition of the notions related to the concept of explainability.

Notion	Description & Reference
Algorithmic transparency	The degree of confidence of a learning algorithm to behave 'sensibly' in general [2,26]
Actionability	The capacity of a learning algorithm to transfer new knowledge to end-users [27,28]
Causality	The capacity of a method for explainability to clarify the relationship between input and output [8,21–25,29]
Completeness	The extent to which an underlying inferential system is described by explanations [27,28,30]
Comprehensibility	The quality of the language used by a method for explainability [9,31–38]
Cognitive relief	The degree to which an explanation decreases the "surprise value" which measures the amount of cognitive dissonance between the explanandum and the user's beliefs. The explanandum is something unexpected by the user that creates dissonance with his/her beliefs [23,39]
Correctability	The capacity of a method for explainability to allow end-users make technical adjustments to an underlying model [27,28]
Effectiveness	The capacity of a method for explainability to support good user decision-making [40–43]
Efficiency	The capacity of a method for explainability to support faster user decision-making [20,41,42]
Explicability	The degree of association between the expected behaviour of a robot to achieve assigned tasks or goals and its actual observed actions [44]
Explicitness	The capacity of a method to provide immediate and understandable explanations [45]
Faithfulness	The capacity of a method for explainability to select truly relevant features [45]
Intelligibility	The capacity to be apprehended by the intellect alone [46–50]
Interactivity	The capacity of an explanation system to reason about previous utterances both to interpret and answer users' follow-up questions [51,52]
Interestingness	The capacity of a method for explainability to facilitate the discovery of novel knowledge and to engage user's attention [33,34,36,53,54]
Interpretability	The capacity to provide or bring out the meaning of an abstract concept [9,18,33,35,55–63]
Informativeness	The capacity of a method for explainability to provide useful information to end-users [21]
Justifiability	The capacity of an expert to assess if a model is in line with the domain knowledge [1,33,40,55,64,65]
Mental Fit	The ability for a human to grasp and evaluate a model [33,66]
Monotonicity	The relationship between a numerical predictor and the predicted class that occurs when increasing the value of the predictor leads to either always increase or decrease the probability of an instance's membership to the class [67]
Persuasiveness	The capacity of a method for explainability to convince users perform certain actions [20,41,42]
Predictability	The capacity to anticipate the sequence of consecutive actions in a plan [44]
Refinement	The capacity of a method to guide experts in improving the model's performance/robustness [68]
Reversibility	The capacity to allow end-users to bring a ML-based system to an original state after it has been exposed to an harmful action that makes its predictions worse [27,28]
Robustness	The persistence of a method for explainability to withstand small perturbations of the input that do not change the prediction of the model [68,69]
Satisfaction	The capacity of a method for explainability to increase the ease of use and usefulness of a ML-based system [20,41,42]
Scrutability/ diagnosis	The capacity of a method for explainability to inspect a training process that fails to converge or does not achieve an acceptable performance [20,41,68]
Security	The reliability of a model to perform to a safe standard across all reasonable contexts [70,71]
Selection/ simplicity	The ability of a method for explainability to select only the causes that are necessary and sufficient to explain the prediction of an underlying model [25]
Sensitivity	The capacity of a method for explainability to reflect the sensitivity of the underlying model with respect to variations in the input feature space [72,73]
Simplification	The capacity to reduce the number of the considered variables to a set of principal ones [74,75]
Soundness	The extent to which each component of an explanation's content is truthful in describing an underlying system [27,28]
Stability	The consistency of a method to provide similar explanations for similar/neighbouring inputs [45]
Transparency	The capacity of a method to explain how the system works even when it behaves unexpectedly [9–12,20,26,40,41,47,58,59,63,64,76–78]
Transferability	The capacity of a method for explainability to transfer prior knowledge to unfamiliar situations [21]
Understandability	The capacity of a method for explainability to make a model understandable [32,33,46,68,79]

the following requirements: (I) *fidelity* - the representation of inputs and models in terms of concepts should preserve and present to end-users their relevant features and structures, (II) *diversity* - inputs and models should be representable with few non-overlapping concepts, and (III) *grounding* - concepts should have an immediate human-understandable interpretation. These requirements were further expanded in [18] by listing a set of characteristics that an explanation should possess:

- *contrastive nature of explanations* - people seek for an explanation when they are presented with counterfactual and/or counter-intuitive events;
- *selectivity of explanations* - people usually do not expect that an explanation contains the actual and complete list of the causes of an event, but only a selection of the few causes deemed to be necessary and sufficient to explain it. Authors point out the risk that this selection might be influenced by cognitive biases;

- *social nature of explanations* - explanations are part of a dialogue aiming at transferring knowledge, therefore, they are based on the beliefs of both the explainer and explainee;
- *irrelevance of probabilities to explanations* - referring to the occurrence probabilities of events or to the statistical relationships between causes and events does not produce a satisfactory and intuitive explanation. Explanations are more effective when they refer to the causes and not to their likelihood.

Four further requirements for enhancing the interpretability of visual explanations were added in [60]: (i) *graphical integrity* - the representations should highlight the features that contribute the most to the final predictions and distinguish those with positive and negative attribution, (ii) *coverage* - a large fraction of the most important features should be visible in the representation, (iii) *morphological clarity* - the important features should be clearly displayed, their visualisation cannot be 'noisy', and (iv) *layer separation* - the representation cannot occlude the raw image which should be visible for human inspection. Other two notions strongly correlated with interpretability are *comprehensibility* [33] and *intelligibility* [46]. However, scholars highlighted some differences. [35] proposed to distinguish between *interpretable systems*, systems in which end-users can mathematically analyse algorithms, and *comprehensible systems* that "emit symbols enabling user-driven explanations of how a conclusion is reached". Two studies [46,50] defined intelligibility as an attribute of user-centric reasoned explanations that are easily interpretable by end-users and that draws from foundational concepts of other disciplines such as philosophy and cognitive psychology. Additionally, both studies recommended exploiting the experience and knowledge of the HCI community in making interfaces that empower people to assure that intelligibility will be one of the core requirements of the next generation of AI systems. Other authors focused on breaking some of the notions identified in Table 1 into sub-notions or on assigning further requirements. For example, three sub-notions related to *transparency* that should be achieved by any learning model were defined in [21,26]:

- *simulatability* - the capacity of a model to allow a user to understand its structure and functioning entirely;
- *decomposability* - the degree to which a model can be decomposed into its individual components (input, parameters and output) and of their intuitive explainability;
- *algorithmic transparency* - the degree of confidence of a learning algorithm to behave 'sensibly' in general (see also Table 1).

However, according to [21], it is not possible to achieve algorithmic transparency in neural networks because of the current incapacity of experts to understand the inferential process of these models and to prove that they work correctly on new, unseen observations. Scholars attempted to overcome this shortcoming by finding methods to trace the predictions of a model to the most influential features of the input. Examples of these methods are heat-maps [80] which are created by back-propagating the predictions of a model to the input space and highlighting relevant pixels. Alternatively, [81] proposed a solution to satisfy the simulatability and decomposability properties by substituting black-box models with Generalised Additive Models (GAMs). GAMs are linear combinations of simple models trained on a single feature of an input dataset, thus allowing end-users to quantify the contribution of each feature to the outcome. However, transparency must be handled with caution because it can be dangerous under certain circumstances, as highlighted in [78]. Requiring that data and models are fully visible to end-users prevents the creation of intellectual properties; this can significantly slow down the development of new technologies. Moreover, data can contain sensitive or personal information which cannot be made public without affecting people's privacy. Finally, the displaying of more information might push a researcher to optimise a model on specific instance(s) but deteriorating its overall performance and degree of generalisability. Scholars extensively investigated *sensitivity* [72,73]. In this context, sensitivity is considered as the sensibility of explanations to variations in the input features, model implementation and, subsequently, in the model's predictions. [72] introduced the requirement of *input invariance* meaning that a method for explainability must mirror the sensitivity of the underlying model with respect to transformations of the inputs in order to ensure a reliable interpretation of their contribution to each prediction. [73] focused on the sensitivity of methods for explainability specifically designed for neural networks, in particular those that quantify the contribution of input features to the predictions, such as DeepLift [82] and Layer-wise Relevance Propagation (LRP) [83]. In this case, a method for explainability satisfies the sensitivity requirement if it assigns a non-zero contribution to an input feature when two instances, in the input space, differ in that feature only but lead to different predictions. According to [73], methods for explainability must also fulfil the requirement of *implementation invariance*. This suggests that a method applied to functionally equivalent neural networks should assign identical contributions to the features of the input. Two neural networks are *functionally equivalent* if their predictions are equal for all inputs despite having different implementations and architectures. Finally, scholars identified various factors that might affect the *interestingness* of a model, in particular of the rule-based ones [36,53]. First, *rule size* is the number of instances satisfied by a rule. According to the authors, small size rules are undesirable as they explain only a few instances [53]. The main aim is to discover rules that cover a large portion of the input data. However, there are situations where small rules might capture exception occurring in the data that can be of interest for scientists. Second, *imbalance of class distributions* occurs when the instances belonging to a class are more frequent than those of another class. It might be more difficult, hence more interesting, to discover those rules aimed at predicting the minority classes. *Attribute costs* represent the cost to get access to the actual value of an attribute of the data. For example, it is easy to assess the gender of a patient but the determination of some health-related attributes can require an expensive investigation. Rules that utilise only 'cheap' attributes are more interesting. Eventually, the interestingness of a rule must take into account the *misclassification costs*. In some domain of application, the erroneous classification of an instance might have a significant impact, not only in terms of money. In case of medical diagnosis, classifying as healthy a patient affected by a lethal disease might lead to premature death. Interestingness was also examined for Reinforcement Learning (RL) agents which are designed to take actions in a specific environment with the aim to maximise a cumulative reward [54]. The authors proposed a framework to make the behaviour of these agents explainable by analysing their historical interactions with the environment and extracting a few *interestingness elements*, such as the portion of environment observed by the agent, the frequency of certain types of interactions and the cost (in terms of a reward) of the interactions carried out.

3.2. Theoretical approaches for structuring explanations

Researchers tried to create a classification system for the types of explanation suitable for interpreting the logic of learning algorithms. A method for explainability should answer several questions to form an exhaustive explanation. The two most common questions are *why* and *how* the model under scrutiny produces its predictions/inferences [2,3,84,85]. However, scholars identified other questions that might arise and that require different answers, thus different types of explanations [86]. Additionally, as pointed out in [87,88], distinct behaviours, distinct problems and distinct types of users require distinct explanations, as shown in a diagrammatic way in Fig. 4. This has led to many ad-hoc classifications that are domain-dependent and are hard to be merged into one.

Two studies [89,90] focused on the types of users of methods for explainability. [89] identified four types of explanations, ordered by the degree of completeness demanded by the different users. *Developer explanations* are for researchers and developers who must understand

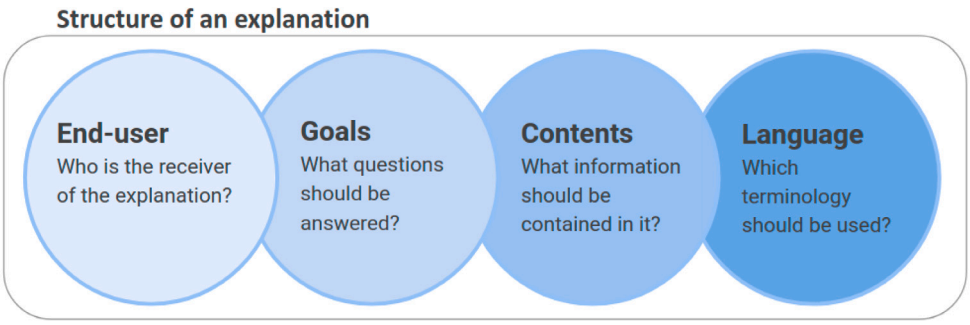


Fig. 4. Diagram of the main factors shaping the structure of a machine-generated explanation.

