Towards a terminology for a fully contextualized XAI

Matthieu Bellucci, Nicolas Delestre, Nicolas Malandain, Cecilia Zanni-Merk

Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 00 (2021) 000–000

Procedia Computer Science

www.elsevier.com/locate/procedia

25th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES 2021)

Towards a terminology for a fully contextualized XAI

Matthieu Bellucci*, Nicolas Delestre, Nicolas Malandain, Cecilia Zanni-Merk

Normandie Université, INSA Rouen, LITIS, Rouen 76000, France

Abstract

Explainable Artificial Intelligence (XAI) has seen a surge in popularity in the past few years, thanks to new legislations that promote the "right to explanation". Many popular methods have been developed recently to help understand black-box models, but it is not clear yet how an explanation is defined. Furthermore, the community agrees to say that many important terms do not have commonly accepted definitions. In this paper, we review the literature and show that there is a major issue concerning the definitions of terms such as explainability or interpretability. There is a lack of consensus that slows the development of this field. To address this problem, we propose a terminology that takes into account the context of an AI system, i.e., its users, purposes or design. This terminology is compatible with the majority of the definitions encountered in the literature so that it can be a foundation for future works.

Keywords: explainability; interpretability; terminology

1. Introduction

Artificial Intelligence (AI) is widely used in the industry and our daily lives. Until recently, its use was not problematic since it was limited to minor decisions such as film recommendations or marketing purposes. However, AI has made its entrance in sectors such as justice, banking or medicine for instance. In these sectors, it is vital to understand why and how an AI made a certain decision. It is natural to wonder why a bank loan was refused, or to understand how a certain diagnosis was made for a patient. It is this right to explanation that the GDPR [1], among others, defends. The use of complex and high-dimensional AI algorithms makes the design of explanations difficult or even impossible. It is this need for explainability that has led to the rise in popularity of the field of eXplainable AI (XAI), with projects such as DARPA's XAI program [2]. Figure 1 presents the most important explainable AI use-case according to these authors: "when in practice, within some context, a final user must understand, trust, and be responsible for the conclusions an AI system draws". The explainable model and the explanation interface of this architecture represent

* Corresponding author.
E-mail address: matthieu.bellucci@insa-rouen.fr

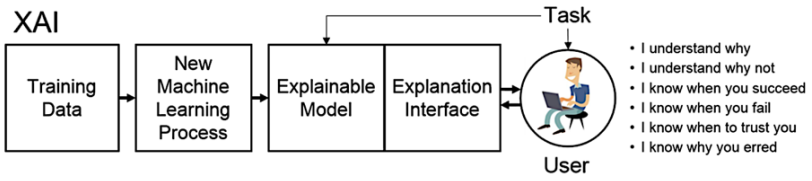

Fig. 1. A typical XAI use-case according to [2]

what we call in the following an explainable intelligent system (XIS). An explainable model associates its outputs to a set of rules (or other formalism) to ground the decisions being made. An explanation interface interacts with a user to provide explanations, based on the inputs, outputs and characteristics of the explainable model.

Before this field of research was called XAI, it was more commonly referred to as Interpretable AI, which suggests a strong connection between those two terms. Nowadays, according to the literature, the terms explainable, interpretable and many other terms have different meanings depending on the authors [3, 4]. There is no consensus yet on the meaning of the terms used, which poses a problem in assessing the quality and usefulness of the methods developed. Many researchers have addressed the issue of the lack of an agreed terminology, outlining the problems this creates [4, 5, 6]. However, to our knowledge, there is no article proposing a terminology that takes up the popular terms of this field and assigns unambiguous definitions to them.

This is what we propose in this article. In Section 2, we will examine the terms found in the literature and the definitions associated with them. Then, in Section 3, we propose a coherent and unambiguous terminology based on the definitions seen in the previous section, whose usefulness is illustrated in a case study. Finally, we conclude this paper in Section 4.

2. Literature review

We have surveyed the literature to determine and understand the terminology of XAI. We noticed many reoccurring terms, but without a clear consensus concerning their definitions [6, 7, 8]. Among these terms, explainability and interpretability are the most important and also the most controversial when it comes to their definition. For some terms, the XAI community has reached a consensus on their meaning, but without providing a definition. This lack of a clear terminology is an issue for the XAI research. Indeed, if researchers do not agree on the terminology, it makes the domain less accessible to newcomers, papers harder to understand and methods harder to use and evaluate.

