AI: To interpret or to explain?

Jinfeng Zhong, Elsa Negre

AI: To interpret or to explain?

Jinfeng ZHONG, Elsa NEGRE

Paris-Dauphine University, PSL Research University,
CNRS UMR 7243, LAMSADE 75016 Paris France
jinfeng.zhong@dauphine.eu
elsa.negre@lamsade.dauphine.fr

ABSTRACT. *Recent years, the need and demand for explainable/interpretable artificial intelligence (AI) has been growing with the ubiquitous application of AI in our daily life. Human beings tend not to trust an AI system that cannot justify how the results have been generated, which is viewed as a "black box" system. People want that an AI system not only can provide high-quality results but also be transparent in the result generating process, which is called "explainable AI" or "interpretable AI". Most of the state-of-art works about what are explanations and interpretations in AI systems are based on researchers' subjective intuitions without solid theory support, neither common consensus nor mathematical definitions have been achieved, which may be the cause of ill definitions and ambiguity in the use of the two terms:* interpret *and* explain*. In this paper, we seek to disambiguate the use of* interpret *and* explain *in the context of AI with the help of solid theory support from knowledge management. We also discuss possible evaluation methods for interpretability and explainability in AI systems respectively.*

RÉSUMÉ. *Ces dernières années, les besoins en intelligence artificielle (IA) explicable/interprétable ont augmenté avec l'utilisation omniprésente de l'IA dans la vie quotidienne. Les femmes/hommes ont tendance à ne pas faire confiance à un système d'IA incapable de justifier la façon dont les résultats ont été générés et le considèrent comme un système de «boîte noire». Les utilisateurs veulent qu'un système d'IA puisse non seulement fournir des résultats de haute qualité, mais aussi qu'il soit transparent dans le processus de génération de résultats, appelé «IA explicable» ou «IA interprétable». La plupart des travaux actuels sur ce que sont les explications et les interprétations dans les systèmes d'IA sont basés sur les intuitions subjectives des chercheurs sans support théorique solide, ni consensus commun, ni définition mathématique, ce qui peut être la cause d'une mauvaise définition et d'une ambiguïté dans l'utilisation des deux termes:* interpréter *et* expliquer*. Dans cet article, nous cherchons à lever cette ambiguïté dans le contexte de l'IA à l'aide d'un solide support théorique issu de la gestion des connaissances. Nous discutons également des méthodes d'évaluation possibles pour* l'interprétabilité *et* l'explicabilité *dans les systèmes d'IA.*

KEYWORDS: *interpretability, explainability, artificial intelligence.*

MOTS-CLÉS : *interprétabilité, explicabilité, intelligence artificielle.*

---

1. Introduction

Recent years have witnessed ubiquitous application of artificial intelligence (AI), it has made great changes to people's daily life and has become the core technique of many real-world applications, such as recommendation, image processing, etc. People may wonder whether they can trust these techniques or will they work in deployment (Lipton, 2018). In some cases, especially where AI is used to make high-stake decisions such as health care and criminal justice, people wish to know why a system makes certain decisions to control risks since it is hard for them to trust a system without explanations. Besides, since 2018, European Union requires that algorithms used in decision support systems should provide explanations, which is known as "right to explanation" (Voigt, Bussche, 2017). People wish that AI systems could provide high-quality results and reasonable explanations at the same time. The definition, design, optimization and evaluation of such AI systems have attracted lots of attention. Researchers frequently claim that their models are interpretable or explainable, indicating *interpretability* and *explainability* respectively. However, no strict definitions concerning what is *interpretability* and what is *explainability* have been achieved. Some researchers distinguish them (Lipton, 2018; Doshi-Velez, Kim, 2017; Montavon *et al*., 2018) while some use them interchangeably (Miller, 2019; Molnar, 2020; Du *et al*., 2019). These claims are usually based on researchers' subjective intuitions (Miller, 2019) without solid theory support and until now no consensus has been achieved. The ill definitions and ambiguity in the use of the two terms: *interpretability* and *explainability* have made problem formulation difficult in defining, designing and evaluating AI systems that can provide explanations for the results generated.

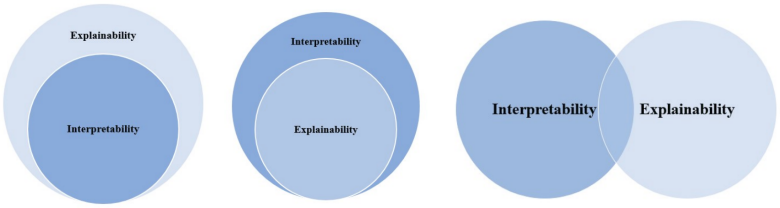
In order to boost the research of *interpretability* and *explainability* in AI, the definitions of these two terms must be critically and seriously engaged and should become a "rigorous science" (Doshi-Velez, Kim, 2017). In this paper, we seek to disambiguate the use of *interpret* and *explain* in the context of AI and propose two frameworks for evaluating *interpretability* and *explainability* in AI systems.