how an entire AI system works to verify or improve it. They require complete explanations that are generally harder to interpret. *Assurance explanations* are for users, other than developers and researchers, who need evidence demonstrating that a system meets a set of requirements assuring that it is fit for purpose. The last two types, *end-user explanations* and *external explanations*, are for those who are somehow affected by the system's operations and they need to know the rationale behind its behaviour, usually in real-time, without being familiar with all its components. The main difference between these explanations is that in the former, end-users are involved in operative situations while, in the latter, users are being affected by the system's behaviour but are not strictly involved in its operations. [90] proposed a two-class system consisting of *traced-based explanations*, useful for system designers, that accurately reflects the reasoning implemented within a model, and *reconstructive explanations*, designed for end-users, based on an active, problem-solving approach. A reconstructive explanation tends to build a 'story' exposing the input features contributing to a prediction. For instance, an image of a bird was assigned to a certain class because of the colour of the bird. However, the model might have analysed other features that did not influence the final assessment, like the image's background. These characteristics can be included in the traced-based explanations but excluded from the reconstructive explanations. The same scholars also developed Reconstructive EXplanation (REX) [90, 91], an explanatory tool capable of producing reconstructive textual explanations for expert systems. These explanations consist of mapping over key elements from the execution trace of an expert system and expanding on them using the more structured textbook knowledge, which is a collection of relationships between cues, hypotheses and goals as illustrated by this example: "The presence of damages to the drainage pipes is a sign that the cause of an excessive high uplift pressures on a concrete dam is internal erosion of soil under the dam. Erosion would lead to broken pipes, therefore slowing drainage and causing high uplift pressures". The goal is to determine the cause of high uplift pressure on a concrete dam, the cues consist of the presence of broken pipes and the hypothesis is the erosion of soil. Another classification of the types of explanations was proposed in [92] for intelligent systems which include intelligent agents, such as those AI assistants utilised in customer support chats, or other support decision systems like those for medical diagnoses. Here, traced-based explanations were defined as *mechanistic explanations* and correspond to the answer of the question "How does it work?". Hence, they must offer insights into the causes and consequences of events and how these events and the different components of the intelligent systems interact to give rise to complex actions. Reconstructive explanations were instead called *ontological explanations* and describe the structural properties of the intelligent systems: its components, their attributes, and how they are related to each other. [92] also added a third category, referred to as *operational explanations* which respond to the question "How do I use it?" by relating goals to the mechanics designed to realise them. A more articulated classification of the types of explanations was introduced in [93] and includes five types of explanations that intelligent systems should produce. The first one, *teaching explanations*, aims at informing humans about the concepts learned by the system such as, for example, the presence of some physical constraints (walls or other obstacles) that can limit its actions. *Introspective tracing explanations* have the goal of finding the cause of and the solution to a fault whilst *introspective informative explanations* aim at explaining predictions based on the system's reasoning process to improve human–system interaction. The last two types of explanations, *post-hoc explanations* and *execution explanations*, are respectively focused on explaining the decisions and their execution without necessarily following the same reasoning process and directly linking them with the inputs. An example of post-hoc and execution explanation is a robot describing the path it wants to follow to go from point A to point B and all the movements it must do to cover that path. This explanation can mention the characteristics of the surrounding environment that have been considered while planning the path, but it does not mention that alternative paths were considered and discarded and the reasons beyond these decisions. Finally, [94] presented a classification of the types of knowledge intrinsically embedded in an explanation. Explanations based on *reasoning domain knowledge* focus on the domain knowledge needed to perform reasoning, including rules and terminology. *Communication domain knowledge* is instead about the domain knowledge needed to inform, clearly and comprehensively, end-users about the underlying domain, and it might include additional information not strictly necessary for reasoning. Eventually, *domain communication knowledge* focuses on how to communicate within a certain domain of application and it deals with practical aspects of the communication process, such as the language to be used, the most effective strategies for effective explanations and the communication medium. This knowledge must be tuned to the prior knowledge and cognitive state of the hearer.

The most effective way to structure explanations is still an open problem despite being tackled by several scholars. As highlighted in [95], two properties of the structure of an explanation can have a significant effect on learning, namely the capacity to "accommodate novel information in the context of prior beliefs and do so in a way that fosters generalization". As prior beliefs greatly vary according to the application field and the domain knowledge of end-users, researchers examined and proposed different structures for explanations which are domain-dependent. The first studies on the most suitable and effective structures of textual explanations were carried out in the 80–90s and focused on interpreting the inferential process of expert models. Most of these explanations were planned as dialogues where end-users were allowed to ask a (limited) number of questions via an explanatory tool. Blah [96], an example of these tools, was primarily concerned with structuring explanations so that they do not appear too complex. It was based on a series of psycho-linguistic studies that analysed how human beings explain decisions, choices, and plans to one another. Different ways to structure a conversational explanation, or dialogue, to successfully transfer knowledge from an explainer to an explainee were listed in [52,97–100]. All these studies proposed to split a dialogue into three stages: opening, explanation and closing stage. Each stage has to obey a set of rules to ensure that the knowledge about the

model's inferential process can be successfully transferred to end-users. On one hand, [52] grounded this three-stage formal protocol on the data collected from almost four hundred real dialogues which were examined to detect the key components of an explanation, the relationships between them and their order of occurrence. These main components can be synthesised by a set of questions (mainly how, why and what) and the relative arguments presented by an explainer to an explainee who, respectively, answer the questions and acknowledge the explanation or challenge it with counterfactual examples. On the other hand, [98–100] focused on the most effective set of rules to manage interactive dialogues with interruptions from the user while maintaining coherence between the different sections of an explanation. They also developed a tool, called EDGE, that generates dialogues based on these rules. EDGE updates assumptions about the user's knowledge based on his/her questions and uses this information to influence the further planning of the explanation. Other studies on interactive dialogues [101–105] focused on their structure, language and main components (what pieces of information must be included). Based in these early studies, [106–109] proposed a modular architecture for explaining the behaviour of simulated entities in military simulations. It consists of three modules: a reasoner, a natural language generator and a dialogue manager. The user can stop simulation and query about what happened at the current time point by selecting questions from a list. The dialogue manager orchestrates the system's response: firstly, by using the reasoner to retrieve the relevant information from a relational database, then producing English responses using the natural language generator. More recently, interactive dialogues were used as the explanation format of choice in knowledge-based systems other than expert systems. AutoTutor [110], designed to be integrated into tutoring systems, is grounded on learning theories and tutoring research. It simulates a human tutor by holding a conversation with the learner in natural language. The explanations of task planning systems, according to [8,111], must contain information on (I) why a planner choose an action, (II) why a planner did not choose another action, (III) why the decisions of a planner are the best among a set of possible alternatives, (IV) why certain actions cannot be executed and (V) why one needs or does not need to change the original plan. The criterion of *episodic memory* was added to the above list by [111], whereby an agent should remember all the factors that influenced the generation and execution of a plan such as "states, actions, and values considered during plan generation, traces of plan execution in the environment, and anomalous events that led to plan revision". A formal framework to generate *preferred explanations* of a plan was introduced in [112]. Preferences over explanations must be contextualised with respect to complex observational patterns. Actions might be affected by several causes and requires reflecting on the past, meaning that explanations must take into consideration previous events and information.

4. Evaluation of methods for explainability

The development of many methods for explainability have led scholars to focus also on their evaluation. Different evaluation metrics were proposed and found in the literature as well as different types of evaluation were conducted, as summarised in Fig. 5. These metrics are not limited to the evaluation of explainability, but they also consider other aspects of explanations, such as accessibility and inclusion of content [113,114].

A thorough review of these studies led to the identification of two main ways to evaluate methods for explainability:

- `objective evaluations` - it includes research studies that employed objective metrics and automated approaches to evaluate methods for explainability;
- `human-centred evaluations` - it contains those studies that evaluated methods for explainability with a human-in-the-loop approach by involving end-users and exploited their feedback or informed judgement.

The same dual categorisation system was suggested in [33], but they named the two classes *heuristic-based* and *user-based* metrics. The former includes quantitative measures which consist of mathematical entities such as, for example, the size of models [26,67,115–117]. This is a simple explainability indicator and it is based upon the assumption that the bigger the size of a model, the harder it becomes for the users to understand it. However, this assumption was proved false. One of the outcomes of the human-centred evaluation study conducted in [67] was that users found some larger models to be more comprehensible than some smaller models because larger models provided more classification-relevant information and users are unlikely to accept weaker, simpler models when the underlying modelled concept is believed to be complex. An alternative categorisation was presented in [2] with three classes: *application-grounded*, *functionally-grounded* and *human-grounded* evaluation metrics where functionally-grounded and human-grounded metrics respectively corresponds to the heuristic-based and user-based metrics proposed in [33]. Application-grounded metrics assess the quality of machine-produced explanations of data-driven models by comparing the increase in productivity of a few users of these models when following these explanations instead of those produced by human engineers, as done in [2,26]. Because they involve humans, they can be merged into the human-grounded ones.

4.1. Objective evaluations

Scholars proposed several metrics to evaluate, formally and objectively, the methods for explainability, listed in Table 1. In the scientific literature, there is consensus that simpler learning techniques, such as linear regression and decision trees, can lead to more transparent inferences than more complex techniques, such as neural network, as they are intrinsically self-interpretable [26]. However, these simpler techniques usually do not lead to the construction of models with the same level of accuracy than those induced by more complex learning techniques. The interpretability of these models depends on many factors such as the learning algorithm, the learning architecture and its configuration (hyper-parameters).

A few sale performance indicators were utilised as a formal metric to assess the increase in productivity of the sales department when two methods of explainability, Explain and Ime [118,119], were applied to a complex real-world business problem of business-to-business sales forecasting [120,121]. The system was tested for a long period in a real-world company and the sale performance indicators were monitored. The indicators showed that the forecasts based on the Explain and Ime explanations outperformed initial sales forecasts, which supported the hypothesis that data-driven explanations better facilitate unbiased decision-making than the mental models of sale experts based upon their previous experience. Two quantitative evaluation metrics to assess the interpretability of methods generating visual explanations of a neural network trained to classify images were presented in [122]. The first metric, *filter interpretability*, considers six types of semantics for CNN filters that must be annotated by humans on testing images at the pixel level: objects, parts, scenes, textures, materials, and colours. The metric measures the intersection areas between these annotations and the distribution of the activation values of a filter over a heat-map. If there are overlapping areas, it can be said that the filter represents these semantic concepts. The second metric, *location instability*, checks if a CNN locates the relevant parts of the same object, shown in different images, at an almost constant relative distance, as the distances between the parts of an object must be almost invariant. Two alternative metrics were proposed in [123], namely *infidelity*, defined as the expected difference between the dot product of a significant input perturbation to the explanation and the subsequent output perturbation, and *sensitivity* which measures the degree to which a visual explanation is affected by insignificant perturbations in the input instances. [124] proposed to use three automated quantitative metrics, designed to assess the

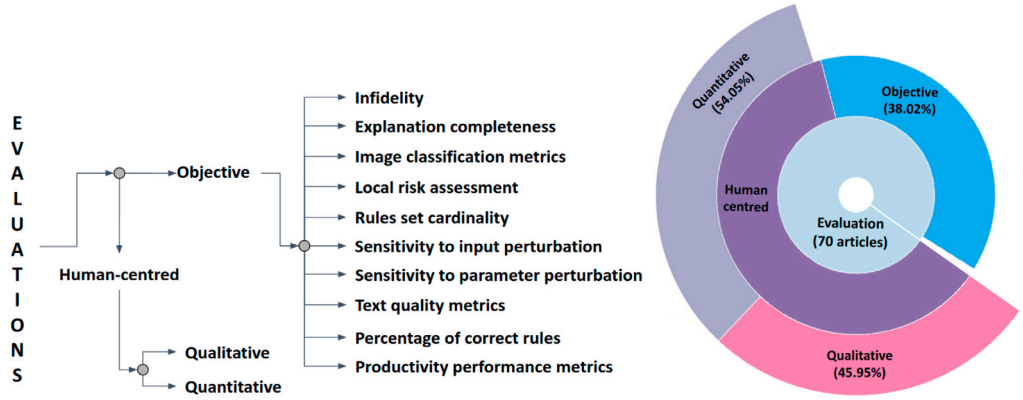


Fig. 5. Classification of the approaches to evaluate methods for explainability (left) and distribution of the relative scientific studies across categories (right).

quality of text documents, to evaluate textual explanations automatically generated by methods for explainability: *BiLingual Evaluation Understudy (BLEU)* that assesses the similarity of sentences based on the average percentage of n-gram matches, *Automatic NT Translation Metric (METEOR)* that evaluates semantically the similarity between words of sentences by using pre-trained word embeddings and *Consensus-based Image Description Evaluation (CIDEr)* that compares sentences generated by neural networks to reference explanations written by humans by counting 'Term Frequency–Inverse Document Frequency' weighted n-grams. The quality of rule-based explanations was assessed in [125] by checking on the entire input instance space the percentage of invalid rules, which are reported as *incorrect*. Those that proved to be valid were instead split between *redundant* rules, meaning that they could be discarded without compromising the prediction accuracy of the rule-set, and rules that proved to be *correct* and minimal. A general evaluation metric for post-hoc methods for explainability was presented in [126] based on the risk of generating unjustified 'counterfactual examples' which are instances that do not depend on previous knowledge but are artefacts of the classifier. This might happen when a model must predict an area not covered by the training set. The algorithm that generates these examples applies the minimal perturbation that changes the predicted classes of observation in such a way that it is still connected to the input data and avoid the construction of examples representing situations that are neither feasible nor logical. The explanations of the predictions made by an underlying model for these observations would not make sense and would not help the understanding of the model's logic.

A number of researchers carried out formal comparisons, based on heuristic-based metrics, between different methods for explainability to evaluate their strengths and weaknesses. The methodologies utilised for the comparisons are (see also Table 2):

- `sensitivity to input perturbation` - some features of the input dataset are removed, masked or altered and the explanations generated by a method for explainability from the model trained on both the original and modified inputs are compared;
- `sensitivity to model parameter randomisation` - the outputs a method for explainability generated from a trained model and another model of the same architecture with some or all parameters replaced by random values are compared;
- `explanation completeness` - these approaches check which method generates explanations that describe the inferential process of the underlying model to the highest extent. This consists of capturing the highest number of features of the input that affect the decision process of the model.