We have studied more than 30 papers related to XAI. We used the keywords “XAI”, “terminology”, “taxonomy”, “survey”, “review”, “explainability”, “interpretability” in Google Scholar accessed between October 2020 and January 2021; especially looking for surveys, taxonomies and reviews, as well as popular papers that propose specific methods of explainability. Our objective was to identify the terms and associated definitions that come up in these surveys. We believe that combining these surveys of the literature gives a good overview of the terminology in XAI. The rest of the papers helped us refine our definitions in order to make them compatible with most existing methods.

In this section, we will study how the reoccurring terms are used and defined in different papers and identify points of convergence and divergence for each of them. We will also determine how terms are related to each other.

2.1. Interpretability

Interpretability is the preferred term of the scientific community to describe AI techniques that are easily understandable [6]. This term is used to describe models or algorithms rather than entire intelligent systems. The implicit definition that some papers use is that an interpretable model is a model that is easy to understand for anyone [3]. In Lipton’s paper [5] dedicated to understanding what interpretability is, he concludes that it is meaningless to qualify any model as intrinsically interpretable. Adadi et al. [6] propose the following definition: “An interpretable system is a system where a user cannot only see but also study and understand how inputs are mathematically mapped to outputs”. This definition fits well with the implicit definition we described above. Gilpin et al. [9] propose a similar

definition. Lipton [5] and Futia et al. [10] introduce the notions of simulatability and decomposability. These notions can be considered as subsets of interpretability. This is coherent with Adadi's definition: if we can decompose or simulate a model, that means that we can understand how the inputs are related to the outputs.

Another component that appears in many definitions is the cognitive effort required to understand a model. For example, Calegari et al. [8] declare "In AI algorithms, interpretability refers to the cognitive effort required by human observers to assign a meaning to the way the algorithm works, or motivate the outcome it produces". This notion of cognitive effort is also retrieved in other papers, Ribeiro et al. [11] mention that the user's limitations must be taken into account, Gilpin et al. [9] add that "the success of this goal is tied to the cognition, knowledge, and biases of the user". This cognitive effort can be measured by the complexity of a model. This notion of complexity is widely used in the literature. It can be defined as a measure of the interpretability of one model [6, 11, 12]. Examples of complexity measures for a variety of models can be found in different papers [9, 10, 11, 12, 13].

Many authors give a very broad definition of interpretability. "The ability to explain or present in understandable terms to a human" is a definition proposed in some papers [4, 7, 14]. We believe that the use of the term *explain* in this definition makes it confusing with the notion of explanation.

We can conclude that interpretability refers to the ability of an object to be understood and studied by a user, with a reasonable cognitive effort. We use the term object rather than model, because as Lipton [5] mentioned, the input must be understandable as well as the model. Therefore, we could also define an input as interpretable. Interpretability also encompasses different notions such as decomposability and simulatability.

2.2. Explainability and explanations

Some scholars use explainability and interpretability as synonyms, but do not give a clear definition. Beaudouin et al. [14] propose the following definition of explainability : "the ability, inclination or suitability to make plain or comprehensible, or explain the meaning of, an algorithm". In other words, explainability is the ability of a system to generate explanations. As discussed in a series of four essays by Hoffman, Klein et al. [15, 16, 17, 18], it is a complex task to define what an explanation is. Explaining something is equivalent to answering questions that a user[1] may ask, to be able to understand what it is observing. The aim is to provide the relevant information so that it can reason on its own about how a model works, or why a model made a specific decision. The notion of interaction with the user is discussed in some papers [10, 18]. Calegari et al. [8] define an explanation as "an activity aimed at making the relevant details of an object clear or easy to understand to some observer". Arrieta et al. [4] and Guidotti et al. [12] share the same definition, "an explanation is an interface between humans and a decision maker that is, at the same time, both an accurate proxy of the decision maker and comprehensible to humans". This definition adds a dimension to what an explanation should be. It should be an accurate proxy, i.e., the explanation must be based on the model's mechanisms and the inputs used. Lipton [5] warns that some ways of generating explanations proposed in the literature do not guarantee to be accurate proxies. Ribeiro et al. [11] define an explanation as presenting "textual or visual artefacts that provide qualitative understanding of the relationship between the instance's components and the model's prediction". We observe that this definition is a subset of the previous ones. Indeed, textual or visual artefacts can be the relevant details we were previously discussing. We consider that their definition is adapted to their own method, but lacks generality. Finally, Tiddi et al. [19] propose an ontology design pattern for explanations, showing that an explanation is generated by an agent, using a theory, to explain an event based on a prior event. This design pattern does not explicitly take into account the user to generate an appropriate explanation. To conclude, from these definitions, an explanation is an interaction between the user and another agent whose goal is to provide details to answer the user's questions or make a system or decision easy to understand.