The remainder of this paper is structured as follows. In Section 2 we summarize state-of-art works concerning *interpretability* and *explainability* in the context of AI by answering the following questions: what is *interpretability* and *explainability*?; why *interpretability* and *explainability*?; how to evaluate *interpretability* and *explainability*?. In Section 3, we propose definitions of *interpretability* and *explainability* in the context of AI and discuss possible evaluation methods. Lastly, we conclude and highlight several major challenges for future work.

2. Related work

2.1. What is interpretability *and* explainability*?*

As described in Section 1, in AI the research community, neither common consensus nor strict definitions concerning *interpretability* and *explainability* exist even if



(a) Interpretable models are explainable by default

(b) Explainable models are interpretable by default

(c) Explainable and interpretable are used interchangeably

Figure 1. Relationships of interpretability and explainability in state-of-art works

lots of efforts have been devoted to this research subject. In AI research community, there have been numerous discussions about the definitions of the two terms.

According to Lipton (2018), interpretable models fall into two categories. The first category is transparent models meaning that how models work exactly can be explained and understood by humans. The second category is the models that can provide post-hoc explanations (suggesting post-hoc interpretability) without elucidating precisely how models work. The author distinguishes *interpretability* and *explainability* in that the former enhances the latter. Gilpin *et al*. (2018) argued that *interpretability* and *explainability* should be distinguished in that explainable models are interpretable by default, but the reverse is not always true. The authors argued that explaining explanations is an approach to evaluating *interpretability*. Apparently, in this claim, explanations are strictly defined in that explanations should be able to be justified. In another line of research, some researchers use *interpretability* and *explainability* interchangeably. Miller (2019) discussed explanations in AI from a social science point of view. To the best of our knowledge, this is the very first attempt to link explanation research in AI to psychology research, social science and cognitive science. The author thoroughly surveyed philosophy, psychology and cognitive research related to explanation. According to him, *interpretability* is the degree to which an observer can understand the cause of a decision, which is widely cited. The author equated *interpretability* and *explainability*. The listed surveys and reviews above are not exhausted, they represent three different relationships concerning *interpretability* and *explainability* identified from state-of-art works, which is presented in Figure 1. For more reviews and surveys about *interpretability* and *explainability* in the context of AI, we refer to (Hoffman *et al*., 2018; Zhang, Chen, 2018; Mittelstadt *et al*., 2019).

Figure 1a shows that *interpretability* is a subset of *explainability* meaning that *interpretability* enhances *explainability*. This is the case for Lipton (2018); Montavon *et al*. (2018); Guidotti *et al*. (2018); Mittelstadt *et al*. (2019). In these works, the authors argued that in the context of AI, interpretation concerns the internal mechanisms of models and how models work while explanation concerns why certain re-

sults have been generated. An interpretable model is explainable by default, since the reasoning process behind the results (e.g., a recommendation or a classification) generation is clear indicating that the results can be explained following understandable logic. Figure 1b shows another suggested relationship between *interpretability* and *explainability* that distinguishes them, the reverse. Gilpin *et al*. (2018) can be referred. The author defined that explainable models are interpretable by default, but the reverse is not always true, suggesting that *explainability* implies *interpretability*. In this sense, explanations correctness should be justified, revealing causality relationships behind. Figure 1c shows that two terms can be used interchangeably. According to Molnar (2020), in terms of models, interpretable is more often used; in terms of results, explainable is more often used. Interpreting a model could also mean producing explanations for individual predictions (Molnar, 2020).

2.2. Why interpretability *and* explainability*?*

Reasons or goals of research in *interpretability* and *explainability* are defined differently depending on application domain, as asserted by (Miller, 2019; Hoffman *et al*., 2018), explanations and interpretations are context-aware. Another reason for this diversity is the elusive definitions of these two terms, the reasons and goals may have been defined according to authors' subjective intuitions.

Many researchers have summarized the reasons why research in the two terms are vital in developing responsible AI (Arrieta *et al*., 2020) in a general and abstract level. Lipton (2018) defined five desiderata of interpretability research: (i) trust, (ii) causality, (iii) transferability, (iv) informativeness, (v) fair and ethical decision making. According to Arrieta *et al*. (2020), the reasons why research about the two terms are needed can be summarized as promoting: (i) trustworthiness, (ii) causality, (iii) transferability, (iv) informativeness, (v) confidence, (vi) fairness, (vii) accessibility, (viii) interactivity and privacy awareness. When it comes to real application, as suggested by Hall (2019), the reasons can be summarized as intellectual and social motivations. We believe that they can further be classified into three drives, namely commercial drive, regulation drive and technique drive.