The vast majority of these scientific articles compared methods for explainability designed to generate visual explanations of the logic followed by neural networks for the classification of either images or texts. All these methods produce maps, like heat-maps or feature maps, and the comparison is carried out by measuring the differences in these maps generated before and after the input or the model's parameters were perturbed. The complete list of the methods that were compared, along with the type of input data that were analysed in these comparisons, is shown in Table 3. [127,128] compared the saliency maps generated by various methods for visual explainability to either a randomly initialised untrained network or from a copy of the dataset in which the labels were randomly permutated. The degree of correlation between the saliency maps was measured by calculating Spearman Rank Correlation coefficients. Similarly, [129] proposed to vary input images by occluding with zero-valued pixels their portions sharing the same relevance level, according to the saliency maps generated by four gradient-based attribution methods. The sensitivity of the four methods to this input perturbation was assessed with a formal metric, *Sensitivity-n*, which quantifies the variation in the output caused when features sharing the same relevance level are removed. The results of this analysis showed that Occlusion Sensitivity is the method that identifies the few most important features, in respect with the other methods, because it suffers the faster variations in the output when the most relevant pixels are removed. Layer-wise Relevance Propagation (LRP) and Sensitivity Analysis were tested in [130,134,135,138] by removing important words from input documents in text classification problems [130,138] or replacing the most relevant pixels (in case of images) by randomly sampling new pixel values from a uniform distribution [134,135]. The metric used in this study assesses the differences in the model's classification accuracy between the original and the perturbed input points when fed into the model. Both studies showed that LRP qualitatively and quantitatively provides a better explanation of what made a DNN reach a specific classification prediction. [131] used the same evaluation approach of [134,135] to compare LRP with Occlusion Sensitivity and Sensitivity Analysis. LRP and Occlusion Sensitivity performed better than Sensitivity Analysis, seconding the findings of [134,135]. Input perturbation was also used in [69,72,132] to test the robustness of several methods that generate visual explanations of the inferential process of DNNs applied to image classification problems.

Robustness concerns variations of a prediction's explanation provided by the method for explainability with respect to changes in the input leading to that prediction. Intuitively, if the input being explained is modified slightly–subtly enough to not change the prediction of the model then the explanation must not change much either [69]. On the one hand, [69] applied a Gaussian noise to input images and measured the relative changes in the output with respect to these perturbations with the Local Lipschitz Continuity metric. On the other hand, [72,132] added/subtracted a constant shift to the input images. Then, [132] used two metrics to assess the similarity between the heat-maps generated from the original and the perturbed images: Spearman Rank Correlation coefficients and Top-κ intersection which measure the

Table 2
Classification of the scientific articles proposing comparative approaches to evaluate methods for explainability, classified according to the methodology followed to carry out the comparison task.

Comparison approach	Reference
Sensitivity to input perturbation	[69,72,127–135]
Sensitivity to model parameter randomisation	[127,128,136]
Explanation completeness	[137]

Table 3
List of the methods for explainability evaluated by comparing their output explanations, which are only saliency masks, with the explanations generated by similar methods (listed in the fourth column). These comparisons were carried out over different types of input data (listed in the third column).

Method for explainability (references)	Acronym	Input type	Compared with (references)
Deep-Taylor Decomposition [139]	DTD	Pictorial	GBP, IG, SA, PM in [72]
DeepLIFT [82]	DLT	Pictorial	IG, LRP in [129] IG, SA in [132]
Gradient*Input [82]	GI	Pictorial	GC, GBP and SG [127] GC, GBP, IG, SG [128] IG, LIME, OS, SM, SHAP in [69]
Grad-CAM [140]	GC	Pictorial	GI, GBP, SG [127] GI, GBP, IG, SG [128]
Guided BackProp [141]	GBP	Pictorial	GI, GC, SG [127] GI, GC, IG, SG [128] DTD, IG, SA, PM in [72]
Integrated Gradients [73]	IG	Pictorial	GI, GC, GBP, SG [128] LRP, LIME, OS, SM, SHAP [69] DLT and LRP in [129] DLT, SA in [132] DTD, GBP, SA, PM in [72]
Layer-wise Relevance Propagation [83]	LRP	Pictorial	IG, LIME, OS, SM, SHAP in [69] DLT, IG in [129] SA in [131,134,135] OS in [131]
Local Interpretable Model-Agnostic Explanations [142]	LIME	Pictorial Numerical/Categorical	GI, IG, OS, SM, SHAP in [69]
Occlusion Sensitivity [143]	OS	Pictorial	GI, IG, LIME, SM, SHAP in [69] LRP in [131]
Sensitivity Analysis [144]	SA	Pictorial	LRP in [131] DLT, IG in [132] DTD, GBP, IG, PM in [72] LRP in [134,135]
PatternNet [145]	PM	Pictorial	DTD, GBP, IG, SA in [72]
Saliency Maps [80]	SM	Pictorial	GI, IG, LIME, OS, SHAP in [69]
SHapley Additive exPlanations [146]	SHAP	Pictorial	GI, IG, LIME, OS, SM in [69]
SmoothGrad [147]	SG	Pictorial	GI, GC, GBP [127] GI, GC, GBP and IG [128]
Layer-wise Relevance Propagation [83]	LRP	Textual	SA in [130,134]
Sensitivity Analysis methods [144]	SA	Textual	LRP in [130,134]

size of the intersection of the κ most important features before and after perturbation. [72] instead measure the differences in the model's predictions by checking whether the methods for explainability under analysis satisfy the requirement of 'input invariance' (see Section 3.1). These studies show that all the tested methods (see Table 3) are vulnerable even to small perturbations that do not affect the prediction of the underlying model but change significantly the heat and saliency maps produced by the explainers. Hence, these methods do not satisfy the 'input invariance' requirement [72], meaning that they do not reflect the sensitivity of the model with respect to input perturbations. A comparison of methods for explainability of time series classification was proposed in [133]. The parts of the time series that are deemed relevant by each method for explainability are perturbed to measure the impact on the classification accuracy of an underlying model. The assumption is that highly informative explanations highlight those parts of the time series that are considered discriminative by a classifier. Lastly, [137] compared the completeness of the explanations generated by seven methods for explainability that interpret the logic followed by DNNs by calculating the partial derivatives of the output according to the input variables, analysing the network's weights or perturbing the input variables in various ways.

4.2. Human-centred evaluation

Explanations are effective when they help end-users to build a complete and correct mental representation of the inferential process of a given model. Many scientific articles focused on testing the degree of explainability of one or multiple methods, with a human-in-the-loop approach shown in a diagrammatic way in Fig. 6. These

experiments involved human participants of two kinds. On the one hand, people randomly selected from the lay public and without any prior technical/domain knowledge who were asked to interact with one or more explanatory tools and give feedback by filling questionnaires. On the other hand, domain experts who were asked to give informed opinions on the explanations produced by these methods and verify the consistency of the explanations with the domain knowledge. Examined scientific articles can be clustered into two categories, depending on the nature of questions administered to people (see Table 4). *Qualitative studies* are based upon open-ended questions aimed at achieving deeper insights whereas *quantitative studies* make use of close-ended questions that can be easily analysed statistically.

The first methods for explainability that were tested by human users are those generating textual explanations for Expert Systems (ES) in the 90's [152,153]. These researches aimed at collecting pieces of evidence on whether and how explanations could enhance the user's confidence in the decisions proposed by an ES. Scholars carried out several human-centred evaluations over the years to assess the effects of textual explanations on end-users to other types of systems employing ML models, such as learning systems [170], intelligent agents [151] and neural networks [164]. In the human-centred experiment carried out in [164], 60 participants were split into 4 groups and were respectively presented with textual, vocal, visual (produced with LIME) and vocal as well as no explanations. Participants' feedback about the explanations were collected by using a combination of a 7-point Likert scale and open-ended questions. Likewise, [159] tested the explanations generated by LIME and SHAP on a random forest with a gradient boosted trees model by asking a mix of open-ended questions and by employing a 5-point Likert scale. The impact of explanations on the reliability, trust and reliance of end-users on automated systems was explored in [19]. The participants were presented with photos of the Fort Sill terrain where the presence of a camouflaged soldier was indicated by an automated decision system. Initially, they considered the inference produced by the system to be trustworthy and reliable but, after observing the system making errors, they distrusted even reliable aids, unless an explanation was provided regarding why the system failed. In conclusion, explanations of these errors increased the trust of the participants in the automated system who were asked to estimate their perceived reliability on a 9-point Likert-format scale, ranging from 'not very well' to 'very well'. Similarly, the System Causability Scale is a questionnaire containing 10 5-point Likert questions (from strongly disagree to strongly agree) to measure the usability of user interfaces showing explanations of a ML model [176]. Other scholars proposed to use the Situation Awareness-based Global Assessment Technique (SAGAT) to measure the situation awareness of end-users interacting with autonomous intelligent agents [148]. The aim is to assess whether the provided explanations contain enough information to allow humans to perform their roles. In the SAGAT test, simulations of representative tasks to be done with the support of an autonomous agent are frozen at randomly selected times and users are asked to provide their perception on the situation. The degree of persuasiveness of automatically-generated explanations was analysed in [181]. The influence of explanation factors over the capacity of end-users to understand and develop a mental representation of the internal reference process of a model was investigated in [155,156,171,172,174]. The explanations produced by the analysed methods for explainability consisted of graphical representation of the most relevant features. [171,172] showed that users of simulation-based training systems with virtual players prefer short explanations to long ones where length is defined by the number of elements in an explanation. An element can be a goal or an event of the training program. This was tested by showing to the participants four explanation alternatives of different length (they contained either one or two elements), in the form of DTs, for each action of the virtual players and asked to indicate which alternative they considered the most useful for increasing end-user understanding. The influence factors analysed in [174] were the number of independent variables of a trained linear regression model and the values of these variables for each instance. Some participants were randomly assigned to check either a model that uses only two features or a model that uses eight features. The coefficients of the linear regression model were also presented only to half of the participants. Then, the participants were asked to estimate what would be the output of the model and to correct it in case it was not accurate. As expected, users can easily simulate models with a small number of features whereas, surprisingly, displaying model internal parameters can hamper their ability to notice unusual inputs and correct inaccurate predictions. The method for explainability tested in [155] listed the two most relevant features and the prediction of the model. The prediction was coded with a solid colour taken from a scale running from red, representing the most negative possible outcome, to green, the most positive one. Participants were asked to interact with the system for two weeks at the end of which they were interviewed to collect their feedback and to check whether they gained some insight on the logic of the model to be explained by the explanatory system under analysis. [156] studied to what extent two methods for interpretability, GAMs and SHAP, help data scientists better understand how machine learned models work. They submitted to 197 participants multiple-choice questions in order to quantify their accuracy in reading their explanations, open-ended questions for testing how well they them and close-ended questions (7-point Likert scale) to state their confidence on the previous questions. [85] studied the explainability of an interactive interface, called *Prospector*, containing a set of diagnostic tools that allows end-user, via visual and numerical representations, to understand how the features of a dataset affect the prediction of a model overall. Users can also inspect a specific instance to check its prediction and can weak feature values to see how the model responds. A team of data-scientists was asked to interact with this tool to debug a set of models designed to predict if a patient is at risk of developing diabetes by using a database of electronic medical records. The human experiment consists of interviewing, at the end of the experiment, the data scientists on whether they feel that it was beneficial for their work.

Methods for evaluating the interpretability of data-driven models with a human-in-the-loop approach was proposed in [168,173,179,180]. The approach proposed in [173] identifies some *proxies* which consist of other, simpler models inherently more explainable. For example, a DT is inherently more interpretable than DNNs and the former methods can be used to explain the latter. The authors presented to participants a list of the coefficients of the features used by each proxy and a graphical depiction of its structure (in the form of a DT) and they asked them to identify the correct prediction. [168] proposed instead to assess the comprehensibility of DTs by asking the participants to perform the following tasks: (I) classify a data-point according to the classification tree, (II) explain the classification of a data-point by pointing out which attributes' values must be changed or retained to modify the instance's class, (III) validate a part of the classification tree by asking the participant to check whether a statement about the domain is confirmed/rejected by the tree, (IV) discover new knowledge by finding a property (attribute-value pair) that is unusual for instances from one class, (V) rate the classification tree by giving a subjective opinion on the comprehensibility of the tree and, lastly, (VI) compare two classification tree by saying which one is more comprehensible. The two studies presented in [179,180] analysed the interpretability of predictive models by asking participants to interact with them and fill self-reporting questionnaires. The surveys carried out in [179] aimed at comparing six supervised learning algorithms. These were ranked in order of preference based on the subjective quantification of understandability obtained from the self-reporting questionnaires filled by participants. Pairs of models trained on the same dataset were generated and participants were asked to rate them on a scale where one extreme is 'the first model is the most understandable' to the other extreme 'the second model is the most understandable' via increasingly positive grades. In the study carried out in [180], participants were

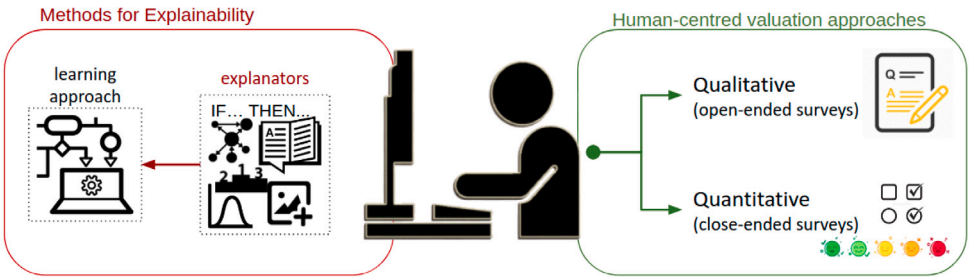


Fig. 6. Diagram of the general process followed by human centred evaluation approaches.

Table 4
List and classification of the scientific articles proposing human-centred approaches to evaluate methods for explainability, classified according to the construction approach, the type of measurement employed (qualitative or quantitative), and the format of their output explanation.