The XAI community appears to agree on a taxonomy of methods to generate explanations. Arya et al. [3] propose a taxonomy based on questions about what is explained, how is it explained and at what level. This taxonomy takes the form of a decision tree. They also introduce the idea that an explanation can either be static or interactive, which is rarely discussed in the literature. We have encountered the notions of global/local explainability and post hoc/direct explanations in a majority of the papers we reviewed.

[1] A user can be a human but could also be a machine. We will therefore use the pronoun "it" to refer to a user.

- Post hoc/Direct explanations: According to Futia et al. [10], post hoc explanations "do not seek to reveal how a model works, but they are focused on how it behaved and why". Lipton [5] says that "these interpretations might explain predictions without elucidating the mechanisms by which models work". Guidotti et al. [12] qualify them of "reverse-engineering approaches". We understand from these definitions that post hoc explanations do not exploit the logic of the system. Arya et al. [3] say that post hoc explanations involve "auxiliary methods to explain a model after it is trained". Among existing methods, we can cite LIME [11] or SHAP [20]. They use the inputs of the system to better understand how they are mapped to the outputs. Post hoc explanation methods are usually model-agnostic, thus they can be applied to any model, whether it is interpretable or not.
  Some scholars discuss that explanation methods opposed to post hoc explanations use the directly interpretable nature of the system they want to explain [3, 4, 8]. This type of explanation is rarely explicitly defined in the literature and thus it is not associated to a specific term. Therefore, we propose the term direct explanations to refer to these methods. Calegari et al. [8] call it explainability by design, "methods in this category aim at creating interpretable or explainable intelligent systems by construction". Adadi et al. name these methods "complexity related methods", which corresponds to designing models that are intrinsically interpretable. This notion of using the intrinsic interpretability of a model is also used in [3]. These direct explanations are obviously model-dependent, because the model is used as a source to generate explanations.
- Global/local explainability: Global explainability refers to the explanation of how a model works. Adadi et al. [6] and Guidotti et al. [12] say that global explainability "facilitates the understanding of the whole logic of a model". Similar definitions are found in [3, 14, 21]. Hoffman et al. [18] describe it as the explanation of "how the conceptual categories and mechanisms are derivable from instances and their attributes". Doshi-Velez et al. [7] share a somewhat identical definition, it "implies knowing what patterns are present in general". Local explainability refers to explaining a single prediction made by an AI. This idea is shared by many scholars [3, 6, 7, 10]. Local explanations are usually focused on analysis of feature importance in learning algorithms (such as LIME [11]) and may use specific reasoning mechanisms to answer "what if" questions.

2.3. Transparency and black-box-ness

Transparency is a controversial term. For some authors, transparency is closely tied to interpretability. Guidotti et al. [12] discuss the transparent box problem which, according to them, "consists in directly providing a model that is locally or globally interpretable". Arya et al. [3] say that "a directly interpretable model is one that by its intrinsic transparent nature is understandable by most consumers". Lipton [5] in his survey, says that certain papers qualify understandable models as transparent, while incomprehensible models are called black-boxes. According to the definitions of interpretability seen in Section 2.1, an understandable model is interpretable. Therefore, the distinction between transparent and interpretable is unclear.

Beaudouin et al. [14] propose a different definition: "transparency generally refers to making information about the inner workings of the algorithm available for scrutiny, including how an AI system is developed, trained and deployed". They also add that "transparency does not necessarily mean that the underlying information is easily comprehensible". This definition makes the distinction between transparent and interpretable clearer. Indeed, an interpretable model is a model a user can understand, whereas a transparent model is a model that provides all the information about its development. Transparency may help to provide explanations, but it rather helps ensure its fairness and unbiasedness.

Futia et al. [10] and Lipton [5] classify three levels of transparency. Transparency considered at the level of the entire model is called simulatability, at the level of individual components such as parameters, called decomposability and at the level of the training algorithm, called algorithmic transparency. We point out that these notions are also linked to interpretability, as discussed in Section 2.1. This is another evidence of the ambiguity between transparency and interpretability.

- Simulatability: Futia et al. [10] define simulatability as "checking through a heuristic approach whether a human reaches the mechanistic understanding of how the model functions, and consequently if the human is able to simulate the decision process". Lipton [5] proposes a similar definition, which is also used by Wu et al. [13]: "a human should be able to take in input data together with the parameters of the model and in reasonable time,

step through every calculation required to produce a prediction". Arrieta et al. [4] give a similar definition and add that the complexity of the model is very important for simulatability. Indeed, the more complex a model is, the harder it is to simulate. We can also deduce that the simulatability of a model depends on the cognitive capacity of the user, in the same fashion as for interpretability.