The most important may be commercial drive. Nowadays, AI has become the core competency of many companies, they rely on AI techniques to provide fascinating services. Human beings are curious and adept at learning, they tend not to trust a decision without logical reasoning. Therefore, consumers will not simply trust a system without explanation especially when AI is used to make high-stake decisions, they hope to get explanations or reasonings to support their decision, which means that explanation concerns trust building (Arrieta *et al*., 2020). For example, it is easier for people to trust a recommendation that is well explained compared to a recommendation made by a black-box model.

More importantly, regulation such as GDPR (General Data Protection Regulation) (Voigt, Bussche, 2017) demands that consumers have the legal right to obtain explanations, which makes it necessary to provide explanations to users.

Another drive is from technique. After years of research and real-world applications, researchers have come to know that a model simply based on prediction accuracy cannot always be trusted, since accuracy is an incomplete description of the real-world tasks (Doshi-Velez, Kim, 2017). Not knowing the reasoning process behind result generation may make model builders end up making wrong models. On the contrary, knowing the real reasoning behind models helps designers to debug and improve models, which requires interpretability of models.

2.3. How to evaluate interpretability *and* explainability*?*

There has been a considerable amount of work on *interpretability* and *explainability* in the context of AI, which urged authors to propose corresponding evaluation methods.

Existing evaluation methods usually concern the following: model *interpretability*, how well people can understand a model; quality of explanations, to what extent the provided explanations meet up with design goals. Nguyen and Martínez (2020) proposed a set of objective measurements for simplicity, broadness and fidelity of interpretations. The authors further proposed a taxonomy for metrics according to feature extractor. Hoffman *et al.* (2018) proposed to evaluate explainable artificial intelligence (XAI) by measures for the goodness of explanations and the curiosity in the search for explanations, users' satisfaction and understanding concerning explanation, users' trust and reliance concerning XAI systems, and human-XAI work system performances. The authors further proposed corresponding investigation forms and rating scales.

State-of-art methods for evaluating *interpretability* and *explainability* can also be resumed as automated quantitative methods and human-centered evaluation methods. The former usually involve metrics defined by authors while the latter usually involve human-centered experiments. It should be noted that automated quantitative methods and human-centered methods are both indispensable. The former is designed to guide the selection of a small subset of tasks in human-centered experiments to reduce the overall financial and time cost of such experiments (Nguyen, Martínez, 2020). Automated quantitative measurements can be referred to as functionally grounded evaluation (Doshi-Velez, Kim, 2017), which does not require human involvement. For example, Abdollahi and Nasraoui (2016) proposed *Explainability Precision* and *Explainability Recall* to measure *explainability* of recommendations. Another line of research requires human involvements. For example, Mohseni *et al.* (2018) proposed six measurements for evaluating XAI: (i) human mental model evaluation, (ii) explanation usefulness and satisfaction evaluation, (iii) user trust and reliance, (iv) human task performance, these four evaluations are aimed at users; (v) explainer fidelity, (vi) model trustworthiness these two methods are aimed at developers. The methods proposed here require human involvement and can be time consuming and expensive.

As we discussed in Section 2.1, neither common consensus nor strict definitions concerning *interpretability* and *explainability* exist. This ambiguity also exists in the

evaluation methods. On the one hand, there is no golden rule for evaluation methods, neither strict objective metrics nor standard human evaluations exist. They vary across application domains. On the other hand, due to the elusive definitions of *interpretability* and *explainability*, state-of-art evaluation methods can be misleading. Some claimed to evaluate interpretability of an AI system may end up evaluating the quality of explanations (Doshi-Velez, Kim, 2017). The ambiguity also urges that the definitions of *interpretability* and *explainability* should be seriously engaged in order that the assertions of evaluations could be meaningful.

2.4. Summaries of state-of-art works

There are two popular directions in *interpretability* and *explainability* research: (i) Developing transparent (interpretable) models, the result generation process follows a certain reasoning process that can be expressed in human understandable terms, explanation for model result can be faithful to the original model; (ii) Simply provide explanations to prediction of models while the internal mechanisms of model are not clear, which constitutes post-hoc explanation techniques. Model results are explained by finding the links between the features of input data and the results or by building a simpler model to approximate the original model. The two methods may stem from human cognitive and psychology science.

As Miller (2019) pointed out that explanation in the context of AI involves AI itself, social science and human computer interaction. Seeking for reasons for decisions is the nature of human beings. When human beings must decide something, if permitted, they may consider all the factors to analyze the situation where they are so that they can control the decision to be made; when situation becomes too complicated for them to reason clearly, they may first decide, and latter try to find an explanation for this decision to convince themselves (Lipton, 2018; Zhang, Chen, 2018).