Measure type	Construction approach	Output format	Reference
Qualitative	Autonomous agents	Textual	[19,148]
Qualitative	Context-aware systems	Textual	[49,149]
Qualitative	Context-aware systems	Visual	[149]
Qualitative	Data clustering	Mixed	[150]
Qualitative	Intelligent agents	Textual	[151]
Qualitative	Expert Systems	Textual	[152,153]
Qualitative	Learning Systems	Textual	[154]
Qualitative	ML (Model agnostic)	Visual	[85,155]
Qualitative	ML (Model agnostic)	Numerical	[156]
Qualitative	ML (Model agnostic)	Textual	[157]
Qualitative	Fisher vector classifiers	Visual	[158]
Qualitative	Random Forest	Visual	[159]
Qualitative	Support Vector Machine	Visual	[160]
Qualitative	Neural networks	Visual	[136,161–164]
Quantitative	Case-based Reasoning	Mixed	[165]
Quantitative	Context-aware systems	Textual	[49]
Quantitative	Neural networks	Visual	[166]
Quantitative	Neural networks	Visual	[167]
Quantitative	Decision trees	Mixed	[168]
Quantitative	Kernel-based neural networks	Visual	[169]
Quantitative	Learning Systems	Textual	[170–172]
Quantitative	ML (Model agnostic)	Mixed	[173,174]
Quantitative	ML (Model agnostic)	Rules	[175]
Quantitative	ML (Model agnostic)	Numerical	[156]
Quantitative	ML (Model agnostic)	Textual	[176]
Quantitative	ML (Model agnostic)	Visual	[177]
Quantitative	Naïve Bayes	Textual	[27,178]
Quantitative	Rule-based classifier	Mixed	[179,180]
Quantitative	Rule-based classifier	Mixed	[181]
Quantitative	Decision trees	Mixed	[179,180]
Quantitative	Decision tables	Mixed	[180]
Quantitative	Neural networks	Visual	[164]
Quantitative	Random Forest	Visual	[159]

required to evaluate the explanations of a credit model, trained to accept or reject loan applications, consisting of IF-THEN rules and displayed as a decision tree. They were asked to predict the model's outcome on a new loan application, answer a few yes/no questions such as "Does the model accept all the people with an age above 60?" and rate, for each question, the degree of confidence in the answer on a scale from 1 (Totally not confident) to 5 (Very confident). The authors measured, besides the answer confidence, other two variables about task performance: accuracy, quantified as the percentage of correct answers, and the time in seconds spent to answer the questions. The effectiveness of why-oriented explanation systems in debugging a naïve Bayes learning model for text classification and in context-aware applications were respectively tested in [27,178] and [49,149]. [27,178] asked participants to train a prototype system, based on a Multinomial naïve Bayes classifier, that can learn from users how to automatically classify emails by manually moving a few of them from the inbox into an appropriate folder. The system was subsequently run over a new set of messages, some of which were wrongly classified. The participants had to debug the system by asking 'why' questions via an interactive explanation tool producing textual answers and by giving two types of feedback: some participants could label the most relevant feature (words) whereas the others could only provide more labelled instances (moving more messages to the appropriate folders). At the end of the session, the participants filled a questionnaire to express their opinions on the prototype. In the experiment run in [49,149], participants were invited to interact with a model that predicts whether a person is doing physical exercise or not, based on the body temperature and the pace. They were shown with some examples of inputs and outputs accompanied by graphical (in the form of decision trees) and textual explanations on the logic followed by the model. Half of the participants were presented with why explanations, such as *"Output classified as Not Exercising, because Body Temperature < 37 and Pace < 3"* whereas the other half were presented with why not explanations, such as *"Output not classified as Exercising, because Pace < 3, but not Body Temperature > 37"*. Subsequently, the participants had to fill two questionnaires to check their understanding by asking questions how the system works and to give feedback on the explanations in terms of understandability, trust and usefulness. Both questionnaires contained a mix of open and close questions, where the close ones consisted of a 7-point Likert scale. In a preliminary study [157], the authors showed a set of Kandinsky Patterns to 271 participants who were asked to classify them and give an explanation of their decisions. The scope

was to compare these with machine-generated explanations in order to verify if they highlight the same properties of the same images. A qualitative evaluation of the interpretability of the Mind the Gap Model (MGM) method for explainability was gathered in [150]. MGM clustered the data of an input dataset according to the most relevant features. The participants were presented with the raw data and the data clustered with MGM and k-means and were asked to write a 2–3 sentence executive summary of each data representation within 5 min. They all found impossible to summarise the raw data, not being able to complete the task in the given amount of time, but they managed to do so on the data with clustered MGM and k-means. To test Bayesian Case Model (BCM), [165] asked the participants to assign sixteen recipes, described only by a set of ingredients, to the right category (so a recipe of cookies had to be classified under 'cookie') and then they counted how many of them were correctly classified. BCM was compared with Latent Dirichlet Allocation (LDA), a clustering approach based on extracting similar characteristics in the data. The need for XAI in Intelligent Tutoring Systems (ITS) was explored in [154]. The participants in the study were asked to use an ITS that provided tools to explore and explain an algorithm solving constraint satisfaction problems in an interactive simulation. The participants were instructed by the exploration tool with a textual hint message. They could select to have the hint explained by the explanatory tool which was also designed to solicit their suggestions on the explanations they would like to see for each hint by presenting them a checkbox list with the following options: 'why the system gave this hint', 'how the system chose this hint', 'some other explanation about this hint' (followed by an open-text field) and 'I do not want an explanation for this hint.'

Many scholars proposed human-centred evaluation approaches for methods for explainability generating visual explanations. The participants selected in the study in [166] were presented with whole images misclassified by a DNN and the visual explanations generated by LIME and MMD-critic. For example, a photo of a Jeep with a red-cross was wrongly classified as an ambulance and the visual explanations show the red-cross with the rest of the image greyed out. Participants were asked to say whether the class predicted by the model was nonetheless relevant where the possible answers to questions like 'Is the label Red-Cross relevant?' were 'yes' and 'no'. A similar experiment was carried out in [167] to test GAN Dissection. Participants were presented with images reporting highlighted patches showing the most highly-activating regions for each unit at each intermediate convolutional layer of a DNN. Each layer was aligned with a semantic and were given labels across a range of objects, parts, scenes, textures, materials, and colours. For example, if a DNN was trained to recognise a list of object in input images, like flowers and cars, the semantic consists of this list and images showing flowers were labelled 'flower'. The participants were asked to say if the highlighted patches were pertinent to the label by answering yes/no questions. The capacity of Anchors and LIME in helping end-users forecasting the predictions of an image classifier was tested in [175]. Participants were asked to predict the output class assigned by the classifier to ten random test instances before and ten instances after seeing two rounds of explanations generated by either Anchors or LIME. A few scholars conducted human-centred studies to test the interpretability of the heat-maps generated with LRP. [158,160] applied it respectively to test models built with Fisher vector classifiers for object recognition in images and to SVMs, trained on videos, to understand and interpret action recognition and to check whether LRP allows identifying in which point of the video the important action happens. By visually inspecting the heat-maps, the authors of the two studies could show a possible weakness of the underlying classifiers by looking at the regions highlighted in the heat-maps and examining whether they were relevant for the recognition of an object (or at least part of it) in images and of the areas of video frames showing the actions performed in the video. Similarly, [163] employed LRP with DNNs for electroencephalography (EEG) data analysis. The predictions of the trained DNNs are transformed with LRP, for every trial, into heat-maps indicating the relevance of each data point. The relevance information can be plotted as a scalp topography that can be visually inspected by experts to check if there are neurophysiologically plausible patterns in the EEG data. [162] used LRP for computing the contribution of contextual words to arbitrary hidden states in the attention-based encoder–decoder framework of neural machine translation (NMT). As per the previous studies, the authors checked if the translation made by the NMT (Japanese–English) were right or wrong and what types of errors were made more frequently. The participants in [177] were asked to interact with explAIner which showed them explanations generated by LRP and Saliency Analysis of both a simple and a complex network trained on the MNIST dataset and, in the meantime, to communicate their thoughts and actions by 'thinking aloud'. The sessions were audio-recorded and screen-captured. At the end of the sessions, participants were also interviewed to provide qualitative feedback on the overall experience. Another method for explainability producing visual explanations as maps, Grad-CAM, was applied to multivariate time series from photovoltaic power plants that were fed into a neural network to forecast the energy production of these plants in different weather conditions [161]. Grad-CAM was used to explain which features, such as environmental temperature, wind bearing or humidity, or any combinations of these features were responsible, at different time intervals, for a given prediction. The results showed that Grad-CAM was able to visualise the network's attention over the time dimension and the features of multivariate time series data. [136] compared other three methods, namely Activation Maximisation, Sampling Unit and Linear Combination, designed to produce explanations as heat-maps of the most relevant features of the input images. Activation Maximisation consists of selecting the input features that maximise the activation of a single hidden neuron. Sampling Unit consists of setting the value of a neuron to one and calculating the probability with which each sample is assigned to a class. Lastly, Linear Combination consists of choosing the largest weights of the connections between neurons of two adjacent layers. The authors did not use any objective measure to compare these methods but from a qualitative visual inspection of the heat-maps and an exploration of connections between all of them.

5. Final remarks and recommendations

The development of methods to explain the inner logic of either a learning algorithm, a model induced from it, or a knowledge-based approach for inference, is now generally recognised as a core area of AI that is often referred to as eXplainable Artificial Intelligence (XAI). Other widely used terms exist, such as 'Interpretable Machine Learning', but with XAI we would like to emphasise the wider applicability of this emerging and fascinating field of research. Several scientific studies are published every year, with many workshops and conferences organised around the world to propose novel methods and disseminate findings. Although this has led to the production of an abundance of knowledge, unfortunately, this is very scattered. Many reviews attempted to fill this gap by organising this vast knowledge but, given the complexity of the task, they ended up focusing only on specific aspects of explainability. Many studies within XAI have focused on improving the quality and widening the variety of explanations for several types of learning approaches with data. To achieve this, scholars expanded their research horizons by incorporating the knowledge developed in other fields, like psychology, philosophy and social sciences, into the design of their novel methods for explainability. The goal was to improve the structure, efficiency and efficacy on people's understanding of automatically generated explanations. This research effort has produced many definitions of explainability and identified numerous notions related to it, such as interpretability, understandability, comprehensibility and justifiability, just to mention a few. Coupled to these notions, different qualitative and quantitative measures have also been coined for their assessment. Scholars have proposed some theories to structure an

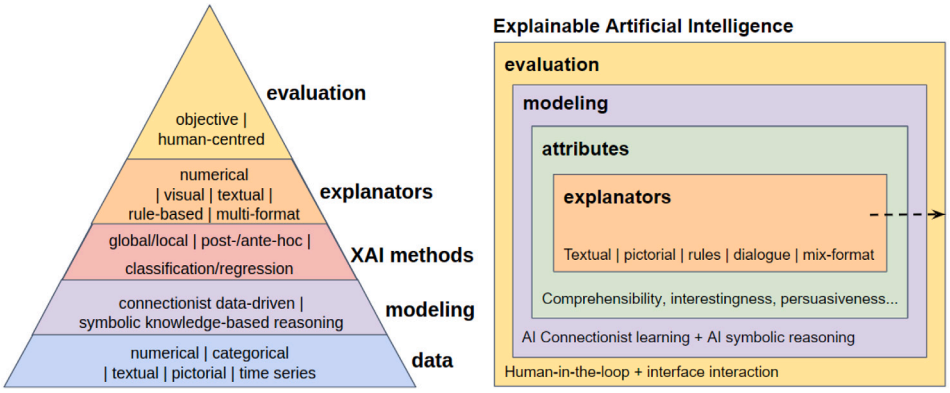


Fig. 7. State of the art (a) and envisioned (b) frameworks for eXplainable Artificial Intelligence.

effective explanation by considering the needs of different types of end-users, the language, the format and its content. To verify these theories, human-centred studies have been mainly carried out to collect people's feedback on machine-generated explanations, usually in the form of open or close-ended questions. Limited research has been done on their automatic assessment without human intervention. The ultimate goal of an explanation is to help end-users build a complete and correct mental model of the inferential process of either a learning algorithm or a knowledge-based system and to promote trust for its outputs.

Despite the large number and variety of methods, notions, theories and qualitative or quantitative approaches for explainability proposed so far, there are still important scientific issues that must be tackled. Firstly, there is no agreement among scholars on what an explanation exactly is and which are the salient properties that should be considered to make it effective and understandable for end-users, in particular non-experts. Therefore, future research should focus on the definition of explainability and structured formats of explanations that incorporate as many notions as possible in order to cover a wide variety of attributes and dimensions of explainability. Secondly, the construct of explainability might be thought as a concept borrowed from psychology, since it is strictly connected to humans, and it is also linked to other constructs such as trust, transparency and privacy. Therefore, the involvement of humans is key, as they are the final consumers of artificial explanations, and thus research on psychometrics should be performed by scholars interested in the development of explainable intelligent systems, as it usually occurs in studies involving psychological measurements. To support this research direction, we recommend exploiting knowledge and experiences belonging to the field of Human–Computer Interaction and its advances for the development of interactive explanatory interfaces [182,183]. Thirdly, the concept of explainability has been invoked in several fields, such as medicine, social sciences and human–computer interaction, just to mention a few. Therefore, future research should focus on the development of a definition of explainability that have a wider applicability across contexts and domains of applications of intelligent systems. This should always take into consideration the existing trade-off between the dimensions of model accuracy and its interpretability/explainability which are currently inversely correlated. One possible suggestion is the use of methods that take advantage of modern learning techniques, to maximise the former dimension, and reasoning approaches to optimise the latter dimension. The assumption is that integrating connectionist and symbolic paradigms might be the most efficient way to produce meaningful and precise explanations [184]. Advances on these two paradigms are immense, however, their intersection is under explored. For example, on one hand, a school of thought suggests to firstly train accurate models from data and then wrap them with a reasoning layer [185]. This layer, for instance, can be produced by exploiting advances in defeasible reasoning and argumentation [186–188] making use of knowledge-bases constructed with a human-in-the-loop approach. On the other hand, another direction is to promote the use of neuro-symbolic learning and reasoning in parallel, each one informing the other at all stages of model construction and evaluation [189].