- Decomposability: Futia et al. [10], Arrieta et al. [4] and Lipton [5] all agree on the definition of decomposability: "each part of the model - input, parameter, and calculation - admits an intuitive explanation". We deduce from this definition of decomposability that a model consists of three components, the input, the parameters and the calculations.

Black-box-ness is related to the opacity of the so-called "black-box" models according to the literature. Lipton [5] defines black-box models as "incomprehensible models". Calegari et al. [8] discusses the definition further: "used to refer to models where knowledge is not explicitly represented, but rather it is distributed among tensors of real numbers, whose complexity seldom fits our cognitive capabilities as humans". We find once again the notion of cognitive capabilities, which echoes the definition of interpretability.

2.4. Reliability and Confidence

Confidence is a term that we often find in the literature. To refer to approximately the same idea, some scholars use reliability. Confusions may arise when trust and confidence are defined in different ways within the same paper, because according to the Oxford dictionary, confidence is defined as "the feeling that you can trust, believe in, and be sure about the abilities or good qualities of someone or something" [22]. Trust is therefore used to define confidence. To circumvent this issue, we will use reliability.

Calegari et al. [8] define reliability as "the assurance that a model is providing the correct answer". Arrieta et al. [4] view it "as a generalization of robustness and stability". This idea is shared in [7, 12]. Robustness is the ability of a system to cope with errors during execution and cope with erroneous input [23]. Stability, or algorithmic stability, refers to how a model reacts to small perturbations to its inputs [24]. We conclude that for a model to be reliable, it needs to provide an output that is not sensitive to small perturbations.

3. Terminology proposition

Now that we have studied the current state of the terminology in the literature, we propose a terminology that uses these terms in an unambiguous way. The proposed definitions for each term are presented in *italics*. At the first appearance of a defined term of the terminology in this section, the term is presented in **bold**.

3.1. General terminology

In this section, we will define what constitutes an XIS, what is **explainability** and **interpretability** and introduce some notions that we have not discussed in Section 2.

User. In the literature, the person using an XIS is called a **user**, an observer or even a customer. We propose the following definition: *A user is an agent that interacts with an XIS*. This simple definition allows us to consider a program, or another agent as a user [25]. The designer of the system can be considered a user as well. As we have seen in the literature, explainability and interpretability need to consider the user in order to provide a useful explanation [6, 10, 11]. This important idea was kept in mind when designing the terminology. Many definitions will depend on the user and its task.

Explanations and Explainability. We consider that the Explanation Ontology Design Pattern from Tiddi et al. [19] is the background for the design of an explanation. We believe that this ontology lacks the concept of user, as the explanation depends on it, therefore it should be taken into account when designing an explanation. Based on this observation and the literature review, we propose the following definition: *an explanation is the result of an interaction between a user and an explainer in order to answer the user's questions*. We define explainability as *the ability of an XIS to be explained to a user or to provide an explanation.*

In Figure 1, the interaction is represented by the arrows between the explanation interface and the user. The explanation interface starts the interaction by providing at least the decision from the explainable model. The interface might also add a first explanation to the decision, in order to answer usual questions; or it might wait for the user to ask questions. The goal of an explanation is to answer the user's questions about the system and/or its decision. An explanation is considered valid when the user has no more questions. This notion of interaction is also what differentiates explainability and interpretability.

Interpretability. We define interpretability as *the ability of an object or an XIS to be seen, understood and studied by a user, with a reasonable cognitive effort*. It is therefore a property of an XIS or an object. This term denotes a property of a system, but considering its importance, we have decided to define it in this section. For an object to be interpretable to a user, the user must have access to enough information so that it can understand this object. This necessity links interpretability to **transparency**. Again, the interpretability of an object depends on the user's prior knowledge, a same model can be interpretable for one user, but not for another.

Trust. Gunning et al. [2] in DARPA's XAI program highlight that a final user must **trust** the conclusions an AI system draws. According to the Cambridge Dictionary [26], *to trust is to believe that something is safe and reliable*. If a user does not trust an XIS, it will never use it, rendering the system useless. For a user to trust an XIS, full information about its structure, function and behaviour should be provided. This is done through explanations, transparency, **reliability** and providing the failure modes or edge cases of the system.

3.2. Explanation terminology

The literature review showed that there are several types of explanations: **global** or **local** explanations, **direct** or **post hoc** explanations and **static** or **interactive** explanations. We characterise explanations with three properties that answer a question about the design of an explanation. The **Focus** answers the question "what do we explain?", **Means** answers the question "By what means do we explain?" and finally, **Modality** answers "how do we explain?". These properties help us understand what the objective of an explanation is. They also clarify which methods are comparable and why. Another property that Tiddi et al. [19] consider in their ontology design pattern is the **Reasoning**. The importance of reasoning to build an explanation is also outlined in this series of essays [15, 16, 17, 18]. Therefore, we also consider reasoning as a property of an explanation. We will now discuss in further details each property.