The first decision process is totally transparent, and human beings can fully control it, this is similar to interpretable models, where the internal mechanisms are clear, and it is possible to trace how a result is generated. The second is similar to the post-hoc explanation techniques mentioned above. How the original model works exactly is not elucidated. The model training and result explanation can be separated, a considerable amount of research has been done to develop post-hoc explanation techniques due to the flexibility it offers. For example, Ribeiro *et al*. (2016) proposed *LIME* (Local Interpretable Model-agnostic Explanation) to seek for explaining a single prediction by training interpretable linear models to approximate the original model. According to Ribeiro *et al*. (2016) the explanations provided can reveal how the original model works. On the other hand, some researchers call for designing interpretable models instead of explaining black box models, especially for high-stake decisions. According to Rudin (2019), some post-hoc explanation techniques simply show the trends of results related to features as explanations, which may not be faithful to the original model and cannot reveal how the original model works exactly. Therefore, it would

be less confusing to call them "summaries of predictions", "summary statistics" or "trends" instead of "explanations" (Rudin, 2019).

In this section, we reviewed state-of-art works concerning *interpretability* and *explainability* in the context of AI. The definitions of the two terms given by former researchers cannot be fitted to general cases and lack solid theory support, therefore the proposed definitions are to some extent subjective. In order to boost the research in *interpretability* and *explainability*, the formulation of problems concerning the two terms must be critically and seriously engaged.

3. Our definitions

In this section, we will first propose general definitions of interpretability and explainability by combining linguistics, interpretative frameworks (Tsuchiya, 1993) and mental model (Jones *et al*., 2011). Then we propose a *Interpret/Explain schema* in AI systems. Then, we give the definitions of *interpretability* and *explainability* in AI systems. Lastly, we propose evaluation frameworks for the definitions proposed.

3.1. A general definition

The ambiguity in use of *interpretability* and *explainability* may root from linguistics. According to *Merriam-Webster Dictionary*, interpret (Dictionary, n.d.-b) means to explain or tell the meaning of: present in understandable terms while explain (Dictionary, n.d.-a) means to make plain or understandable, or to give the reason for or cause of. Indeed, the two words have similar meanings, which may have been the reason why they are used interchangeably. However, there are subtle differences that exist between them. For example, some articles in GDPR can be interpreted as "right to explanations", meanwhile explaining GDPR may mean answering why certain articles should have been regularized as such. It seems that to interpret means answering a "what" type question while to explain tends to answer a "why" type question. Accordingly, something being interpretable means that it is decided to have a certain meaning and can be presented in understandable terms; something being explainable means that it can be made understandable.

Human beings possess interpretative framework (Jones *et al*., 2011) also called mental model (Tsuchiya, 1993) through which new information is filtered and stored, thus allowing them to interact with the world around them, filter information and finally create knowledge. Each person has his own interpretative framework, so two men, even though they see the same thing and have the same data, can interpret it differently. Arduin *et al*. (2015) further pointed out that interpretation is central in knowledge management. Constantly, human beings are interpreting information in the process of sense-reading (Polanyi, 1967). For example, right now I am texting information out to you and each of you are receiving this information that you are going to interpret. Through my interpretative frameworks, I give meaning to the information I create to share my knowledge, sense-giving (Polanyi, 1967); and each of you,

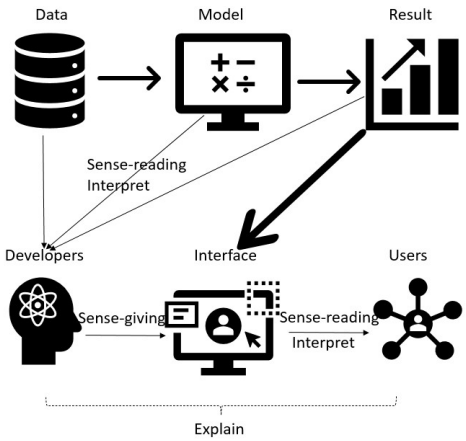


Figure 2. Explain and interpret in AI systems

through the information you perceive and interpret, you will read a meaning to create your knowledge, sense-reading (Polanyi, 1967). This means that it is through interpretation that human beings select data that they have perceived from information and they interpret that data, allowing them to create their own knowledge. By combining the sense-giving and sense-reading, an explanation is realized. This accords with the assertion of Miller (2019):"Explanations are social and involve conversations." From this philosophical perspective, interpretation is a subjective action, while explanation involves interactions.