In summary, an high-level structure of the current state-of-the-are in XAI is depicted in Fig. 7 (part a). On one hand, here, emphasis has been placed so far on the sequence of research activities currently and often performed by several scholars, their dependencies and order (from bottom to top, Fig. 7 (part a)). This sequence usually starts from input data of different formats that is then used for modelling purposes, employing connectionist data-driven learning, or symbolic reasoning knowledge-based paradigms. After a model has been formed, then an XAI method is designed and applied for its analysis, knowledge discovery, supporting its interpretability. This phase provides the end-users of these models with one or more explanators for the purpose of supporting its explainability and interpretability. Eventually, very few scholars have proposed approaches for evaluating such layer of explainability and explanators, either proposing formal, objective metrics, or performing human-centred evaluations involving model designers and end-users. On the other end, what we believe is an ideal framework for XAI is depicted in Fig. 7 (part b). Here, the main focus should be on the explanators, which is what end-users will ultimately interact with (from centre outward). The development of explanators should be designed by taking into account the multiple attributes and notions that are linked to the psychological construct of explainability. Subsequently, scholars should focus on the modelling phase, preferably using both connectionist and symbolic paradigms from Artificial Intelligence. This will allow the development of models that are both robust in terms of accuracy, but also intrinsically interpretable during all the stages of construction. Eventually, the last phase should focus on the evaluation of explainability of such models with a human-in-the-loop approach, involving both model designers and end-users, as well as the development of interactive interfaces for supporting model interpretability and inference explainability.

CRediT authorship contribution statement

Giulia Vilone: Conceptualization, Methodology, Investigation, Visualization, Writing - original draft. **Luca Longo:** Supervision, Writing - review & editing.

Declaration of competing interest

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

References

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

[2] Alun Preece, Asking "Why" in AI: Explainability of intelligent systems–perspectives and challenges, Intell. Syst. Account. Finance Manag. 25 (2) (2018) 63–72, http://dx.doi.org/10.1002/isaf.1422.

[3] Weiquan Wang, Izak Benbasat, Recommendation agents for electronic commerce: Effects of explanation facilities on trusting beliefs, J. Manage. Inf. Syst. 23 (4) (2007) 217–246, http://dx.doi.org/10.2753/mis0742-1222230410.

[4] Sebastian Lapuschkin, Stephan Wäldchen, Alexander Binder, Grégoire Montavon, Wojciech Samek, Klaus-Robert Müller, Unmasking clever hans predictors and assessing what machines really learn, Nat. Commun. 10 (1) (2019) 1096, http://dx.doi.org/10.1038/s41467-019-08987-4.

[5] Cynthia Rudin, Algorithms for interpretable machine learning, in: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, New York, New York, USA, 2014, p. 1519, http://dx.doi.org/10.1145/2623330.2630823.

[6] Cynthia Rudin, Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, Nat. Mach. Intell. 1 (5) (2019) 206, http://dx.doi.org/10.1038/s42256-019-0048-x.

[7] Jean-Marc Fellous, Guillermo Sapiro, Andrew Rossi, Helen S. Mayberg, Michele Ferrante, Explainable artificial intelligence for neuroscience: Behavioral neurostimulation, Front. Neurosci. 13 (2019) 1346, http://dx.doi.org/10.3389/fnins.2019.01346.

[8] Maria Fox, Derek Long, Daniele Magazzeni, Explainable planning, in: IJCAI 2017 Workshop on Explainable Artificial Intelligence, XAI, International Joint Conferences on Artificial Intelligence, Inc, Melbourne, Australia, 2017, pp. 24–30.

[9] Filip Karlo Došilović, Mario Brčić, Nikica Hlupić, Explainable artificial intelligence: A survey, in: 41st International Convention on Information and Communication Technology, Electronics and Microelectronics, MIPRO, IEEE, Opatija, Croatia, 2018, pp. 0210–0215, http://dx.doi.org/10.23919/mipro.2018.8400040.

[10] Eva Thelisson, Kirtan Padh, L. Elisa Celis, Regulatory mechanisms and algorithms towards trust in AI/ML, in: IJCAI Workshop on Explainable AI, XAI, International Joint Conferences on Artificial Intelligence, Inc, Melbourne, Australia, 2017, pp. 53–57.

[11] Eva Thelisson, Towards trust, transparency, and liability in AI/AS systems, in: Proceedings of the 26th International Joint Conference on Artificial Intelligence, International Joint Conferences on Artificial Intelligence, Inc, Melbourne, Australia, 2017, pp. 5215–5216, http://dx.doi.org/10.24963/IJCAI.2017/767.

[12] Sandra Wachter, Brent Mittelstadt, Luciano Floridi, Transparent, explainable, and accountable AI for robotics, Sci. Robot. 2 (6) (2017) http://dx.doi.org/10.1126/scirobotics.aan6080.

[13] Wojciech Samek, Klaus-Robert Müller, Towards explainable artificial intelligence, in: Explainable AI: Interpreting, Explaining and Visualizing Deep Learning, Springer, Cham, Switzerland, 2019, pp. 5–22, http://dx.doi.org/10.1007/978-3-030-28954-6_1.

[14] Carmen Lacave, Francisco J. Díez, A review of explanation methods for Bayesian networks, Knowl. Eng. Rev. 17 (2) (2002) 107–127, http://dx.doi.org/10.1017/s026988890200019x.

[15] David Martens, Bart Baesens, Tony Van Gestel, Jan Vanthienen, Comprehensible credit scoring models using rule extraction from support vector machines, European J. Oper. Res. 183 (3) (2007) 1466–1476, http://dx.doi.org/10.1016/j.ejor.2006.04.051.

[16] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al., Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI, Inf. Fusion 58 (2020) 82–115, http://dx.doi.org/10.1016/j.inffus.2019.12.012.

[17] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, Dino Pedreschi, A survey of methods for explaining black box models, ACM Comput. Surv. (CSUR) 51 (5) (2018) 93:1–93:42, http://dx.doi.org/10.1145/3236009.

[18] Tim Miller, Explanation in artificial intelligence: Insights from the social sciences, Artificial Intelligence 267 (2019) 1–38, http://dx.doi.org/10.1016/j.artint.2018.07.007.

[19] Mary T. Dzindolet, Scott A. Peterson, Regina A. Pomranky, Linda G. Pierce, Hall P. Beck, The role of trust in automation reliance, Int. J. Hum.-Comput. Stud. 58 (6) (2003) 697–718, http://dx.doi.org/10.1016/s1071-5819(03)00038-7.

[20] Nava Tintarev, Judith Masthoff, A survey of explanations in recommender systems, in: IEEE 23rd International Conference on Data Engineering Workshop, IEEE, Istanbul, Turkey, 2007, pp. 801–810, http://dx.doi.org/10.1109/icdew.2007.4401070.

[21] Zachary C. Lipton, The mythos of model interpretability, Commun. ACM 61 (10) (2018) 36–43.

[22] Taehyun Ha, Sangwon Lee, Sangyeon Kim, Designing explainability of an artificial intelligence system, in: Proceedings of the Technology, Mind, and Society, ACM, Washington, District of Columbia, USA, 2018, p. 14:1, http://dx.doi.org/10.1145/3183654.3183683.

[23] Urszula Chajewska, Joseph Y. Halpern, Defining explanation in probabilistic systems, in: Proceedings of the Thirteenth Conference on Uncertainty in Artificial Intelligence, Morgan Kaufmann Publishers Inc., Providence, Rhode Island, USA, 1997, pp. 62–71.

[24] Andreas Holzinger, Georg Langs, Helmut Denk, Kurt Zatloukal, Heimo Müller, Causability and explainability of artificial intelligence in medicine, Wiley Interdiscip. Rev.: Data Min. Knowl. Discovery 9 (2019) e1312, http://dx.doi.org/10.1002/widm.1312.

[25] Tim Miller, Piers Howe, Liz Sonenberg, Explainable AI: Beware of inmates running the asylum or: How I learnt to stop worrying and love the social and behavioural sciences, in: IJCAI Workshop on Explainable AI, XAI, International Joint Conferences on Artificial Intelligence, Inc, Melbourne, Australia, 2017, pp. 36–42.

[26] Hoa Khanh Dam, Truyen Tran, Aditya Ghose, Explainable software analytics, in: Proceedings of the 40th International Conference on Software Engineering: New Ideas and Emerging Results, ACM, Gothenburg, Sweden, 2018, pp. 53–56, http://dx.doi.org/10.1145/3183399.3183424.

[27] Todd Kulesza, Margaret Burnett, Weng-Keen Wong, Simone Stumpf, Principles of explanatory debugging to personalize interactive machine learning, in: Proceedings of the 20th International Conference on Intelligent User Interfaces, ACM, Atlanta, Georgia, USA, 2015, pp. 126–137, http://dx.doi.org/10.1145/2678025.2701399.

[28] Todd Kulesza, Simone Stumpf, Margaret Burnett, Sherry Yang, Irwin Kwan, Weng-Keen Wong, Too much, too little, or just right? Ways explanations impact end users' mental models, in: IEEE Symposium on Visual Languages and Human-Centric Computing, VL/HCC, IEEE, Raleigh, NC, USA, 2013, pp. 3–10, http://dx.doi.org/10.1109/vlhcc.2013.6645235.

[29] Raha Moraffah, Mansooreh Karami, Ruocheng Guo, Adrienne Raglin, Huan Liu, Causal interpretability for machine learning-problems, methods and evaluation, ACM SIGKDD Explor. Newsl. 22 (1) (2020) 18–33, http://dx.doi.org/10.1145/3400051.3400058.

[30] Xiaocong Cui, Jung Min Lee, J. Hsieh, An integrative 3C evaluation framework for explainable artificial intelligence, in: AI and Semantic Technologies for Intelligent Information Systems, SIGODIS, AIS eLibrary, Cancún, Mexico, 2019, pp. 1–10.

[31] Irit Askira-Gelman, Knowledge discovery: Comprehensibility of the results, in: Proceedings of the Thirty-First Hawaii International Conference on System Sciences, vol. 5, IEEE, Hawaii, 1998, pp. 247–255, http://dx.doi.org/10.1109/hicss.1998.648319.

[32] Jose M. Alonso, Ciro Castiello, Corrado Mencar, A bibliometric analysis of the explainable artificial intelligence research field, in: International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems, Springer, Cádiz, Spain, 2018, pp. 3–15.

[33] Adrien Bibal, Benoît Frénay, Interpretability of machine learning models and representations: An introduction, in: Proceedings on the European Symposium on Artificial Neural Networks, ESANN, i6doc, Bruges, Belgium, 2016, pp. 77–82.

[34] Ivan Bratko, Machine learning: Between accuracy and interpretability, in: Learning, Networks and Statistics, Springer, Vienna, Austria, 1997, pp. 163–177, http://dx.doi.org/10.1007/978-3-7091-2668-4_10.

[35] Derek Doran, Sarah Schulz, Tarek R. Besold, What does explainable AI really mean? A new conceptualization of perspectives, in: 16th International Conference of the Italian Association of Artificial Intelligence, 2017. Workshop on Comprehensibility and Explanation in AI and ML, Cex, University of Bremen, Germany, Bari, Italy, 2017, pp. 1–8.

[36] Alex A. Freitas, Are we really discovering interesting knowledge from data? Expert Update BCS-SGAI Mag. 9 (1) (2006) 41–47.

[37] Randy Goebel, Ajay Chander, Katharina Holzinger, Freddy Lecue, Zeynep Akata, Simone Stumpf, Peter Kieseberg, Andreas Holzinger, Explainable AI: The new 42? in: International Cross-Domain Conference for Machine Learning and Knowledge Extraction, Springer, Hamburg, Germany, 2018, pp. 295–303, http://dx.doi.org/10.1007/978-3-319-99740-7_21.

[38] David S. Watson, Jenny Krutzinna, Ian N. Bruce, Christopher E.M. Griffiths, Iain B. McInnes, Michael R. Barnes, Luciano Floridi, Clinical applications of machine learning algorithms: Beyond the black box, BMJ 364 (2019) l886, http://dx.doi.org/10.1136/bmj.l886.

[39] Alexander Jung, Pedro Henrique Juliano Nardelli, An information-theoretic approach to personalized explainable machine learning, IEEE Signal Process. Lett. 27 (2020) 825–829, http://dx.doi.org/10.1109/LSP.2020.2993176.

[40] Karl de Fine Licht, Jenny de Fine Licht, Artificial intelligence, transparency, and public decision-making, AI Soc. (2020) 1–10, http://dx.doi.org/10.1007/s00146-020-00960-w.

[41] Nava Tintarev, Judith Masthoff, Designing and evaluating explanations for recommender systems, in: Recommender Systems Handbook, Springer, Boston, Massachusetts, USA, 2011, pp. 479–510, http://dx.doi.org/10.1007/978-0-387-85820-3_15.