Focus. Focus can be declined into global and local explanations. *A global explanation describes the behaviour of the entire system.* The explanation may also include what training data was used, the performance of the system (evaluated by adequate metrics) and what its limitations or failure modes are.

A local explanation aims at explaining a single prediction. It answers "what", "why", "what if" questions. Counterfactual reasoning is especially relevant for this kind of explanations, as it consists in exploring "what if" questions.

Means. Means can be broken down into direct and post hoc explanations. *A direct explanation is an explanation issued directly from the AI algorithm at the core of the XIS.* For this explanation to be satisfactory, the system and its components should be interpretable.

A post hoc explanation uses an auxiliary method to explain an XIS or its decisions. This type of explanation is appropriate to explain **black-boxes**, although it can explain any type of system. These auxiliary methods are usually model-agnostic, which makes them adaptable to any XIS. Post hoc explanations are less flexible than direct explanations because to answer a particular question from the user, the probability that one auxiliary method can answer it alone is very low. Therefore, a post hoc explanation should use a combination of these auxiliary methods to be satisfactory to the user. This adds a layer of complexity for the designer of the explanation, which in turn, may make the explanation harder to understand for the user. These definitions already make consensus in the literature.

Modality. Modality indicates whether the explanation is static or interactive, as introduced in [3]. These terms are self-explanatory, *a static explanation does not change depending on the user*, whereas *an interactive explanation varies depending on the user.* Several methods to generate static explanations have been proposed in the literature. They give new information on the system, or make the available information easy to understand. Unfortunately, a static explanation cannot adapt to the user on its own.

An interactive explanation adapts to the level of expertise of the user which makes it more powerful. When a combination of static explanations is needed, a single interactive explanation can be sufficient. Unfortunately, at the time

of writing, no method of generating interactive explanations exists. This is due to the fact that making an interactive explanation is much more complex than making a static one, as it needs to interact with a user. This interaction requires the explanation interface to understand a request and to map requests to different explanations.

Reasoning. According to the Cambridge dictionary [27], *reasoning is the process of thinking about something in order to make a decision*. Hoffman et al. [15] define **abduction** and **counterfactual** reasoning, that humans use and are not deductive logic. These methods are the most used by humans to reason, that is why they should be used to generate explanations. Abduction is defined as *rendering what might be thought of as a unique experience into an instance of a more general phenomenon*. This is an intuitive way of reasoning, based on the prior knowledge and experience of the user. This method of reasoning can be a way of providing satisfying explanations without guaranteeing absolute truth. *Counterfactual reasoning consists in imagining how an event could change if a previous event that led to the observed event was modified or removed.* It answers the question "What would happen if...?" [18]. This is also useful to provide a satisfying explanation, as it allows the user to validate or refute its hypotheses on why a particular prediction was made. Counterfactual reasoning is often mentioned in the literature as being an important method to generate explanations [4, 6, 16].

3.3. System terminology

Many notions we have studied are bound to an XIS. In numerous papers, a model or a system are used as synonyms. But in our case, we want to encompass the full context of XAI, from end to end. That is to say, from the input to the user of the system. The model is the object that calculates the prediction from the input, which means it is part of this system. We defined interpretability as being a property of an object or system, therefore, any component of a system can be interpretable on its own, without guaranteeing that the entire system is interpretable. In Figure 1, a list of goals of the XIS is proposed. We divide these notions into two parts, the first part concerns the properties of an XIS that are essential to generate explanations. The second part is about concepts to increase trust in an XIS.

Recalling the DARPA's XAI program main statement evoked in Section 3.1, trust and responsibility are goals of explainability [8, 11], in order to create a *responsible AI* [14, 6]. Responsible AI is an AI that is accountable, i.e., has the ability to demonstrate and accept responsibility. Explainability is only a step towards a responsible AI, other notions such as transparency, traceability and auditability play a role to move towards a responsible AI. Among these notions, we define transparency and **fairness** as they are often seen in the XAI literature and not always clearly defined.

3.3.1. Towards more interpretable systems

The notions defined in this section are aimed at understanding the underlying concepts of interpretability and how to measure interpretability. In our literature review, we saw that some scholars define transparency as the opposite of **black-box-ness**. We argue that the actual opposite of black-box is interpretable, since a black-box is a model a user does not understand, whereas an interpretable model is a model understood by a user. The levels of transparency that we studied in the literature then become levels of interpretability.