Combining the linguistic definitions and interpretative framework, definitions of *interpretability* and *explainability* from a general and philosophical perspective can be given as below:

DEFINITION 1. — *Explainability: the ability to make an event understandable; the ability to give the reason or cause of an event.*

DEFINITION 2. — *Interpretability: the degree to which an observer can understand the meaning of an event.*

3.2. Interpret/Explain schema *in AI system*

Since AI technologies aim at creating human intelligence in machines enabling them to think like humans and mimic their actions, it is logical to define *interpretability* and *explainability* in the context of AI based on interpretative frameworks and mental model.

Usually, an AI system has three components (apart from people), the input data, the model and results, as presented in Figure 2. The results of an AI system are displayed to users via an interface. System developers are expected to interpret information from data, model and results through their interpretative frameworks (sense-reading). This information is displayed via an interface to users through developers' interpretative frameworks (sense-giving). Users perceive the information (sense-reading) displayed via this interface. The information conveyed by developers and displayed through the interface constitutes an explanation, as presented in Figure 2. To conclude, in an AI system, developers interpret information from data, model and results to explain to users why certain results have been generated.

3.3. Proposed definitions

In an AI system presented in Figure 2, the input data is interpretable in the sense that developers can extract information from it through statistical analysis or data visualization. Regarding the models, some are transparent while others are not. For transparent models, developers can tell the explicit meanings of each part of models, for example the weights in linear models. In this case, concerning a certain result, they can easily explain to users why this result has been generated to him or even at a higher level, how the whole system works. If the adopted models are not transparent, even developers cannot tell the explicit meanings of parameters. For example, the exact meanings of parameters in deep neural networks. In this case, they may explain why the results are such by finding the links between the features of input data and the results or by building a simpler model to approximate the original model, which constitutes post-hoc explanation techniques.

Therefore, we argue that in the context of AI, explanation is aimed at users and concerns why certain results have been generated. In terms of interpretation, it contains the following aspects: (i) interpretation of input data; (ii) interpretation of model; (iii) interpretation after model training (post-hoc interpretation); (iv) interpretation of explanation. Interpretation of results is closely related to interpretation of data and models. Combining the general definitions of *interpretability* and *explainability* proposed in Section 3.1, we propose the following definitions:

DEFINITION 3. — *Explainability of model results: the ability to make model results understandable; the ability to give the reason or cause of model results.*

DEFINITION 4. — *Interpretability of data: the degree to which one (mainly a developer) can understand the information contained in data, which usually consists of data analysis and data visualization.*

DEFINITION 5. — *Interpretability of model: the degree to which one (mainly a developer) can understand the mechanisms of model*

DEFINITION 6. — *Post-hoc interpretability: the degree to which one (mainly a developer) can explain model results without elucidating precisely how it works*

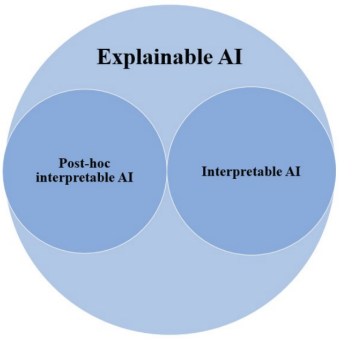


Figure 3. Explainable AI system, Interpretable AI system, Post-hoc interpretable AI system

DEFINITION 7. — *Interpretability of explanations: the degree to which one (mainly a user) can understand a given explanation.*

Based on the definitions above, we believe that "interpret" and "explain" should be distinguished. Interpretation is a subjective action, while explanation involves interactions. Therefore, an explainable AI system refers to an AI system that can explain why certain results have been generated. If the model used in this system is interpretable then the system can further be viewed as an interpretable AI system. If the model used in the system is post-hoc interpretable, the system is a post-hoc interpretable AI system. The relationship is presented in Figure 3.

Besides, we argue the following three questions should be clarified when using the definitions of *interpretability* and *explainability* in the context of AI:

1 Who is concerned about interpretation and explanation?

2 Why *interpretability* and *explainability*?

3 When using interpretable or explainable models?

(1) Concerning the first question, who is concerned about interpretation and explanation? From system developers' perspective, they are interested in all parts of an AI system since they conceptualize the system, they are expected to be clear about the meaning of each part of the system while plain users usually care more about why certain results appear. *Interpretability* of data is mainly aimed at developers and concerns the information contained; *interpretability* of models is mainly aimed at developers and concerns the internal mechanisms of models; model's post-hoc *interpretability* is needed when internal mechanisms of models are not clear and is aimed at developers; *interpretability* of explanations is mainly aimed at users and concerns the quality of explanations. Therefore, how well users can interpret an explanation can reflect the quality of this explanation. *Explainability* of model results is mainly aimed at users and concerns why certain results have been generated.

(2) In terms of the second question, why *interpretability* and *explainability*? The reasons why human beings need interpretability and explainability have been discussed in Section 2.2.