[42] Nava Tintarev, Judith Masthoff, Explaining recommendations: Design and evaluation, in: Recommender Systems Handbook, Springer, Boston, Massachusetts, USA, 2015, pp. 353–382, http://dx.doi.org/10.1007/978-1-4899-7637-6_10.
[43] Ajay Chander, Ramya Srinivasan, Evaluating explanations by cognitive value, in: International Cross-Domain Conference for Machine Learning and Knowledge Extraction, Springer, Hamburg, Germany, 2018, pp. 314–328.
[44] Yu Zhang, Sarath Sreedharan, Anagha Kulkarni, Tathagata Chakraborti, Hankz Hankui Zhuo, Subbarao Kambhampati, Plan explicability and predictability for robot task planning, in: IEEE International Conference on Robotics and Automation, ICRA, IEEE, Singapore, 2017, pp. 1313–1320, http://dx.doi.org/10.1109/icra.2017.7989155.
[45] David Alvarez-Melis, Tommi S. Jaakkola, Towards robust interpretability with self-explaining neural networks, in: Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems, NeurIPS, Neural Information Processing Systems Foundation, Inc., Montréal, Canada, 2018, pp. 7786–7795.
[46] Ashraf Abdul, Jo Vermeulen, Danding Wang, Brian Y Lim, Mohan Kankanhalli, Trends and trajectories for explainable, accountable and intelligible systems: An HCI research agenda, in: Proceedings of the CHI Conference on Human Factors in Computing Systems, ACM, Montréal, Canada, 2018, pp. 582–599, http://dx.doi.org/10.1145/3173574.3174156.
[47] Michael Chromik, Malin Eiband, Sarah Theres Völkel, Daniel Buschek, Dark patterns of explainability, transparency, and user control for intelligent systems, in: Joint Proceedings of the ACM IUI 2019 Workshops Co-Located with the 24th ACM Conference on Intelligent User Interfaces ACM IUI, vol. 2327, CEUR-WS.org, Los Angeles, California, USA, 2019.
[48] Jonathan Dodge, Sean Penney, Andrew Anderson, Margaret M Burnett, What should be in an XAI explanation? What IFT reveals, in: IUI Workshop 7: Explainable Smart Systems - ExSS, CEUR-WS.org, Tokyo, Japan, 2018.
[49] Brian Y. Lim, Anind K. Dey, Daniel Avrahami, Why and why not explanations improve the intelligibility of context-aware intelligent systems, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM, Boston, Massachusetts, USA, 2009, pp. 2119–2128, http://dx.doi.org/10.1145/1518701.1519023.
[50] Brian Y. Lim, Qian Yang, Ashraf M. Abdul, Danding Wang, Why these explanations? Selecting intelligibility types for explanation goals, in: Joint Proceedings of the ACM IUI 2019 Workshops Co-Located with the 24th ACM Conference on Intelligent User Interfaces ACM IUI, vol. 2327, CEUR-WS.org, Los Angeles, California, USA, 2019.
[51] Johanna D. Moore, Cécile L. Paris, Planning text for advisory dialogues: Capturing intentional and rhetorical information, Comput. Linguist. 19 (4) (1993) 651–694.
[52] Prashan Madumal, Tim Miller, Liz Sonenberg, Frank Vetere, A grounded interaction protocol for explainable artificial intelligence, in: Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems, International Foundation for Autonomous Agents and Multiagent Systems, Montréal, Canada, 2019, pp. 1033–1041.
[53] Alex A. Freitas, On rule interestingness measures, in: Research and Development in Expert Systems XV, Springer, United Kingdom, 1999, pp. 147–158, http://dx.doi.org/10.1016/s0950-7051(99)00019-2.
[54] Pedro Sequeira, Eric Yeh, Melinda T. Gervasio, Interestingness elements for explainable reinforcement learning through introspection, in: Joint Proceedings of the ACM IUI 2019 Workshops Co-Located with the 24th ACM Conference on Intelligent User Interfaces ACM IUI, vol. 2327, CEUR-WS.org, Los Angeles, California, USA, 2019.
[55] Or Biran, Courtenay Cotton, Explanation and justification in machine learning: A survey, in: IJCAI 2017 Workshop on Explainable Artificial Intelligence, XAI, International Joint Conferences on Artificial Intelligence, Inc, Melbourne, Australia, 2017, pp. 8–13.
[56] André Carrington, Paul Fieguth, Helen Chen, Measures of model interpretability for model selection, in: International Cross-Domain Conference for Machine Learning and Knowledge Extraction, Springer, Hamburg, Germany, 2018, pp. 329–349.
[57] Grégoire Montavon, Wojciech Samek, Klaus-Robert Müller, Methods for interpreting and understanding deep neural networks, Digit. Signal Process. 73 (2017) 1–15, http://dx.doi.org/10.1016/j.dsp.2017.10.011.
[58] Ribana Roscher, Bastian Bohn, Marco F. Duarte, Jochen Garcke, Explainable machine learning for scientific insights and discoveries, IEEE Access 8 (2020) 42200–42216, http://dx.doi.org/10.1109/ACCESS.2020.2976199.
[59] Isabel Sassoon, Nadin Kökciyan, Elizabeth Sklar, Simon Parsons, Explainable argumentation for wellness consultation, in: International Workshop on Explainable, Transparent Autonomous Agents and Multi-Agent Systems, Springer, Cham, Switzerland, 2019, pp. 186–202.
[60] Mukund Sundararajan, Jinhua Xu, Ankur Taly, Rory Sayres, Amir Najmi, Exploring principled visualizations for deep network attributions, in: Joint Proceedings of the ACM IUI 2019 Workshops Co-Located with the 24th ACM Conference on Intelligent User Interfaces, ACM IUI, vol. 2327, CEUR-WS.org, Los Angeles, California, USA, 2019.
[61] Vanya Van Belle, Paulo Lisboa, Research directions in interpretable machine learning models, in: 21st European Symposium on Artificial Neural Networks, ESANN, i6doc, Bruges, Belgium, 2013, pp. 533–541.
[62] Alfredo Vellido, José David Martín-Guerrero, Paulo J.G. Lisboa, Making machine learning models interpretable, in: European Symposium on Artificial Neural Networks, ESANN, vol. 12, i6doc, Bruges, Belgium, 2012, pp. 163–172.
[63] Shang-Ming Zhou, John Q. Gan, Low-level interpretability and high-level interpretability: A unified view of data-driven interpretable fuzzy system modelling, Fuzzy Sets and Systems 159 (23) (2008) 3091–3131, http://dx.doi.org/10.1016/j.fss.2008.05.016.
[64] Mark Coeckelbergh, Artificial intelligence, responsibility attribution, and a relational justification of explainability, Sci. Eng. Ethics 26 (4) (2020) 2051–2068, http://dx.doi.org/10.1007/s11948-019-00146-8.
[65] Shirley Gregor, Izak Benbasat, Explanations from intelligent systems: Theoretical foundations and implications for practice, MIS Q. 23 (4) (1999) 497–530, http://dx.doi.org/10.2307/249487.
[66] Claus Weihs, Um Sondhauss, Combining mental fit and data fit for classification rule selection, in: Exploratory Data Analysis in Empirical Research, Springer, Munich, Germany, 2003, pp. 188–203, http://dx.doi.org/10.1007/978-3-642-55721-7_21.
[67] Alex A. Freitas, Comprehensible classification models: A position paper, ACM SIGKDD Explor. Newslett. 15 (1) (2014) 1–10, http://dx.doi.org/10.1145/2594473.2594475.
[68] Shixia Liu, Xiting Wang, Mengchen Liu, Jun Zhu, Towards better analysis of machine learning models: A visual analytics perspective, Vis. Inform. 1 (1) (2017) 48–56, http://dx.doi.org/10.1016/j.visinf.2017.01.006.
[69] David Alvarez-Melis, Tommi S. Jaakkola, On the robustness of interpretability methods, in: Proceedings of the 2018 ICML Workshop in Human Interpretability in Machine Learning, ICML, Stockholm, Sweden, 2018, pp. 66–71.
[70] Rowan McAllister, Yarin Gal, Alex Kendall, Mark Van Der Wilk, Amar Shah, Roberto Cipolla, Adrian Vivian Weller, Concrete problems for autonomous vehicle safety: Advantages of Bayesian deep learning, in: Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, International Joint Conferences on Artificial Intelligence, Inc, Melbourne, Australia, 2017, pp. 4745–4753, http://dx.doi.org/10.24963/ijcai.2017/661.
[71] Kacper Sokol, Peter Flach, Explainability fact sheets: A framework for systematic assessment of explainable approaches, in: Proceedings of the Conference on Fairness, Accountability, and Transparency, Barcelona, Spain, 2020, pp. 56–67, http://dx.doi.org/10.1145/3351095.3372870.
[72] Pieter-Jan Kindermans, Sara Hooker, Julius Adebayo, Maximilian Alber, Kristof T. Schütt, Sven Dähne, Dumitru Erhan, Been Kim, The (un)reliability of saliency methods, in: NIPS Workshop on Explaining and Visualizing Deep Learning, NIPS, Long Beach, California, USA, 2017, pp. 93–101.
[73] Mukund Sundararajan, Ankur Taly, Qiqi Yan, Axiomatic attribution for deep networks, in: Proceedings of the 34th International Conference on Machine Learning-Volume 70, JMLR.org, Sydney, Australia, 2017, pp. 3319–3328.
[74] Fabian Offert, "I know it when i see it". Visualization and intuitive interpretability, in: NIPS Symposium on Interpretable Machine Learning, NIPS, Long Beach, California, USA, 2017, pp. 43–46.
[75] Koji Maruhashi, Masaru Todoriki, Takuya Ohwa, Keisuke Goto, Yu Hasegawa, Hiroya Inakoshi, Hirokazu Anai, Learning multi-way relations via tensor decomposition with neural networks, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32, no. 1, 2018.
[76] Stefan Larsson, Fredrik Heintz, Transparency in artificial intelligence, Internet Policy Rev. 9 (2) (2020) http://dx.doi.org/10.14763/2020.2.1469.
[77] Joseph B. Lyons, Being transparent about transparency, in: AAAI Spring Symposium, AAAI Press, Palo Alto, California, USA, 2013, pp. 48–53.
[78] Adrian Weller, Challenges for transparency, in: Proceedings of the ICML Workshop on Human Interpretability in Machine Learning, ICML, Sydney, Australia, 2017, pp. 55–62.
[79] Andrés Páez, The pragmatic turn in explainable artificial intelligence (XAI), Minds Mach. 29 (2019) 1–19, http://dx.doi.org/10.1007/s11023-019-09502-w.
[80] Karen Simonyan, Andrea Vedaldi, Andrew Zisserman, Deep inside convolutional networks: Visualising image classification models and saliency maps, in: Proceedings of ICLR Workshop, ICLR, Banff, Canada, 2014.
[81] Yin Lou, Rich Caruana, Johannes Gehrke, Intelligible models for classification and regression, in: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, Beijing, China, 2012, pp. 150–158, http://dx.doi.org/10.1145/2339530.2339556.
[82] Avanti Shrikumar, Peyton Greenside, Anshul Kundaje, Learning important features through propagating activation differences, in: Proceedings of the 34th International Conference on Machine Learning-Volume 70, JMLR.org, Sydney, Australia, 2017, pp. 3145–3153.
[83] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, Wojciech Samek, On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation, PLoS One 10 (7) (2015) e0130140.
[84] Jonathan L. Herlocker, Joseph A. Konstan, John Riedl, Explaining collaborative filtering recommendations, in: Proceedings of the 2000 ACM Conference on Computer Supported Cooperative Work, ACM, Philadelphia, Pennsylvania, USA, 2000, pp. 241–250, http://dx.doi.org/10.1145/358916.358995.

[85] Josua Krause, Adam Perer, Kenney Ng, Interacting with predictions: Visual inspection of black-box machine learning models, in: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, ACM, San Jose, California, USA, 2016, pp. 5686–5697, http://dx.doi.org/10.1145/2858036.2858529.

[86] Mireia Ribera, Àgata Lapedriza, Can we do better explanations? A proposal of user-centered explainable AI, in: Joint Proceedings of the ACM IUI 2019 Workshops Co-Located with the 24th ACM Conference on Intelligent User Interfaces, ACM IUI, vol. 2327, CEUR-WS.org, Los Angeles, California, USA, 2019.

[87] Maartje M.A. de Graaf, Bertram F. Malle, How people explain action (and autonomous intelligent systems should too), in: AAAI Fall Symposium on Artificial Intelligence for Human-Robot Interaction, AAAI Press, Arlington, Virginia, USA, 2017, pp. 19–26.

[88] Maaike Harbers, Karel van den Bosch, John-Jules Ch Meyer, A study into preferred explanations of virtual agent behavior, in: International Workshop on Intelligent Virtual Agents, Springer, Amsterdam, Netherlands, 2009, pp. 132–145, http://dx.doi.org/10.1007/978-3-642-04380-2_17.

[89] Jon Arne Glomsrud, André Ødegårdstuen, Asun Lera St Clair, Øyvind Smogeli, Trustworthy versus explainable AI in autonomous vessels, in: Proceedings of the International Seminar on Safety and Security of Autonomous Vessels (ISSAV) and European STAMP Workshop and Conference (ESWC), Sciendo, Espoo, Finland, 2020, pp. 37–47, http://dx.doi.org/10.2478/9788395669606-004.

[90] Michael R. Wick, William B. Thompson, Reconstructive explanation: Explanation as complex problem solving, in: Proceedings of the Eleventh International Joint Conference on Artificial Intelligence, International Joint Conferences on Artificial Intelligence, Inc, Detroit, Michigan, USA, 1989, pp. 135–140.

[91] Michael R. Wick, Second generation expert system explanation, in: Second Generation Expert Systems, Springer, Berlin, Germany, 1993, pp. 614–640, http://dx.doi.org/10.1007/978-3-642-77927-5_26.