Black-box-ness. We consider that *black-box-ness is the opposite of interpretability*. It can also be referred to as opacity. This definition is directly tied to the definition of interpretability, thus black-box-ness depends on the user. Some models can be judged as too complex for anyone to understand. Deep neural networks, for instance, are notoriously hard to understand and it is commonly agreed that they are black-boxes, regardless of the user.

Complexity. **Complexity** is *the measure of the interpretability of an object, i.e., the measure of how easy it is for a user to simulate and/or understand an object*. This measure is model-dependent. Many examples are given in the literature, for instance the complexity of a decision-tree could be the number of nodes of the tree, or the maximum length of path. The complexity of a linear regression could be the number of variables, and so on. More examples are given in [12].

Simulatability. **Simulatability** refers to *the ability of a model to be simulated or replicated by a user*. Measuring simulatability requires to define metrics to assess whether a user can simulate or not the system. Simulatability allows a user to perform many experiments and manipulations of the system on its own, therefore it does not have to ask an explanation interface to get answers. This also increases its trust in the model, as it knows how the system works. Some authors argue that if a model is simulatable, then the system using this model is interpretable. The two notions

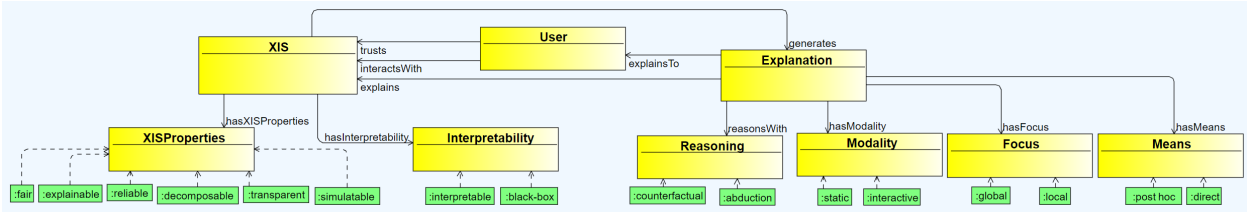

Fig. 2. A terminology for a fully contextualized XAI.

are linked, but being able to simulate a system does not mean that the user understands each component. This leads us to **decomposability** which addresses this particular problem.

Decomposability. *A system is decomposable if each of its components (inputs, parameters and calculations) are interpretable*. Being able to understand each component is vital if the user needs to understand how a prediction was made. Likewise, understanding parameters and calculations can help understand and simulate the model. It contributes to trust and interpretability, in the same way simulatability does. The combination of simulatability and decomposability leads a system towards interpretability, though we cannot conclude that their combination guarantees the interpretability of the system.

3.3.2. Towards trustworthy systems

This section contains terms that do not directly serve explainability or interpretability, but whose purpose is to provide properties that a system should have to gain the trust of a user. As seen in Section 3.1, trust is of utmost importance to ensure that a system is used.

Reliability. *An XIS is reliable if it is at least robust and stable, i.e., it is not sensitive to any type of perturbations or errors to its input or parameters*. Robustness and stability have been defined in Section 2. A reliable system provides the confidence that it will perform in the same manner in any condition. A user will therefore be more inclined to trust a system that is known to be reliable.

Fairness. Fairness is *the ability of a system to avoid any form of unjustified discrimination at each level of the system*. We have decided to specify "unjustified discrimination", because, as discussed in [28], when there is a proven difference between two groups, it can be discriminatory not to use this difference in the system for both groups. Designing fair systems is incredibly difficult, intuitive solutions to prevent discriminations are not always valid, according to Corbett-Davies and Goel [28]. One goal of explainability is to detect when a system is discriminating against a certain group or using variables that should not be relevant. Ensuring the fairness of a system contributes to the trust given to this system. Indeed, if the user is a member of a group that is discriminated against by the system, this user will not trust it nor use it.

Transparency. As discussed in Section 3.3.1, we do not use transparency as the opposite of black-box. We prefer the more global definition of transparency, proposed by Beaudouin et al. [14]. Therefore, *a system is transparent if it provides all the information about its design and functioning for scrutiny*. Transparency does not guarantee that a user will be able to understand the system, but that it will have access to all the information concerning the training data, how the data was preprocessed, the performances of the system and so on. Transparency of a system is a notion similar to open-source programs, in which the whole code can be seen and studied by anyone, which prevents the designer to use this program for malicious purposes. Providing this information is likely to increase the trust in a system, as any experienced user can verify that the system is properly designed and does not present any flaw or bias that could alter the predictions and possibly discriminate.

