Agent-Based Explanations in AI: Towards an Abstract Framework

Giovanni Ciatto[1](✉), Michael I. Schumacher[2], Andrea Omicini[1], and Davide Calvaresi[2]

[1] University of Bologna, 47521 Cesena, FC, Italy
{giovanni.ciatto,andrea.omicini}@unibo.it
[2] HES-SO Valais, 3960 Sierre, Switzerland
{michael.schumacher,davide.calvaresi}@hevs.ch

Abstract. Recently, the eXplainable AI (XAI) research community has focused on developing methods making Machine Learning (ML) predictors more *interpretable* and *explainable*. Unfortunately, researchers are struggling to converge towards an unambiguous definition of notions such as *interpretation*, or, *explanation*—which are often (and mistakenly) used interchangeably. Furthermore, despite the sound metaphors that Multi-Agent System (MAS) could easily provide to address such a challenge, and agent-oriented perspective on the topic is still missing. Thus, this paper proposes an abstract and formal framework for XAI-based MAS, reconciling notions, and results from the literature.

Keywords: Explainable artificial intelligence · Multi-agent systems · Understandability · Explainability · Interpretability

1 Introduction

The adoption of intelligent systems (IS) in modern society is booming: the trend is mostly due to the recent momentum gained by Machine Learning (ML). In the past decades, disruptive results from ML dictated several waves of temporary yet massive adoption of AI systems, in both academia and industry. Therefore, some authors refer to the current era as the third *spring of AI*—stressing that AI has already lived two *winters*.

As in the previous springs of AI, the expectations are being inflated by the promising predictive capabilities showed by ML-based IS. Besides the remarkable computational capability characterising this era, the vast availability of data is the second key aspect enabling the new spring. However, also modern researchers and stakeholders are experiencing problems stemming from the *opacity* of ML-based solutions.

The opacity of numeric predictors (i.e., the outcome of ML techniques applied on data) is a broadly acknowledged issue, which has been studied even before

This paper is the full version of the extended abstract [9] soon to be appearing on the AAMAS '20 Proceedings.

D. Calvaresi et al. (Eds.): EXTRAAMAS 2020, LNAI 12175, pp. 3–20, 2020.
https://doi.org/10.1007/978-3-030-51924-7_1

the current spring of AI. However, mostly due to the unprecedented pace and extent of ML adoption in several, often critical domains (e.g., finance, healthcare, and law), the need for addressing such opacity issues is more compelling than ever [2].

The opaqueness of ML-based solutions is an unacceptable condition in a world where ML is involved in many (safety-)critical activities. Indeed, performing automatic, good predictions (resp. provide useful decisions) is essential as much as letting the humans involved in those contexts *understand* why and how such predictions (resp. decisions) have been obtained. When humans cannot understand the outcome or the behaviour of ML predictors involved in some business processes, bad consequences can follow. This is because, in the current society, the liability of decisions/actions is still mainly associated with human beings (even if the outcomes have been obtained via IS). To make the picture even more complicated, modern regulations recognise citizens right to receive meaningful explanations when automatic decisions may affect their lives [12]. For all the above reasons, the problem of understanding ML results is rapidly gaining momentum in recent AI research [5].

The topic of understandability in AI is nowadays the main concern of the eXplainable AI community (XAI henceforth), whose name is due to a successful project of DARPA [24]. There, the authors review the main approaches to make AI more understandable to human beings. However, as further discussed in this paper, we argue that studies in this field are flawed by a fundamental issue—namely, they lack an unambiguous definition for the concept of *explanation* and, consequently, a clear understanding of what X in XAI actually means. Indeed, the notion of *explanation* is not clearly established in the literature, nor is there a consensus on what the property named "explainability" should imply. This is especially true for ML-based solutions, where knowledge is represented in a *sub-symbolic*, unintelligible way.

Similar issues exist as far as the notion of *interpretation* is concerned. The two terms are sometimes used interchangeably in the literature, whereas other times they carry different meanings. To face such issues, we argue that since multi-agent systems (MAS) offer a coherent yet expressive set of abstractions, promoting *conceptual integrity* in the engineering of complex software systems [18] – and of socio-technical systems (STS) in particular –, they can be exploited to define a sound and unambiguous reference framework for XAI.

In this paper, we propose an abstract framework for XAI relying on notions and results from the MAS literature. The framework is mostly targetting sub-symbolic AI and ML-based intelligent systems. In particular, our framework introduces a clear distinction among two orthogonal, yet interrelated, activities – i.e., *interpretation* and *explanation* – which can be performed on sub-symbolic predictors to make them more understandable in the eyes of human beings. Thus, it provides a formal definition for such activities in the MAS perspective, thus stressing the *objective* nature of explanation, other than the *subjective* nature of interpretation.