(3) Concerning the third question, when using interpretable or explainable models? In cases where the price of a decision made by a system is negligible, less transparent models such as deep learning techniques can be applied to guarantee the quality of results and provide explanations to users using post-hoc explanation techniques. For example, this is the case for movie recommendation. In cases where the price of a wrong decision can be high, interpretable models are preferred to black-box models (Rudin, 2019).

3.4. Evaluation of interpretability and explainability

Having defined *interpretability* and *explainability* in the context of AI, we now discuss possible evaluation methods for the two terms in order that different systems can be compared meaningfully. According to the definitions in Section 3.3, interpretation and explanation are human-centered process. Besides, explanations and interpretations are context-aware (Miller, 2019; Hoffman *et al*., 2018), they are heavily influenced by humans' prior knowledge. They can vary from person to person and across application domains. Therefore, it is difficult to find general metrics that fit all cases. As discussed in Section 2.3, few objective metrics have been proposed due to the ill definitions of these two terms. State-of-art research focuses on human-centered methods to evaluate the two terms in AI systems, which requires human's participation. Under our definitions, the evaluation of *interpretability* and *explainability* should be separated. Instead of proposing a specific evaluation method whose application may be limited to a certain domain, we will lay out two potential evaluation frameworks.

3.4.1. Evaluation of interpretability

As defined in Section 3.3, interpretation in an AI system contains: (i) interpretation of input data; (ii) interpretation of model; (iii) interpretation after model training (post-hoc interpretation); (iv) interpretation of explanation. Methods for evaluating interpretation of explanations will be discussed in Section 3.4.2. We now discuss methods for evaluating *interpretability* of input data, *interpretability* of model and post-hoc *interpretability* which usually aim at developers as presented in Figure 2.

Evaluation: interpretability of input data Interpretation of input data is independent of the model adopted in an AI system and is usually conducted before model construction. It aims to explore data to extract useful information such as interactions between features. Data analysis techniques such as *Principal Component Analysis*, *Clustering* are widely adopted. Data visualization can represent these interpretations via graphics to simplify the understanding of information contained in data. Therefore, this level evaluation aims to determine whether humans can correctly understand information extracted from input data. User studies can be potential experiments to evaluate these interpretations. Here is a concrete example: Given a dataset, humans

are presented interpretations of it and must indicate the degree to which they can understand these interpretations. This evaluation can be conducted even with lay humans without turning to developers and domain experts who are not always available.

Evaluation: interpretability of model As defined in Section 3.3, *interpretability* of model is the degree to which one (mainly a developer) can understand the mechanisms of model. Lipton (2018) defined three levels of transparency of interpretable models: (i) simulatability; (ii) decomposability; (iii) algorithmic transparency. Simulatability means that humans can produce a prediction in a reasonable time with input data and parameters of models. Evaluating simulatability concerns whether the entire model can be contemplated at once by a human. Decomposability means understanding of a model on a modular level, for example how each feature affects the final results, positively or negatively. Algorithmic transparency applies to the learning algorithm that generates a model. Therefore, we argue that evaluation of model *interpretability* requires application domain knowledge, expertise in AI. Interviewing independent domain experts and developers would be a potential approach. Experiments with them are non-trivial, therefore the questions in the interview should be well designed and should be adapted to the application domain.

Evaluation: post-hoc *interpretability* As discussed in Section 2.4, post-hoc *interpretability* can be achieved by finding the links between the features of input data and the results or by building a simpler model to approximate the original model. Therefore, interpretation of data can be aggregated to yield post-hoc interpretation. However, the assertions made from these interpretations should be careful. Since post-hoc interpretations do not elucidate a model's internal mechanism, they may not be faithful to the original model and cannot reveal how the original model exactly works. Evaluating these interpretations would be non-trivial for lay persons. Therefore, we suggest the involvement of independent domain experts and developers in condition that the experiments are adapted to the application domain.

3.4.2. Evaluation of explainability

Under our definitions, a "good" explanation should: (i) be easy to understand, this is in accordance with the assertion of Miller (2019): "simplicity is one important criterion to evaluate explanation"; (ii) be able to help users understand why certain results have been generated and gain their trust, etc.; (iii) help improve system performances or human performances according to application domain.

When designing explanations concerning a certain result for a user, the following questions are usually considered: how to explain, what to present and what is the effectiveness of the explanations provided in a real-world application. Here, we propose a multilevel evaluation framework for evaluating explanations to guide the design of explanation in an AI system. As presented in Figure 2, how well users interpret the explanations implies the quality of explanation. Therefore, evaluating *interpretability* of explanation also means evaluating the quality of explanation. As explanation is a human-centered process, before an automated quantitative evaluation metric has been adopted, human-centered evaluation methods would be a practical approach.