[92] Steven R. Haynes, Mark A. Cohen, Frank E. Ritter, Designs for explaining intelligent agents, Int. J. Hum.-Comput. Stud. 67 (1) (2009) 90–110, http://dx.doi.org/10.1016/j.ijhcs.2008.09.008.

[93] Raymond Sheh, Isaac Monteath, Introspectively assessing failures through explainable artificial intelligence, in: IROS Workshop on Introspective Methods for Reliable Autonomy, iliad-project.eu, Vancouver, Canada, 2017, pp. 40–47.

[94] Regina Barzilay, Daryl McCullough, Owen Rambow, Jonathan DeCristofaro, Tanya Korelsky, Benoit Lavoie, A new approach to expert system explanations, in: Proceedings of the Ninth International Workshop on Natural Language Generation, Association for Computational Linguistics, Niagara-on-the-Lake, Ontario, Canada, 1998, pp. 78–87.

[95] Tania Lombrozo, The structure and function of explanations, Trends Cognitive Sci. 10 (10) (2006) 464–470, http://dx.doi.org/10.1016/j.tics.2006.08.004.

[96] J.L. Weiner, BLAH, a system which explains its reasoning, Artificial Intelligence 15 (1–2) (1980) 19–48, http://dx.doi.org/10.1016/0004-3702(80)90021-1.

[97] Douglas Walton, A dialogue system specification for explanation, Synthese 182 (3) (2011) 349–374.

[98] Alison Cawsey, Generating interactive explanations, in: Proceedings of the 9th National Conference on Artificial Intelligence, vol. 1, Citeseer, Anaheim, California, USA, 1991, pp. 86–91.

[99] Alison Cawsey, Planning interactive explanations, Int. J. Man-Mach. Stud. 38 (2) (1993) 169–199, http://dx.doi.org/10.1006/imms.1993.1009.

[100] Alison Cawsey, User modelling in interactive explanations, User Model. User-Adapt. Interact. 3 (3) (1993) 221–247, http://dx.doi.org/10.1007/bf01257890.

[101] Martha E. Pollack, Julia Hirschberg, Bonnie Webber, User participation in the reasoning processes of expert systems, in: Proceedings of Second National Conference Artificial Intelligence, MIT Press, Cambridge, Massachusetts, Menlo Park, California, USA, 1982, pp. 358–361.

[102] Hilary Johnson, Peter Johnson, Explanation facilities and interactive systems, in: Proceedings of the 1st International Conference on Intelligent User Interfaces, ACM, Orlando, Florida, USA, 1993, pp. 159–166, http://dx.doi.org/10.1145/169891.169951.

[103] Johanna D. Moore, Cecile L. Paris, Planning text for advisory dialogues, in: Proceedings of the 27th Annual Meeting on Association for Computational Linguistics, Association for Computational Linguistics, Vancouver, British Columbia, Canada, 1989, pp. 203–211, http://dx.doi.org/10.3115/981623.981648.

[104] Johanna D. Moore, William R. Swartout, A reactive approach to explanation, in: IJCAI, International Joint Conferences on Artificial Intelligence, Inc, Detroit, Michigan, USA, 1989, pp. 1504–1510.

[105] Johanna D. Moore, William R. Swartout, A reactive approach to explanation: Taking the user's feedback into account, in: Natural Language Generation in Artificial Intelligence and Computational Linguistics, Springer, USA, 1991, pp. 3–48.

[106] Mark G. Core, H. Chad Lane, Michael Van Lent, Dave Gomboc, Steve Solomon, Milton Rosenberg, Building explainable artificial intelligence systems, in: Proceedings of the 21st National Conference on Artificial Intelligence, AAAI Press, Boston, Massachusetts, USA, 2006, pp. 1766–1773, http://dx.doi.org/10.21236/ada459166.

[107] Dave Gomboc, Steve Solomon, Mark G. Core, H. Chad Lane, Michael Van Lent, Design recommendations to support automated explanation and tutoring, in: Proceedings of the Fourteenth Conference on Behavior Representation in Modelling and Simulation, Simulation Interoperability Standards Organization (SISO), Universal City, California, USA, 2005, pp. 331–340.

[108] H. Chad Lane, Mark G. Core, Michael Van Lent, Steve Solomon, Dave Gomboc, Explainable artificial intelligence for training and tutoring, in: Proceedings of the 12th International Conference on Artificial Intelligence in Education, AIED, IOS Press, Amsterdam, The Netherlands, 2005, pp. 762–764.

[109] Michael Van Lent, William Fisher, Michael Mancuso, An explainable artificial intelligence system for small-unit tactical behavior, in: Proceedings of the National Conference on Artificial Intelligence, Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, San Jose, California, USA, 2004, pp. 900–907.

[110] Arthur C. Graesser, Patrick Chipman, Brian C. Haynes, Andrew Olney, Autotutor: An intelligent tutoring system with mixed-initiative dialogue, IEEE Trans. Educ. 48 (4) (2005) 612–618, http://dx.doi.org/10.1109/te.2005.856149.

[111] Pat Langley, Ben Meadows, Mohan Sridharan, Dongkyu Choi, Explainable agency for intelligent autonomous systems, in: Proceedings of the Thirty-First Conference on Artificial Intelligence, AAAI Press, San Francisco, California, USA, 2017, pp. 4762–4764.

[112] Shirin Sohrabi, Jorge A. Baier, Sheila A. McIlraith, Preferred explanations: Theory and generation via planning, in: Proceedings of the Twenty-Fifth Conference on Artificial Intelligence, AAAI Press, San Francisco, California, USA, 2011, pp. 261–267.

[113] Natalia Díaz-Rodríguez, Galena Pisoni, Accessible cultural heritage through explainable artificial intelligence, in: Adjunct Publication of the 28th ACM Conference on User Modeling, Adaptation and Personalization, 2020, pp. 317–324.

[114] Galena Pisoni, Natalia Díaz-Rodríguez, Hannie Gijlers, Linda Tonolli, Human-centred artificial intelligence for designing accessible cultural heritage, Appl. Sci. 11 (2) (2021) 870.

[115] Maria Jose Gacto, Rafael Alcalá, Francisco Herrera, Interpretability of linguistic fuzzy rule-based systems: An overview of interpretability measures, Inform. Sci. 181 (20) (2011) 4340–4360, http://dx.doi.org/10.1016/j.ins.2011.02.021.

[116] Salvador García, Alberto Fernández, Julián Luengo, Francisco Herrera, A study of statistical techniques and performance measures for genetics-based machine learning: Accuracy and interpretability, Soft Comput. 13 (10) (2009) 959.

[117] Fernando E.B. Otero, Alex A. Freitas, Improving the interpretability of classification rules discovered by an ant colony algorithm: Extended results, Evol. Comput. 24 (3) (2016) 385–409, http://dx.doi.org/10.1162/evco_a_00155.

[118] Marko Robnik-Šikonja, Igor Kononenko, Explaining classifications for individual instances, IEEE Trans. Knowl. Data Eng. 20 (5) (2008) 589–600, http://dx.doi.org/10.1109/tkde.2007.190734.

[119] Marko Robnik-Šikonja, Explanation of prediction models with explain prediction, Informatica 42 (1) (2018) 13–22.

[120] Marko Bohanec, Marko Robnik-Šikonja, Mirjana Kljajić Borštnar, Decision-making framework with double-loop learning through interpretable black-box machine learning models, Ind. Manag. Data Syst. 117 (7) (2017) 1389–1406, http://dx.doi.org/10.1108/imds-09-2016-0409.

[121] Marko Bohanec, Mirjana Kljajić Borštnar, Marko Robnik-Šikonja, Explaining machine learning models in sales predictions, Expert Syst. Appl. 71 (2017) 416–428, http://dx.doi.org/10.1016/j.eswa.2016.11.010.

[122] Quan-shi Zhang, Song-Chun Zhu, Visual interpretability for deep learning: A survey, Front. Inf. Technol. Electron. Eng. 19 (1) (2018) 27–39.

[123] Chih-Kuan Yeh, Cheng-Yu Hsieh, Arun Suggala, David I. Inouye, Pradeep K. Ravikumar, On the (in)fidelity and sensitivity of explanations, in: Advances in Neural Information Processing Systems, Vancouver, BC, Canada, pp. 10965–10976, 2019.

[124] Shane Barratt, InterpNET: Neural introspection for interpretable deep learning, in: NIPS Symposium on Interpretable Machine Learning, NIPS, Long Beach, California, USA, 2017, pp. 47–53.

[125] Alexey Ignatiev, Towards trustable explainable AI, in: Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI, Yokohama, Japan, 2020, pp. 5154–5158, http://dx.doi.org/10.24963/ijcai.2020/726.

[126] Thibault Laugel, Marie-Jeanne Lesot, Christophe Marsala, Xavier Renard, Marcin Detyniecki, The dangers of post-hoc interpretability: Unjustified counterfactual explanations, in: Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI, International Joint Conferences on Artificial Intelligence Organization, Macao, China, 2019, pp. 2801–2807, http://dx.doi.org/10.24963/ijcai.2019/388.

[127] Julius Adebayo, Justin Gilmer, Ian Goodfellow, Been Kim, Local explanation methods for deep neural networks lack sensitivity to parameter values, in: 6th International Conference on Learning Representations, ICLR, Vancouver, Canada, 2018.

[128] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, Been Kim, Sanity checks for saliency maps, in: Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS, Neural Information Processing Systems Foundation, Inc., Montréal, Canada, 2018, pp. 9505–9515.

[129] Marco Ancona, Enea Ceolini, Cengiz Oztireli, Markus Gross, Towards better understanding of gradient-based attribution methods for deep neural networks, in: 6th International Conference on Learning Representations, ICLR, Vancouver, Canada, 2018.

[130] Leila Arras, Franziska Horn, Grégoire Montavon, Klaus-Robert Müller, Wojciech Samek, Explaining predictions of non-linear classifiers in NLP, in: Proceedings of the 1st Workshop on Representation Learning for NLP, Association for Computational Linguistics, Berlin, Germany, 2016, pp. 1–7, http://dx.doi.org/10.18653/v1/w16-1601.

[131] Alexander Binder, Wojciech Samek, Grégoire Montavon, Sebastian Bach, Klaus-Robert Müller, Analyzing and validating neural networks predictions, in: Proceedings of the ICML Workshop on Visualization for Deep Learning, ICML, New York City, New York, USA, 2016, pp. 118–121.

[132] Amirata Ghorbani, Abubakar Abid, James Zou, Interpretation of neural networks is fragile, in: NIPS Workshop on Machine Deception, NIPS, Long Beach, California, USA, 2017.

[133] Thu Trang Nguyen, Thach Le Nguyen, Georgiana Ifrim, A model-agnostic approach to quantifying the informativeness of explanation methods for time series classification, in: International Workshop on Advanced Analytics and Learning on Temporal Data, Springer, 2020, pp. 77–94, [online], http://dx.doi.org/10.1007/978-3-030-65742-0_6.

[134] Wojciech Samek, Thomas Wiegand, Klaus-Robert Müller, Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models, ITU J.: ICT Discov. 1 (2017) 1–10.

[135] Wojciech Samek, Alexander Binder, Grégoire Montavon, Sebastian Lapuschkin, Klaus-Robert Müller, Evaluating the visualization of what a deep neural network has learned, IEEE Trans. Neural Netw. Learn. Syst. 28 (11) (2017) 2660–2673, http://dx.doi.org/10.1109/tnnls.2016.2599820.

[136] Dumitru Erhan, Yoshua Bengio, Aaron Courville, Pascal Vincent, Visualizing higher-layer features of a deep network, Univ. Montr. 1341 (3) (2009) 1.

[137] Muriel Gevrey, Ioannis Dimopoulos, Sovan Lek, Review and comparison of methods to study the contribution of variables in artificial neural network models, Ecol. Model. 160 (3) (2003) 249–264, http://dx.doi.org/10.1016/s0304-3800(02)00257-0.

[138] Leila Arras, Franziska Horn, Grégoire Montavon, Klaus-Robert Müller, Wojciech Samek, "What is relevant in a text document?": An interpretable machine learning approach, PLoS One 12 (8) (2017) e0181142, http://dx.doi.org/10.1371/journal.pone.0181142.

[139] Grégoire Montavon, Sebastian Lapuschkin, Alexander Binder, Wojciech Samek, Klaus-Robert Müller, Explaining nonlinear classification decisions with deep taylor decomposition, Pattern Recognit. 65 (May) (2017) 211–222.

[140] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, Dhruv Batra, Grad-CAM: Visual explanations from deep networks via gradient-based localization, in: Proceedings of the IEEE International Conference on Computer Vision, IEEE, Venice, Italy, 2017, pp. 618–626, http://dx.doi.org/10.1109/ICCV.2017.74.

[141] Yash Goyal, Akrit Mohapatra, Devi Parikh, Dhruv Batra, Towards transparent AI systems: Interpreting visual question answering models, in: ICML Workshop on Visualization for Deep Learning, ICML, New York City, New York, USA, 2016.

[142] Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin, Why should I trust you?: Explaining the predictions of any classifier, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, San Francisco, CA, USA, 2016, pp. 1135–1144, http://dx.doi.org/10.1145/2939672.2939778.

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

[144] David Baehrens, Timon Schroeter, Stefan Harmeling, Motoaki Kawanabe, Katja Hansen, Klaus-Robert Mäžller, How to explain individual classification decisions, J. Mach. Learn. Res. 11 (Jun) (2010) 1803–1831.

[145] Pieter-Jan Kindermans, Kristof T. Schütt, Maximilian Alber, Klaus-Robert Müller, Dumitru Erhan, Been Kim, Sven Dähne, Learning how to explain neural networks: PatternNet and PatternAttribution, in: 6th International Conference on Learning Representations, ICLR, Vancouver, Canada, 2018.