Accordingly, the paper is structured as follows. In Sect. 2 we provide an overview of the XAI research domain. In Sect. 3 we present our abstract framework. Then, Sect. 4 assesses our framework by showing how it can help in unambiguously defining the main problems in XAI. Conversely, Sect. 5 speculates on some future directions. Finally, Sect. 6 concludes the paper.

2 Background

Most IS today leverage on *sub-symbolic* predictive models that are trained from data through ML. The reason for such wide adoption is easy to understand. We live in an era where the availability of data is unprecedented, and ML algorithms make it possible to detect useful statistical information hidden into such data semi-automatically. Information, in turn, supports decision making, monitoring, planning, and forecasting in virtually any human activity where data is available.

However, ML is not the silver bullet. Despite the increased predictive power, ML comes with some well-known drawbacks which make it perform poorly in some use cases. One blatant example is algorithmic *opacity*—that is, essentially, the difficulty of the human mind in *understanding* how ML-based IS function or compute their outputs. This represents a serious issue in all those contexts where human beings are liable for their decision, or, when they are expected to provide some sort of *explanation* for it—even if the decision has been supported by some IS. For instance, think about a doctor willing to motivate a serious, computer-aided diagnosis, or, a bank employee in need of explaining to a customer why his/her profile is inadequate for a loan. In all contexts, ML is at the same time an enabling – as it aids the decision process by automating it – and a limiting factor—as opacity prevents human awareness of *how* the decision process works.

Opacity is why ML predictors are also referred to as *black boxes* into the literature. The “black box” expression refers to models where knowledge is not explicitly represented [15]. The lack of some explicit, symbolic representation of knowledge is what makes it hard for humans to *understand* the functioning of black boxes, and why they led to suggest or undertake a given decision. Clearly, troubles in understanding black-box content and functioning prevent people from fully trusting – therefore accepting – them. To make the picture even more complex, current regulations such as the GDPR [25] are starting to recognise the citizens’ *right to explanation* [12]—which implicitly requires IS to eventually become *understandable*. Indeed, understanding IS is essential to guarantee algorithmic fairness, to identify potential bias/problems in the training data, and to ensure that IS perform as designed and expected.

Unfortunately, the notion of understandability is neither standardised nor systematically assessed, yet. At the same time, there is no consensus on what exactly providing an *explanation* should mean when decisions are supported by a black box. However, several authors agree that not all black boxes are equally *interpretable*—meaning that some black boxes are more susceptible to understand than others for our minds. For example, Fig. 1 is a common way to illustrate the differences in black-box interpretability.

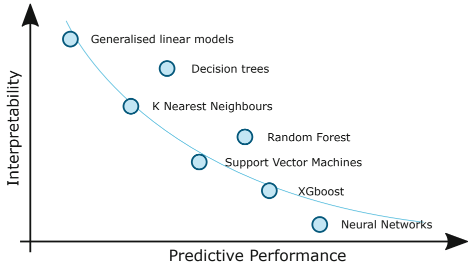

Fig. 1. Interpretability/performance trade-off for some common sorts of black-box predictors

Even though informal – as pointed on in [22], given the lack of ways to measure "interpretability" – Fig. 1 effectively expresses why more research is need on understandability. In fact, the image essentially states how the better performing black boxes are also the less interpretable ones. This is a problem in practice since only rarely predictive performances can be sacrificed in favour of a higher degree of interpretability.

To tackle such issues, the XAI research field has recently emerged. Among the many authors and organisations involved in the topic, DARPA has proposed a comprehensive research road map [24], which reviews the main approaches to make black boxes more understandable. There, DARPA categorises the many currently available techniques aimed at building meaningful interpretations or explanations for black-box models, it summarises the open problems and challenges, and it provides a successful reference framework for the researchers interested in the field. Unfortunately, despite the great effort in defining terms, objects, and methods for the research line, a clear definition of fundamental notions such as *interpretation* and *explanation* is still missing.

2.1 Related Work

Notions such as explanation, interpretation, and transparency are mentioned, introduced, or informally defined in several works. However, a coherent framework has not yet emerged. This subsection recalls some significant contributions from the literature discussing concepts of explanation and interpretation – or any variant of theirs. Our goal here is to highlight the current lack of consensus on the meaning of such terms, for which we propose a possible, unambiguous alternative in the next sections.