Figure 2 presents the relationships among the retained terms that compose our terminology. This figure was made with OWLGrEd 1.6.11 [29] using the notations proposed by [30]. This figure shows that the `XIS` generates an `Explanation`. This explanation is characterized by the four properties discussed in Section 3.2. This explanation can be further adapted to the `User`, via the interactions between the `User` and the `XIS`. The `XIS` can possess many properties, defined as instances of `XISProperties`. One mandatory property is its `Interpretability`, which can be either `:interpretable` or `:black box`.

3.4. An illustrative example

We will illustrate the different kinds of explanations that an XIS will provide for different user profiles (a plain user, a domain expert and an AI expert). Our case will concern the task of predicting whether a patient has a tumour, based on the patient's X-ray.

A plain user: a patient with no knowledge in medicine and AI:. for this type of user, the XIS would provide a **local post hoc** explanation. **Local** explanations are focused on the user's case, which is what they will most likely want. **Post hoc** explanations usually simplify how the model works and restrict the amount of information given. An example of explanation that could be given to a patient is: "A tumour was detected on the left side of the body, the system predicted a tumour because the shape is round, with a diameter of 2cm. The shape, size and position of this object correspond to a tumour."

A domain expert: a doctor that makes the diagnosis:. the doctor should be able to make the decision by themselves. The system only helps them decide. The XIS should be **transparent** and **simulatable**, in order for the doctor to have access to similar cases and reproduce the prediction to verify that no mistake was made. The XIS would provide a **local** explanation, that can be **direct** or **post hoc**. Indeed, the doctor needs to focus on the case being predicted, which requires a **local** explanation. If the system's **complexity** is low, a **direct** explanation could be given, otherwise a **post hoc** explanation is preferred. The explanations would also mostly use **counterfactual** reasoning. **Counterfactual** reasoning allows the doctor to compare with other cases and check their hypotheses. Here is an example of explanation given to the doctor, along with an image highlighting the tumour detected: "A tumour was detected in this place. It has a radius of 2cm. Here are similar cases. The most important criteria used to predict the tumour were its position, its size and its opacity on the X-ray. Round objects with a radius higher than 1cm have a 90% chance of being a tumour according to the training data. If the object was 5 cm on the right, it would not have been detected as a tumour".

An AI expert that studies the system to improve it or make sure it is ready for deployment:. the XIS will provide **global** explanations about how the system works, how it performs, where it fails the most. This explanation requires that the system is **transparent**. An example of explanation for this type of user could be: "The system uses a deep neural network architecture and achieves 85% of accuracy. The input is the image of the patient's X-ray in shades of grey with a resolution of 300×300 pixels".

These examples show that the XIS is able to provide different explanations, based on the user's goals and prior knowledge. We also observe from these examples that many properties such as **transparency** or **simulatability** are desirable to generate appropriate explanations.

4. Conclusion

We have introduced the notion of eXplainable Intelligent System that is composed of an explainable model and an explanation interface. Then, we surveyed the literature to identify the definitions of reoccuring terms proposed by the community. Afterwards, we proposed a terminology for a fully contextualized XAI. We defined general concepts of XAI, especially explainability and interpretability. We also highlighted the importance of the **user** in the design of explanations. Following this, we described four properties of an explanation from the literature (Focus, Means, Modality and Reasoning) that will help design explanations, by deciding the value of each of them, giving implicitly the context of the XAI task. We argue that there is a difference between understanding how an algorithm works and understanding its behaviour. For black-boxes, it is impossible to understanding how they work; but through post hoc explanations it is possible to understand their behaviour. Finally, we defined concepts related to the interpretability and trustworthiness of an XIS. The former concepts will help determine the interpretability of an XIS, through a variety of different metrics, such as complexity, simulatability and decomposability.

The concepts related to trust do not contribute to explainability, but they participate in increasing trust in an XIS. We have noted from the literature that the design of a *Responsible AI* is a common goal of the XAI community. *Responsible AI* will allow XIS to be used broadly because it will be trusted and accountable for its decisions. Explainability is the first milestone to reach this goal.

This terminology will allow the proposal of some metrics for XIS using these concepts. It is the necessary prior background for the development of a formal ontology that we are designing. Many terms still need to be clearly

defined, especially the notions linked to the cognitive effort required by a user to understand a system. We need to identify what factors come into place to determine the capability of a user to understand or not an XIS. We believe that the prior knowledge of the user, (e.g. its curriculum, experiences...) has an impact on the required cognitive effort. Other factors that fall within the scope of psychology and sociology could also be taken into account. We will also explore interactive explanations which have not yet been studied in depth by the XAI community. In this way, we hope to provide new and improved definitions that will take into account the latest progress in the field.