Table 1. Scaled response to questions

5	4	3	2	1
I agree strongly	I agree somewhat	I am neutral about it	I disagree somewhat	I disagree strongly

Evaluation: how to present? The first level evaluation concerns humans' comprehension of explanation. Mental models are personal and inner presentations of external facts that allow people to interact with the world around them (Tsuchiya, 1993; Jones *et al*., 2011; Arduin *et al*., 2015). Therefore, this level involves evaluating how well humans can understand an explanation presented to them, which can be heavily influenced by presentation style (e.g., text, graphic, etc.) of explanations (Gedikli *et al*., 2014). Since this evaluation is a subjective perspective of users, asking them directly would be a useful way (Mohseni *et al*., 2018; Hoffman *et al*., 2018). Such an evaluation can be conducted by lay humans. A concrete example can be: humans are presented different types of explanations (e.g., text, graphic, etc.) containing the same information, question such as (not limited to) "The explanation is easily understandable to me?" (Hoffman *et al*., 2018) can be asked. Humans must select an answer presented in Table 1 to reflect their agreement degree for the question posed, also suggested by Hoffman *et al*. (2018). According to answers of users the most suitable present style can be selected for a certain application domain. The questions in experiments can vary from different application domains and should be well adapted.

Evaluation: what to present? This evaluation concerns whether humans are satisfied with the explanations provided, whether the explanations provided help them understand why certain results have been generated and whether they trust the explanations, which can be reflected by satisfaction and trust of humans (Mohseni *et al*., 2018). Therefore, it is necessary to evaluate whether explanations have helped to reach these goals, these goals depend heavily on what has been included in explanations (Gedikli *et al*., 2014), for example, scores of other movies and characteristics of movies in a movie recommender system. Satisfaction, trust and reliance are personal feelings, therefore asking humans directly would be an intuitive method. Such methods include but are not limited to Likert-scale questionnaires with scaled responses presented in Table 1, where humans are asked to what extent they are satisfied with or trust the explanations provided. A concrete example question in the Likert-scale questionnaire for evaluating trust can be: "I trust this explanation?" Humans must select an answer presented in Table 1 to reflect the degree to which they trust the explanation. It should be noted that the questions in the Likert-scale questionnaire should be adapted according to the application domain. Users' trust, satisfaction can also be implicitly evaluated by real task applications, which will be discussed later.

Evaluation: real application experiment Application-level evaluation involves real application to verify whether explanations have improved system performances or

human performances depending on the applications domain. For example, in a news recommender system, system performances can be evaluated by CTR (Click Through Rate). Therefore, explanation evaluations can be realized through A/B tests (Dixon *et al*., 2011), etc. Measuring the difference of performance of system without explanation and system with explanation would be a potential approach. While for human performances, explanations are expected to help humans gain performances when doing specific tasks. Imagine an AI system has been created to train medical students recognizing tumor images. The performance gain of students such as prediction accuracy when the system provides explanations can be a criterion for evaluating the quality of explanations. Better system performances or human performances usually mean improved user satisfaction, trust and reliance.

The framework proposed above is incremental and is ordered by the workflow of designing explanations in AI systems. User studies and real application experiments can be costly and time consuming, they should be well designed to minimize these costs (Doshi-Velez, Kim, 2017). The first level evaluation helps developers know the most suitable style of explanation given an application domain. Then by carrying out the second level evaluation, what to present in explanations can be decided. With the former two evaluations, possible explanations are selected in real application experiments, for example, A/B tests (Dixon *et al*., 2011). Designing questionnaires is challenging for the first level evaluation and the second level evaluation, it should be adapted to the application domain. Application-level evaluation is not simple since simulation of realistic settings is non-trivial, which requires expertise in AI and human-computer interaction.

In Section 3.4, we discussed potential evaluation methods for *interpretability* and *explainability* in the context of AI. Since no general objective metrics have been adopted, the two frameworks we propose above are both human-centered. The potential experiments should be well designed and should be adapted depending on the application domain.

4. Conclusions and future works

The key contributions of this paper are the following: (i) a review of state-of-art works on *interpretability* and *explainability* in the context of AI; (ii) a *Interpret/Explain schema* in AI system to present *interpret* and *explain* in an AI system; (iii) based on this schema we propose the definition of *interpretability* and *explainability* in the context of AI. Our definitions are based on interpretative frameworks (Tsuchiya, 1993), mental model (Jones *et al*., 2011), with solid theory support from knowledge management domain. With the definitions proposed, problem formulation such as definition, design, evaluation of explainable AI can be seriously engaged, which will in turn make AI system more transparent when making decisions; (iv) proposition of two potential evaluation frameworks for *interpretability* and *explainability* in AI systems. The limits of our work are: (i) we limit our discussion in the context of AI. For non-AI models such as physics based or symbolic models, similar issues concerning

interpretability and *explainability* exist and worth further studies by domain experts; (ii) the evaluation frameworks we proposed here require the involvements of human, which can be time consuming and costing.