Similarly to what we do here, Lipton [15] starts his discussion by recognising how most definitions of ML interpretability are often inconsistent and underspecified. In his clarification effort, Lipton essentially maps interpretability on the notion of *transparency*, and explanation on the notion of *post-hoc* interpretation. Then, he enumerates and describes the several possible variants of transparency, that are *(i)* simulatability – i.e., the *practical* possibility, for a human being, to "contemplate the entire model at once" and simulate its functioning on some data – which characterises, for instance, generalised linear models; *(ii)* decomposability – i.e., the possibility, for the model to be decomposed in elementary parts whose functioning is intuitively understandable for humans and helpful in understanding the whole model – which characterises, for instance, decision trees; and *(iii)* algorithmic transparency – i.e., the possibility, for a human being, to intuitively understand how a given learning algorithm, or the predictors it produces, operate – which characterises, for instance, k-nearest-neighbors techniques. Similarly, *post-hoc* interpretability is defined as an approach where some information is extracted from a black box in order to ease its understanding. Such information have not necessarily to expose the internal functioning of the black box. As stated in the paper: "examples of post-hoc interpretations include the verbal explanations produced by people or the saliency maps used to analyze deep neural networks".

Conversely, Besold et al. [3] discuss the notion of explanation at a fundamental level. There, the authors provide a philosophical overview on such topic, concluding that "explanation is an epistemological activity and explanations are an epistemological accomplishment—they satisfy a sort of epistemic longing, a desire to know something more than we currently know. Besides satisfying this desire to know, they also provide the explanation-seeker a direction of action that they did not previously have". Then they discuss the topic of explanation in AI from a historical perspective. In particular, when focussing on ML, they introduce the following classification of IS systems: *(i)* opaque systems – i.e., black boxes acting as oracles where the logic behind predictions is not observable or understandable –, *(ii)* interpretable systems – i.e., white boxes whose functioning is understandable to humans, also thanks to the expertise, resources, or tools –, and *(iii)* comprehensible systems—i.e., "systems which emit *symbols* along with their outputs, allowing the user to relate properties of the input to the output". According to this classification, while interpretable systems can be inspected to be understood – thus letting observer draw their explanations by themselves– comprehensible systems must explicitly provide a symbolic explanation of their functioning. The focus is thus on *who* produces explanations, rather than *how*.

In [10], the interpretability of ML systems is defined as "the ability to explain or to present in understandable terms to a human". Interpretations and explanations are therefore collapsed in this work, as confirmed by the authors using the two terms interchangeably. The remainder of that paper focuses *(i)* on identifying under which circumstances interpretability is needed in ML, and *(ii)* how to assess the quality of some explanation.

The survey by Guidotti et al. [13] is a nice entry point to explainable ML. It consists of an exhaustive and recent survey overviewing the main notions, goals, problems, and (sub-)categories in this field, and it encompasses a taxonomy of existing approaches for "opening the black box"—which may vary a lot depending on the sort of data and the family of predictors at hand. There, the authors define the verb to interpret as the act of "providing some meaning of explaining and presenting in understandable terms some concepts", borrowing such a definition from the Merriam-Webster[1] dictionary. Consequently, they define interpretability as "the ability to explain or to provide the meaning in understandable terms to a human"— a definition they again borrow from [10]. So, in this case as well the notions of *interpretation* and *explanations* are collapsed.

In [22], Rudin does not explicitly define explainability or interpretability, and she refers to interpretable or explainable ML almost interchangeably. However, she states some interesting properties of *interpretability*, which influenced our work. In particular, she acknowledges that "interpretability is a domain-specific notion". Furthermore, she links interpretability of information with its complexity – and, in particular, its *sparsity* –, as the amount of cognitive entities the human mind can handle at once is minimal ($\sim 7 \pm 2$ according to [16]). As far as explainability is concerned, apparently, Rudin adopts a *post-hoc* perspective similar to the one in [15], as she writes, "an explanation is a separate model that is supposed to replicate most of the behaviour of a black box". In the remainder of that paper, the author argues how the path towards interpretable ML steps through broader adoption of inherently interpretable predictors – such as generalised linear models or decision trees – rather than relying on *post-hoc* explanations which do not reveal what is inside black boxes—thus preventing their full understanding.

Finally, the recent article by Rosenfeld et al. [21] is similar in its intents to our current work. There, the authors attempt to formally define what explanation and interpretation respectively are in the case of ML-based classification. However, their work differs from ours in several ways. In particular, they define interpretation and explanation differently from what we do. In fact, according to the authors, "interpretation" is a function mapping data, data schemes, and predictors to some representation of the predictors internal logic, whereas "explanation" is defined as "the human-centric objective for the user to understand" a predictor using the aforementioned interpretation function. Other notions are formally defined into the paper, such as for instance, *(i)* explicitness, *(ii)* fairness, *(iii)* faithfulness, *(iv)* justification, and *(v)* transparency. Such concepts are formally defined in terms of the aforementioned interpretation and explanation functions. The reminder of that paper then re-interprets the field of XAI in terms of all the notions mentioned so far.