[146] Scott M. Lundberg, Su-In Lee, A unified approach to interpreting model predictions, in: Advances in Neural Information Processing Systems, Neural Information Processing Systems Foundation, Inc., Long Beach, California, USA, 2017, pp. 4765–4774.

[147] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, Martin Wattenberg, Smoothgrad: Removing noise by adding noise, in: International Conference on Machine Learning 2017 - Workshop on Visualization for Deep Learning, ICML, Sydney, Australia, 2017, pp. 15–24.

[148] Lindsay Sanneman, Julie A. Shah, A situation awareness-based framework for design and evaluation of explainable AI, in: International Workshop on Explainable, Transparent Autonomous Agents and Multi-Agent Systems, Springer, Auckland, New Zealand, 2020, pp. 94–110, http://dx.doi.org/10.1007/978-3-030-51924-7_6.

[149] Brian Y. Lim, Anind K. Dey, Assessing demand for intelligibility in context-aware applications, in: Proceedings of the 11th International Conference on Ubiquitous Computing, ACM, Orlando, Florida, USA, 2009, pp. 195–204, http://dx.doi.org/10.1145/1620545.1620576.

[150] Been Kim, Julie A. Shah, Finale Doshi-Velez, Mind the gap: A generative approach to interpretable feature selection and extraction, in: Advances in Neural Information Processing Systems, Neural Information Processing Systems Foundation, Inc., Montréal, Québec, Canada, 2015, pp. 2260–2268.

[151] Sam Hepenstal, David McNeish, Explainable artificial intelligence: What do you need to know? in: International Conference on Human-Computer Interaction, Springer, Copenhagen, Denmark, 2020, pp. 266–275, http://dx.doi.org/10.1007/978-3-030-50353-6_20.

[152] Henri J. Suermondt, Gregory F. Cooper, An evaluation of explanations of probabilistic inference, Comput. Biomed. Res. 26 (3) (1993) 242–254, http://dx.doi.org/10.1006/cbmr.1993.1017.

[153] L. Richard Ye, Paul E. Johnson, The impact of explanation facilities on user acceptance of expert systems advice, MIS Q. 19 (2) (1995) 157–172, http://dx.doi.org/10.2307/249686.

[154] Vanessa Putnam, Cristina Conati, Exploring the need for explainable artificial intelligence (XAI) in intelligent tutoring systems (ITS), in: Joint Proceedings of the ACM IUI 2019 Workshops Co-Located with the 24th ACM Conference on Intelligent User Interfaces, ACM IUI, vol. 2327, CEUR-WS.org, Los Angeles, California, USA, 2019.

[155] Joe Tullio, Anind K. Dey, Jason Chalecki, James Fogarty, How it works: A field study of non-technical users interacting with an intelligent system, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM, San Jose, California, USA, 2007, pp. 31–40, http://dx.doi.org/10.1145/1240624.1240630.

[156] Harmanpreet Kaur, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna Wallach, Jennifer Wortman Vaughan, Interpreting interpretability: Understanding data scientists' use of interpretability tools for machine learning, in: Proceedings of the CHI Conference on Human Factors in Computing Systems, Honolulu, HI, USA, 2020, pp. 1–14, http://dx.doi.org/10.1145/3313831.3376219.

[157] Andreas Holzinger, Michael Kickmeier-Rust, Heimo Müller, KANDINSKY patterns as IQ-test for machine learning, in: International Cross-Domain Conference for Machine Learning and Knowledge Extraction, Springer, Canterbury, United Kingdom, 2019, pp. 1–14, http://dx.doi.org/10.1007/978-3-030-29726-8_1.

[158] Sebastian Lapuschkin, Alexander Binder, Grégoire Montavon, Klaus-Robert Müller, Wojciech Samek, Analyzing classifiers: Fisher vectors and deep neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE, Las Vegas, Nevada, USA, 2016, pp. 2912–2920, http://dx.doi.org/10.1109/cvpr.2016.318.

[159] Avleen Malhi, Samanta Knapic, Kary Främling, Explainable agents for less bias in human-agent decision making, in: International Workshop on Explainable, Transparent Autonomous Agents and Multi-Agent Systems, Springer, Auckland, New Zealand, 2020, pp. 129–146, http://dx.doi.org/10.1007/978-3-030-51924-7_8.

[160] Vignesh Srinivasan, Sebastian Lapuschkin, Cornelius Hellge, Klaus-Robert Müller, Wojciech Samek, Interpretable human action recognition in compressed domain, in: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, IEEE, New Orleans, Louisiana, USA, 2017, pp. 1692–1696, http://dx.doi.org/10.1109/ICASSP.2017.7952445.

[161] Roy Assaf, Anika Schumann, Explainable deep neural networks for multivariate time series predictions, in: Proceedings of the 28th International Joint Conference on Artificial Intelligence, International Joint Conferences on Artificial Intelligence Organization, Macao, China, 2019, pp. 6488–6490, http://dx.doi.org/10.24963/ijcai.2019/932.

[162] Yanzhuo Ding, Yang Liu, Huanbo Luan, Maosong Sun, Visualizing and understanding neural machine translation, in: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), vol. 1, Association for Computational Linguistics, Vancouver, Canada, 2017, pp. 1150–1159, http://dx.doi.org/10.18653/v1/p17-1106.

[163] Irene Sturm, Sebastian Lapuschkin, Wojciech Samek, Klaus-Robert Müller, Interpretable deep neural networks for single-trial EEG classification, J. Neurosci. Methods 274 (2016) 141–145, http://dx.doi.org/10.1016/j.jneumeth.2016.10.008.

[164] Katharina Weitz, Dominik Schiller, Ruben Schlagowski, Tobias Huber, Elisabeth André, "Let me explain!": Exploring the potential of virtual agents in explainable AI interaction design, J. Multimodal User Interfaces (2020) 1–12, http://dx.doi.org/10.1007/s12193-020-00332-0.

[165] Been Kim, Cynthia Rudin, Julie A. Shah, The Bayesian case model: A generative approach for case-based reasoning and prototype classification, in: Advances in Neural Information Processing Systems, Neural Information Processing Systems Foundation, Inc., Montréal, Québec, Canada, 2014, pp. 1952–1960.

[166] Pierre Stock, Moustapha Cisse, Convnets and imagenet beyond accuracy: Understanding mistakes and uncovering biases, in: Proceedings of the European Conference on Computer Vision, ECCV, Springer, Munich, Germany, 2018, pp. 498–512, http://dx.doi.org/10.1007/978-3-030-01231-1_31.

[167] David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, Antonio Torralba, Network dissection: Quantifying interpretability of deep visual representations, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE, Honolulu, Hawaii, USA, 2017, pp. 6541–6549, http://dx.doi.org/10.1109/cvpr.2017.354.

[168] Mitja Luštrek, Matjaž Gams, Sanda Martinčić-Ipšić, et al., Comprehensibility of classification trees–survey design validation, in: 6th International Conference on Information Technologies and Information Society-ITIS2014, ITIS, Šmarješke toplice, Slovenia, 2014, pp. 46–61.

[169] Katja Hansen, David Baehrens, Timon Schroeter, Matthias Rupp, Klaus-Robert Müller, Visual interpretation of kernel-based prediction models, Mol. Inform. 30 (9) (2011) 817–826, http://dx.doi.org/10.1002/minf.201100059.

[170] Vincent A.W.M.M. Aleven, Kenneth R. Koedinger, An effective metacognitive strategy: Learning by doing and explaining with a computer-based cognitive tutor, Cogn. Sci. 26 (2) (2002) 147–179, http://dx.doi.org/10.1016/s0364-0213(02)00061-7.

[171] Maaike Harbers, Joost Broekens, Karel Van Den Bosch, John-Jules Meyer, Guidelines for developing explainable cognitive models, in: Proceedings of ICCM, Citeseer, Berlin, Germany, 2010, pp. 85–90.

[172] Maaike Harbers, Karel van den Bosch, John-Jules Meyer, Design and evaluation of explainable BDI agents, in: 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, vol. 2, IEEE, Toronto, Canada, 2010, pp. 125–132, http://dx.doi.org/10.1109/wi-iat.2010.115.

[173] Isaac Lage, Andrew Ross, Samuel J. Gershman, Been Kim, Finale Doshi-Velez, Human-in-the-loop interpretability prior, in: Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS, Neural Information Processing Systems Foundation, Inc., Montréal, Canada, 2018, pp. 10180–10189.

[174] Forough Poursabzi-Sangdeh, Daniel G. Goldstein, Jake M. Hofman, Jennifer Wortman Vaughan, Hanna Wallach, Manipulating and measuring model interpretability, in: NIPS Women in Machine Learning Workshop, NIPS, Long Beach, California, USA, 2017.

[175] Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin, Anchors: High-precision model-agnostic explanations, in: Thirty-Second AAAI Conference on Artificial Intelligence, AAAI Press, New Orleans, Louisiana, USA, 2018, pp. 1527–1535.

[176] Holzinger Andreas, Carrington André, Müller Heimo, Measuring the quality of explanations: The system causability scale (SCS): Comparing human and machine explanations, KI-Künstliche Intell. 34 (2) (2020) 193–198, http://dx.doi.org/10.1007/s13218-020-00636-z.

[177] Thilo Spinner, Udo Schlegel, Hanna Schäfer, Mennatallah El-Assady, Explainer: A visual analytics framework for interactive and explainable machine learning, IEEE Trans. Vis. Comput. Graph. 26 (2019) 1064–1074, http://dx.doi.org/10.1109/TVCG.2019.2934629.

[178] Todd Kulesza, Simone Stumpf, Weng-Keen Wong, Margaret M Burnett, Stephen Perona, Andrew Ko, Ian Oberst, Why-oriented end-user debugging of naive Bayes text classification, ACM Trans. Interact. Intell. Syst. (TiiS) 1 (1) (2011) 2:1–2:31, http://dx.doi.org/10.1145/2030365.2030367.

[179] Hiva Allahyari, Niklas Lavesson, User-oriented assessment of classification model understandability, in: 11th Scandinavian Conference on Artificial Intelligence, IOS Press, Trondheim, Norway, 2011, pp. 11–19, http://dx.doi.org/10.3233/978-1-60750-754-3-11.

[180] Johan Huysmans, Karel Dejaeger, Christophe Mues, Jan Vanthienen, Bart Baesens, An empirical evaluation of the comprehensibility of decision table, tree and rule based predictive models, Decis. Support Syst. 51 (1) (2011) 141–154, http://dx.doi.org/10.1016/j.dss.2010.12.003.

[181] Mauro Dragoni, Ivan Donadello, Claudio Eccher, Explainable AI meets persuasiveness: Translating reasoning results into behavioral change advice, Artif. Intell. Med. (2020) 101840, http://dx.doi.org/10.1016/j.artmed.2020.101840.

[182] William F. Lawless, Ranjeev Mittu, Donald Sofge, Laura Hiatt, Artificial intelligence, autonomy, and human-machine teams: Interdependence, context, and explainable AI, AI Mag. 40 (3) (2019) 5–13.

[183] Danding Wang, Qian Yang, Ashraf Abdul, Brian Y. Lim, Designing theory-driven user-centric explainable AI, in: Proceedings of the CHI Conference on Human Factors in Computing Systems, Glasgow, Scotland, UK, 2019, pp. 1–15, http://dx.doi.org/10.1145/3290605.3300831.

[184] Adrien Bennetot, Jean-Luc Laurent, Raja Chatila, Natalia Díaz-Rodríguez, Towards explainable neural-symbolic visual reasoning, in: NeSy Workshop IJCAI, International Joint Conferences on Artificial Intelligence, Inc, Macao, China, 2019, pp. 71–75.

[185] Hadrien Bride, Jie Dong, Jin Song Dong, Zhé Hóu, Towards dependable and explainable machine learning using automated reasoning, in: International Conference on Formal Engineering Methods, Springer, Gold Coast, Australia, 2018, pp. 412–416, http://dx.doi.org/10.1007/978-3-030-02450-5_25.

[186] Lucas Rizzo, Luca Longo, A qualitative investigation of the explainability of defeasible argumentation and non-monotonic fuzzy reasoning, in: Proceedings for the 26th AIAI Irish Conference on Artificial Intelligence and Cognitive Science Trinity College Dublin, CEUR-WS.org, Dublin, Ireland, 2018, pp. 138–149.

[187] Lucas Rizzo, Luca Longo, Inferential models of mental workload with defeasible argumentation and non-monotonic fuzzy reasoning: A comparative study, in: 2nd Workshop on Advances in Argumentation in Artificial Intelligence, CEUR-WS.org, Trento, Italy, 2018, pp. 11–26.

[188] Zhiwei Zeng, Chunyan Miao, Cyril Leung, Jing Jih Chin, Building more explainable artificial intelligence with argumentation, in: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th Innovative Applications of Artificial Intelligence (IAAI-18), and the 8th Symposium on Educational Advances in Artificial Intelligence (EAAI-18), AAAI Press, New Orleans, Louisiana, USA, 2018, pp. 8044–8046.

[189] Artur d'Avila Garcez, Tarek R. Besold, Luc De Raedt, Peter Földiak, Pascal Hitzler, Thomas Icard, Kai-Uwe Kühnberger, Luis C. Lamb, Risto Miikkulainen, Daniel L. Silver, Neural-symbolic learning and reasoning: Contributions and challenges, in: AAAI Spring Symposium Series, AAAI Press, Palo Alto, California, USA, 2015, pp. 20–23.