For future work, the following direction can be promising: (i) more in-depth work to be continued to verify the utility and applicability of the proposed definitions such as in recommender systems **(information systems for decision support)**; (ii) objective metrics to evaluate *interpretability* and *explainability* in AI systems should be proposed to make up for the disadvantages of human-centered experiments: time consuming and costly; (iii) design and use of interpretable models should be encouraged. More and more researchers have proposed to design responsible AI (Arrieta *et al.*, 2020), which requires logical reasoning and transparency of models. This is especially required for high-stake decisions, where the prices of wrong decisions are high; (iv) since explanations involve interactions, the way and style of displaying them is a promising direction which worth further research.

Acknowledgements

This paper would not have been possible without the insightful discussions with Michel Grundstein and Camille Rosenthal-Sabroux. The authors sincerely appreciate their support and feedback.

References

Abdollahi B., Nasraoui O. (2016). Explainable matrix factorization for collaborative filtering. In *Proceedings of the 25th international conference companion on world wide web*, pp. 5–6.

Arduin P.-E., Grundstein M., Rosenthal-Sabroux C. (2015). *Information and knowledge systems* (Vol. 2). Wiley Online Library.

Arrieta A. B., Díaz-Rodríguez N., Del Ser J., Bennetot A., Tabik S., Barbado A. *et al*. (2020). Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, Vol. 58, pp. 82–115.

Dictionary M.-W. (n.d.-a). *Definition of explain.* https://www.merriam-webster.com/dictionary/explain.

Dictionary M.-W. (n.d.-b). *Definition of interpret.* https://www.merriam-webster.com/dictionary/interpret.

Dixon E., Enos E., Brodmerkle S. (2011, July 5). *A/b testing of a webpage.* Google Patents. (US Patent 7,975,000)

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

Du M., Liu N., Hu X. (2019). Techniques for interpretable machine learning. *Communications of the ACM*, Vol. 63, No. 1, pp. 68–77.

Gedikli F., Jannach D., Ge M. (2014). How should i explain? a comparison of different explanation types for recommender systems. *International Journal of Human-Computer Studies*, Vol. 72, No. 4, pp. 367–382.

Gilpin L. H., Bau D., Yuan B. Z., Bajwa A., Specter M., Kagal L. (2018). Explaining explanations: An approach to evaluating interpretability of machine learning. *arXiv preprint arXiv:1806.00069*.

Guidotti R., Monreale A., Ruggieri S., Turini F., Giannotti F., Pedreschi D. (2018). A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, Vol. 51, No. 5, pp. 1–42.

Hall P. (2019). *An introduction to machine learning interpretability*. O'Reilly Media, Incorporated.

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

Jones N. A., Ross H., Lynam T., Perez P., Leitch A. (2011). Mental models: an interdisciplinary synthesis of theory and methods. *Ecology and Society*, Vol. 16, No. 1.

Lipton Z. C. (2018). The mythos of model interpretability. *Queue*, Vol. 16, No. 3, pp. 31–57.

Miller T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, Vol. 267, pp. 1–38.

Mittelstadt B., Russell C., Wachter S. (2019). Explaining explanations in ai. In *Proceedings of the conference on fairness, accountability, and transparency*, pp. 279–288.

Mohseni S., Zarei N., Ragan E. D. (2018). A multidisciplinary survey and framework for design and evaluation of explainable ai systems. *arXiv*, pp. arXiv–1811.

Molnar C. (2020). *Interpretable machine learning*. Lulu. com.

Montavon G., Samek W., Müller K.-R. (2018). Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*, Vol. 73, pp. 1–15.

Nguyen A.-p., Martínez M. R. (2020). On quantitative aspects of model interpretability. *arXiv preprint arXiv:2007.07584*.

Polanyi M. (1967). Sense-giving and sense-reading. *Philosophy*, Vol. 42, No. 162, pp. 301–325.

Ribeiro M. T., Singh S., Guestrin C. (2016). " 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*, pp. 1135–1144.

Rudin C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, Vol. 1, No. 5, pp. 206–215.

Tsuchiya S. (1993). Improving knowledge creation ability through organizational learning. In *Ismick'93 proceedings, international symposium on the management of industrial and corporate knowledge*, pp. 87–95.

Voigt P., Bussche A. Von dem. (2017). The eu general data protection regulation (gdpr). *A Practical Guide, 1st Ed., Cham: Springer International Publishing*.

Zhang Y., Chen X. (2018). Explainable recommendation: A survey and new perspectives. *arXiv preprint arXiv:1804.11192*.