3 Explanation *vs.* Interpretation

This section introduces the preliminary notions, intuitions, and notations we leverage upon in Sect. 3.1 and subsequent sections, in order to formalise our abstract framework for agent-based explanations. We start by providing an intuition for the notion of *interpretation*, and, consequently, for the *act* of interpreting something. Accordingly, we provide an intuition for the property of "being interpretable" as well, stressing its comparative nature. Analogously to what we did with *interpretation*, we then provide intuitions for terms such as *explanation* and its derivatives.

[1] https://www.merriam-webster.com/dictionary/interpret.

About Interpretation. Taking inspiration from the field of Logics, we define the *act* of "interpreting" some object X as the activity performed by an agent A – either human or software – assigning a *subjective* meaning to X. Such meaning is what we call *interpretation*. Roughly speaking, an object X is said to be *interpretable* for an agent A if it is *easy* for A to draw an interpretation for X—where "easy" means A requires a low *computational* (or *cognitive*) effort to understand X. For instance, consider the case of road signs, which contain symbols instead of scripts to be easily, quickly, and intuitively interpretable.

We model such intuition through a function $I_A(X) \mapsto [0, 1]$ providing a *degree of interpretability* – or simply interpretability, for short – for X, in the eyes of A. The value $I_A(X)$ is not required to be directly observable or measurable in practice, since agents' mind may be inaccessible in most cases. This is far from being an issue, since we are not actually interested in the absolute value of $I_A(X)$, for some object X, but rather we are interested in being able to order different objects w.r.t. their subjective interpretability. For instance, we write $I_A(X) > I_A(Y)$, for two objects X and Y, meaning that the former is more interpretable than the latter, according to A. For example, consider the case of a neural network and a decision tree, both trained on the same examples to solve the same problem with similar predictive performances. Both objects may be represented as graphs. However, it is likely for a human observer to see the decision tree as more interpretable—as their nodes bring semantically meaningful, high-level information.

Summarising, we stress the subjective nature of interpretations, as agents assign them to objects according to their State of Mind (SoM) [19] and background knowledge, and they need not be formally defined any further.

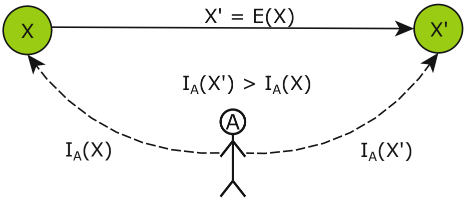

Fig. 2. Explanation vs. Interpretation: a simple framework

About Explanation. We define "explaining" as the activity of producing a more interpretable object X' out of a less interpretable one, namely X, performed by agent A. More formally, we define *explanation* as a function $E(X) \mapsto X'$ mapping objects into other objects, possibly, in such a way that $I_A(X') > I_A(X)$, for some agent A. The simple framework described so far is summarised in Fig. 2.

Notice that human beings tend to collapse into the concept of "explanation" the whole sequence of steps actually involving both explaining and interpreting, according to our framework. This happens because, if the explained object X' is as interpretable for the listening agent B as it is for the explaining agent A, then both A and B are likely to be satisfied with X'. Conversely, it may also happen the explanation E adopted by A produces an object X', which is more interpretable than X for A but not for B. Similarly to how two persons would handle such an unpleasant situation, we envision that interaction and communication may be adopted to break such *impasses* in multi-agent systems.

In the following sections, we develop such an idea, describing how our simple framework could be extended to support ML-based intelligent systems.

3.1 A Conceptual Framework for XAI

In AI several tasks can be reduced to a functional model $M : X \rightarrow Y$ mapping some input data $X \subseteq \mathcal{X}$ from an input domain $\mathcal{X}$ into some output data $Y \subseteq \mathcal{Y}$ from an output domain $\mathcal{Y}$.

In the following, we denote as $\mathcal{M}$ the set of all *analogous* models $M' : X \rightarrow \mathcal{Y}$, which attempts to solve the same problem on the same input data—usually, in (possibly slightly) different ways. For instance, according to this definition, a decision tree and a neural network, both trained on the same data-set to solve the same classification problem with similar accuracies, are analogous—even if they belong to different families of predictors.

At a very high abstraction level, many tasks in AI may be devoted to compute, for instance:

- the best $M^* \in \mathcal{M}$, given $X \subseteq \mathcal{X}$ and $Y \subseteq \mathcal{Y}$ (e.g. supervised ML),
- the best M^* and Y, given X (e.g. unsupervised ML),
- the best Y^*, given X and M (e.g. informed/uninformed search),
- the best X^*, given Y and M (e.g. abduction, most likely explanation), etc.

according to some goodness criterion which is specific for the task at hand.

In the reminder of this section, we discuss how explanation may be defined as a function searching or building a – possibly more interpretable – model w.r.t. the one to be explained. For this process to even make sense, of course, we require the resulting model to be not only analogous to the original but also similar in the way it behaves on the same data. We formalise such a concept through the notion of *fidelity*.