References

[1] General data protection regulation (gdpr) (2016).
URL https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32016R0679

[2] D. Gunning, Explainable artificial intelligence (xai), Defense Advanced Research Projects Agency (DARPA), nd Web 2 (2) (2017).

[3] V. Arya, R. K. Bellamy, P.-Y. Chen, A. Dhurandhar, M. Hind, S. C. Hoffman, S. Houde, Q. V. Liao, R. Luss, A. Mojsilović, et al., One explanation does not fit all: A toolkit and taxonomy of ai explainability techniques, arXiv preprint arXiv:1909.03012 (2019).

[4] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, et al., Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai, Information Fusion 58 (2020) 82–115.

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

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

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

[8] R. Calegari, G. Ciatto, A. Omicini, On the integration of symbolic and sub-symbolic techniques for xai: A survey, Intelligenza Artificiale 14 (1) (2020) 7–32.

[9] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, L. Kagal, Explaining explanations: An overview of interpretability of machine learning, in: 2018 IEEE 5th International Conference on data science and advanced analytics (DSAA), IEEE, 2018, pp. 80–89.

[10] G. Futia, A. Vetrò, On the integration of knowledge graphs into deep learning models for a more comprehensible ai—three challenges for future research, Information 11 (2) (2020) 122.

[11] M. T. Ribeiro, S. Singh, C. 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, 2016, pp. 1135–1144.

[12] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, D. Pedreschi, A survey of methods for explaining black box models, ACM computing surveys (CSUR) 51 (5) (2018) 1–42.

[13] M. Wu, M. Hughes, S. Parbhoo, M. Zazzi, V. Roth, F. Doshi-Velez, Beyond sparsity: Tree regularization of deep models for interpretability, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 32, 2018.

[14] V. Beaudouin, I. Bloch, D. Bounie, S. Clémençon, F. d'Alché Buc, J. Eagan, W. Maxwell, P. Mozharovskyi, J. Parekh, Flexible and context-specific ai explainability: a multidisciplinary approach, Available at SSRN 3559477 (2020).

[15] R. R. Hoffman, G. Klein, Explaining explanation, part 1: theoretical foundations, IEEE Intelligent Systems 32 (3) (2017) 68–73.

[16] R. R. Hoffman, S. T. Mueller, G. Klein, Explaining explanation, part 2: Empirical foundations, IEEE Intelligent Systems 32 (4) (2017) 78–86.

[17] G. Klein, Explaining explanation, part 3: The causal landscape, IEEE Intelligent Systems 33 (2) (2018) 83–88.

[18] R. Hoffman, T. Miller, S. T. Mueller, G. Klein, W. J. Clancey, Explaining explanation, part 4: a deep dive on deep nets, IEEE Intelligent Systems 33 (3) (2018) 87–95.

[19] I. Tiddi, M. d'Aquin, E. Motta, An ontology design pattern to define explanations, in: Proceedings of the 8th International Conference on Knowledge Capture, 2015, pp. 1–8.

[20] S. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, arXiv preprint arXiv:1705.07874 (2017).

[21] O.-M. Camburu, Explaining deep neural networks, arXiv preprint arXiv:2010.01496 (2020).

[22] O. L. Dictionaries, Confidence noun - definition, pictures, pronunciation and usage ..., visited on 10/03/2021.
URL https://www.oxfordlearnersdictionaries.com/definition/american_english/confidence

[23] Ieee standard glossary of software engineering terminology, IEEE Std 610.12-1990 (1990) 1–84doi:10.1109/IEEESTD.1990.101064.

[24] O. Bousquet, A. Elisseeff, Stability and generalization, The Journal of Machine Learning Research 2 (2002) 499–526.

[25] Y. Mualla, Explaining the behavior of remote robots to humans: An agent-based approach, Ph.D. thesis, Université Bourgogne Franche-Comté (2020).

[26] C. Dictionary, Trust — meaning in the cambridge english dictionary, visited on 16/03/2021.
URL https://dictionary.cambridge.org/dictionary/english/trust

[27] C. Dictionary, Reasoning — meaning in the cambridge english dictionary, visited on 16/04/2021.
URL https://dictionary.cambridge.org/dictionary/english/reasoning

[28] S. Corbett-Davies, S. Goel, The measure and mismeasure of fairness: A critical review of fair machine learning, arXiv preprint arXiv:1808.00023 (2018).

[29] J. Bārzdiņš, G. Bārzdiņš, K. Čerāns, R. Liepiņš, A. Sprogis, Uml style graphical notation and editor for owl 2, in: International Conference on Business Informatics Research, Springer, 2010, pp. 102–114.

[30] LUMII, Graphical ontology notation - owlgred, visited on 09/04/2021.
URL http://owlgred.lumii.lv/notation