Let $M, M' \in \mathcal{M}$ be two analogous models. We then say M has a *locally* good *fidelity* w.r.t. M' and Z if and only if $\Delta f(M(Z), M'(Z)) < \delta$ for some arbitrarily small threshold $\delta \geq 0$ and for some subset of the input data $Z \subset X$. There, $\Delta f : 2^{\mathcal{Y}} \times 2^{\mathcal{Y}} \rightarrow \mathbb{R}_{\geq 0}$ is a function measuring the performance *difference* among two analogous models.

Local Interpretations. When an observer agent A is *interpreting* a model M behaviour w.r.t. some input data $Z \subseteq X$, it is actually trying to assign a subjective interpretability value $I_A(R)$ to some representation $R = r(M, Z)$ of choice, aimed at highlighting the behaviour of M w.r.t. the data in Z. There, $r : \mathcal{M} \times 2^{\mathcal{X}} \rightarrow \mathcal{R}$ is *representation means*, i.e., a function mapping models into *local* representations w.r.t. a particular subset of the input domain, whereas $\mathcal{R}$ is the set of model representations. For instance, in the case M is a classifier, R may

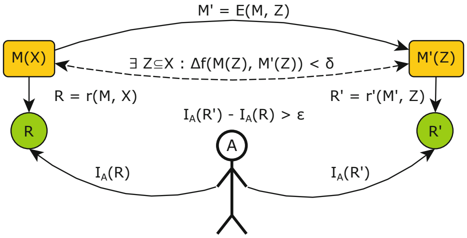

Fig. 3. Local explanation and interpretation of model M

be a graphical representation of (a portion of) the decision boundary/surface for a couple of input features.

There may be more or less interpretable *representations* of a particular model for the same observer A. Furthermore, representations may be either global or local as well, depending on whether they represent the behaviour of the model for the whole input space, or for just a portion of it. For example, consider the case of a plot showing the decision boundary of a neural network classifier. This representation is likely far more interpretable to the human observer than a graph representation showing the network structure, as it synthesise the global behaviour of the network concisely and intuitively. Similarly, saliency maps are an interpretable way to *locally* represent the behaviour of a network w.r.t. some particular input image. So, a way for easing interpretation for a given model behaviour w.r.t. a particular sort of inputs is about looking for the right representation in the eyes of the observer.

Local Explanations. Conversely, when an observer A is *explaining* a model M w.r.t. some input data $Z \subseteq X$, it is actually trying to produce a model $M' = E(M, Z)$ through some function $E : \mathcal{M} \times 2^{\mathcal{X}} \rightarrow \mathcal{M}$. In this case, we say M' is a *local explanation* for M w.r.t. to Z. We also say that M' is produced through the explanation strategy E.

Furthermore, we define an explanation M' as *admissible* if it has a valid fidelity w.r.t. the original model M and the data in Z—where Z is the same subset of the input data used by the explanation strategy. More precisely, we say M' is δ-admissible in Z w.r.t. M if $\mathit{\Delta} f(M(Z), M'(Z)) < \delta$.

Finally, we define an explanation M' as *clear* for A, in Z, and w.r.t. the original model M, if there exists some representation $R' = r(M', Z)$ which is more interpretable than the original model representation R. More precisely, we say M' is ε-clear for A, in Z, and w.r.t M if $I_A(R') - I_A(R) > \varepsilon$ for some arbitrarily big threshold $\varepsilon > 0$.

Several *explanations* may actually be produced for the same model M. For each explanation, there may be again more or less interpretable *representations*. Of course, explanations are useful if they ease the seek for more interpretable representations. Thus, providing an explanation for a given model behaviour w.r.t. a particular class of inputs is about creating *ad-hoc* metaphors aimed at easing the observer's understanding.

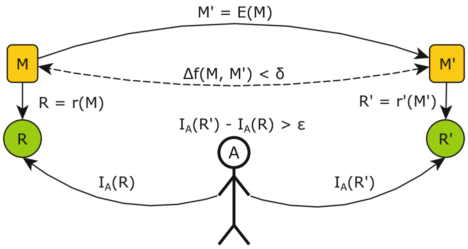

Fig. 4. Global explanation and interpretation of model

Global/Local Explanations. The theoretical framework described so far – which is graphically synthesised in Fig. 3 – is aimed at modelling *local* interpretations and explanations, that are, the two means an explanator agent may exploit in order to make AI tasks' *outcomes* more understandable in the eyes of some explanee.

Conversely, when the goal is not to understand some model outcome, but the model itself, from a *global* perspective – or, equivalently, when the goal is to understand the model outcome w.r.t the whole set of input data X –, the theoretical framework described so far is simplified as shown in Fig. 4, where the dependency on the input data is omitted from functions E, Δf, and r. This is possible because we consider the global case as a particular case of the local one, where $Z \equiv X$.

Finally, we remark that the case where a model M is to be understood on a single input-output pair, say x and $y = M(x)$, is simply captured by the aforementioned local model, through the constraint $Z = \{x\}$ and $M(Z) = \{y\}$.

3.2 Discussion

Our framework is deliberately abstract in order to capture a number of features we believe to be essential in XAI. First of all, our framework acknowledges – and properly captures – the orthogonality of interpretability w.r.t. explainability. This is quite new, indeed, considering that most authors tend to use the two concepts as if they were equivalent or interchangeable.

Furthermore, our framework explicitly recognises the *subjective* nature of interpretation, as well as the subtly *objective* nature of explanation. Indeed, interpretation is a subjective activity directly related to agents' perception and SoM, whereas explanation is an epistemic, computational action which aims at producing a high-fidelity model. The last step is objective in the sense that it does not depend on the agent's perceptions and SoM, thus being reproducible in principle. Of course, the *effectives* of an explanation is again a subjective aspect. Indeed, a clear explanation (for some agent) is a more interpretable variant of some given model—thus, the subjective activity of interpretation is again implicitly involved.

The proposed framework also captures the importance of representations. This is yet another degree of freedom that agents may exploit in their seek for a wider understandability of a given model. While other frameworks consider interpretability as an intrinsic property of AI models, we stress the fact that a given model may be represented in several ways, and each representation may be interpreted differently by different agents. As further discussed in the remainder of this paper, this is far from being an issue. This subjectivity is deliberate, and it is the starting point of some interesting discussions.

Finally, our framework acknowledges the global/local duality of both explanation and interpretation, thus enabling AI models to be understood either general or with respect to a particular input/output pair.

3.3 Practical Remarks

The ultimate goal of our framework is to provide a general, flexible, yet minimal framework describing the many aspects concerning AI understandability in the eyes of a *single* agent. We here illustrate several practical issues affecting our framework in practice, and further constraining it.

According to our conceptual framework, a *rational* agent seeking to understand some model M (or make it understandable) may either choose to elaborate on the *interpretation axis* – thus looking for a (better) representation R of M – or it can elaborate on the *explainability axis*—thus producing a novel, high fidelity model M', coming with a representation R' which is more interpretable than the original one (i.e., R).

Notice that, in practice, the nature of the model constrains the set of admissible representations. This means that a rational agent is likely to exploit both the explanation and interpretation axes in the general case—because novel representations may become available through an explanation. we argue and assume that each family of AI models comes with just a few *natural* representations. Because of this practical remark, we expect that, in real-world scenarios, an agent seeking for understandability is likely to "work" on both the interpretation and the explanation axes.

For instance, consider decision trees, which come with a natural representation as a tree of subsequent choices leading to a decision. Conversely, neural networks can either be represented as graphs or as algebraic combinations of tensors. In any case, neural network models are commonly considered less interpretable than other models. In such situation, a rational agent willing to make a neural network more understandable may choose to combine decision trees extraction (explanation) – possibly focusing on methods from the literature [1,4] – to produce a decision tree whose tree-like structure (representation) could be presented to the human observer to ease his/her interpretation. The decision-tree like representation is not ordinarily available for neural networks, but it may become available provided that an explanation step is performed.

Another interesting trait of our framework concerns the semantics of clear explanations. The current definition requires explanation strategies to consume a model M with a given representation R and to produce a high-fidelity model M' for which a representation R' exists, which is more interpretable than R. Several semantics may fit this definition. This is deliberate, since different semantics may come with different computational requirements, properties, and guarantees. For instance, one agent may be interested in finding the *best* explanation—that is, the one for which *each* representation is more interpretable than the most interpretable representation of the original model. Similarly, in some cases, it may be sufficient – other than more feasible – to find an *admissible* explanation—that is, a high-fidelity model for which *some* representation exists that is more interpretable than *some* representation of the original model. However, the inspection of the possible semantics and their properties falls outside the scope of this paper and is going to be considered as a future research direction.

4 Assessment of the Framework

The abstraction level of the presented framework has also been conceived in order to capture most of the current state of the art. Along this line, this section aims at validating the fitting of the existing contributions w.r.t. the framework presented in Sect. 3.1: if our framework is expressive enough, it should allow most (if not all) existing approaches to be uniformly framed, to be easily understood and compared. To this end, we leverage on the work by Guidotti et al. [13], where the authors perform a detailed and extensive survey on the state-of-the-art methods for XAI, by categorising the surveyed methods according to an elegant taxonomy. Thus, hereafter, we adopt their taxonomy as a reference for assessing our framework.

The taxonomy proposed by Guidotti et al. essentially discriminates among two main categories of XAI methods. These are the "transparent box design" and the "black-box explanation" categories. While the former category is not further decomposed, the latter comes with three more sub-categories, such as "model explanation", the "outcome explanation", and the "model inspection". Notice that, despite the authors' definition of "explanation" does not precisely match the one proposed in this paper, we maintained the original categorisation.

The remainder of this section navigates such a taxonomy accordingly, by describing how each (sub-)category – along with the methods therein located – fits our abstract framework.

4.1 Model Explanation

The mapping of the methods classified as part of the "model explanation" subcategory into our framework is seamless. Hence, it can be defined as follows:

> Let M be a sub-symbolic classifier whose internal functioning representation R is poorly interpretable in the eyes of some explanee A, and let $E(\cdot)$ be some *global* explanation strategy. Then, the model explanation problem consists of computing some *global* explanation $M' = E(M)$ which is δ-admissible and ε-clear w.r.t. to A, for some $\delta, \varepsilon > 0$.

For instance, according to Guidotti et al., possible sub-symbolic classifiers are neural (possibly deep) networks, support vector machines, and random forests.

Conversely, explanation strategies may consist of algorithms aimed at *(i)* extracting decision trees/rules out of sub-symbolic predictors and the data they have been trained upon, *(ii)* compute feature importance vectors, *(iii)* detecting saliency masks, *(iv)* detecting partial dependency plots, etc.

In our framework, all the algorithms mentioned above can be described as *explanation strategies*. Such mapping is plausible given their ability to compute an admissible, and possibly more explicit models out of black boxes and the data they have been trained upon. However, it is worth to highlight that the clarity gain produced by such explanation strategies mostly relies on the implicit assumption that their output models come with a natural representation which is intuitively interpretable to the human mind.

4.2 Outcome Explanation

Methods classified as part of the "outcome explanation" sub-category can be very naturally described in our framework as well. In fact, it can be defined as follows:

> Let M be some sub-symbolic classifier whose internal functioning representation $R = r(M, Z)$ in some subset $Z \subset \mathcal{X}$ of the input domain is poorly interpretable to some explanee A, and let $E(\cdot, \cdot)$ be some *local* explanation strategy. Then, the outcome explanation problem consists of computing some *local* explanation $M' = E(M, Z)$ which is δ-admissible and ε-clear w.r.t. to A, for some $\delta, \varepsilon > 0$.

Summarising, while input black boxes may still be classifiers of any sort, explanation, and explanation strategies differ from the "model explanation" case. In particular, explanation strategies in this sub-category may rely on techniques leveraging on attention models, decision trees/rules extraction, or well-established algorithms such as LIME [20], and its extensions—which are essentially aimed at estimating the contribution of every input feature of the input domain to the particular outcome of the black box to be explained.

Notice that the explanation strategies in this category are only required to be admissible and clear in the portion of the input space surrounding the input data under study. Such a portion is implicitly assumed to be relatively small in most cases. Furthermore, the explanation strategy is less constrained than in the global case, as it is not required to produce explanations elsewhere.

4.3 Model Inspection

Methods classified as part of the "model inspection" sub-category can be naturally defined as follows:

> Let M be a sub-symbolic classifier whose available *global* representation $R = r(M)$ is poorly interpretable to some explanee A, and let $r(\cdot), r'(\cdot)$ be two different representation means. Then, the model inspection problem consists of computing some representation $R' = r'(M)$ such that $I_A(R') > I_A(R)$.

Of course, solutions to the model inspection problem vary a lot depending on which specific representation means $r(\cdot)$ is exploited by the explanator, other than the nature of the data the black box is trained upon. Guidotti et al. also provide a nice overview of the several sorts of representations means which may be useful to tackle the model inspection problem, like, for instance, sensitivity analysis, partial dependency plots, activation maximization images, tree visualisation, etc.

It is worth pointing out the capability of our framework to reveal the actual nature of the inspection problem. Indeed, it clearly shows how this is the first problem among the ones presented so far, which only relies on the interpretation axis alone to provide understandability.

4.4 Transparent Box Design

Finally, methods classified as part of the "transparent box design" sub-category can be naturally defined as follows:

> Let $X \subseteq \mathcal{X}$ be a dataset from some input domain $\mathcal{X}$, let $r(\cdot)$ be a representation means, and let A be the explanee agent. Then the transparent box design problem consists of computing a classifier M for which a global representation $R = r(M, X)$ exists such that $I_A(R) > 1 - \delta$, for some $\delta > 0$.

Although very simple, the transparent-box design is of paramount importance in XAI systems as it is the basic brick of most general explanation strategies. Indeed, it may be implicit in the functioning of some explanation strategy E to be adopted in some other model or outcome explanation problem.

For instance, consider the case of a local explanation strategy $E(M, X) \mapsto M'$. In the general case, to compute M', it relies on some input data X and the internal of the to-be-explained model M. However, there may be cases where the actual internal of M are not considered by the particular logic adopted by E. Instead, in such cases, E may only rely on X and the outcomes of M, which are $Y = M(X)$. In this case, the explanation strategy E is said *pedagogical*—whereas in the general case it is said *decompositional* (cf. [1]).

In other words, as made evident by our framework, the pedagogical methods exploited to deal with the model or outcome explanation problems must internally solve the transparent box design problem, as they must build an interpretable model out of some sampled data-set and nothing more.

5 Towards the Social Dimension of Explainability

In previous sections, we mostly focus on understandability from the single-agent perspective. Conversely, in this section we move from the *intra*-agent perspective – relying on the framework presented in Sect. 3.1 – to the *inter*-agent one—where two or more interacting agents are involved [8].

Our discussion stems from the observation that the agent extracting/eliciting information In other words, no agent explains something to itself. Furthermore, in a multi-agent setup, it is plausible to have agents characterised by heterogeneous (potentially exclusive) capabilities and knowledge bases. In this situation, transferring knowledge and demanding for explanations may not even be possible without a *social* connotation. Indeed, the social, interactive dimension of understandability is well recognised (e.g., in the social sciences), and some authors are already suggesting the XAI research community should take it into account [17]. Accordingly, we argue that our framework should be extended in this direction.

In particular, we envision two main actors

explanator—formulating and sharing an explanation, and the
explanee—consuming and possibly demanding the explanation

needing to establish a *mutual-understanding.* The explanator and the explanee can be a software agent, a human, or a grouped combination of them.

The possible social scenarios to share explanations can be generalised in 1-to-1, 1-to-n, and m-to-n. Thus, the framework presented in Sect. 3.1 – which mostly focuses on the single-agent perspective – needs further extensions to tackle the challenge of understandability in a multi-agent scenario.

Mutual understanding is not just an algorithm, nor is it some cognitive activity that an agent can perform by itself; it is instead a formalised protocol involving two or more parties. Therefore, aiming at scaling our framework to the MAS setup, we envision the following behaviours to be modelled.

As the interpretation function is subjective by construction, a piece of given information can be considered interpretable by agent A but not by another agent B. Consequently, if agent A is willing to make a model X understandable by another agent B, a joint *agreement* about the representation of the explanation has to be established. We define *mutual understanding* as a request-response protocol involving at least one agent acting as *explanator* and one agent acting as *explanee*—both either virtual or humans. Such an agreement may involve the establishment of a common taxonomy and knowledge reconciliation [14,23].

The protocol can begin with the explanator taking the initiative to share an explanation or with an explainee requiring it. The object of the explanation is the desire to understand the behaviour of a given model M w.r.t some data X—which is naturally represented through $R = r(M, X)$. Assuming that the explanator can rely on a wider dataset $X' \supseteq X$ than the one the explanee is relying upon (i.e., X), it may respond in several ways:

- it may produce an alternative representation $R' = r'(M, X')$ of M on some data $X' \supseteq X$, expecting that R' may result more interpretable than R in the eyes of the explanee
- it may produce an explanation for M in X' by leveraging on some internal strategy E, hoping that the natural representation $R'' = r''(E(M), X')$ of $E(M)$ in X' may result more interpretable than R in the eyes of the explanee

In turn, the explanee may provide feedback based on its subjective interpretation of the proposed representation. The protocol may thus go through one or more request-response rounds. The object of the further iteration(s) can be *(i)* a specific component of the explanation – possibly demanding for a new level of granularity of the explanation – or *(ii)* the entire explanation that might need a complete rehearsal to be eventually understood. To prevent possible endless *(diverging)* explanations, we have to discriminate their underlying scenario. E.g.:

1-to-1—Once reached the most granular representation of an information, the agent say "no more additional information are available" concluding the iterations and declaring the *failure* of the explanation;

1-to-*n*—In case of misalignment on the understanding of a given explanation, techniques from defeasible reasoning [11] might be exploited to avoid the failure of the explanation;

m*-to-*n—Likewise the previous scenario, it is envisioned to possibly implement defeasible reasoning. Moreover, mechanisms enabling explanation-support among the n explanator might be developed to overcome the *failure* for lack of specification.

Another factor raising the complexity of the *mutual-understanding* is the possible heterogeneous composition of the explanator(s) or explanee(s) (e.g., a composition of both virtual and humans actors). A possible solution might be to generate clusters (e.g., sub-pools of explanators and explainees) and generate reconciled and personalised explanations. Including the human factor in the social explainability demands to consider elements such as *expectations*, *trust*, *State of Mind* (SoM), *emotions* and multi-modal *formats* of the explanation (e.g., natural language and graphical).

Finally, it is worth to be mentioned that the idea of leveraging on interaction to reach mutual understanding shares some similarities with several works from the planning literature, such as [6,7], For instance, in [7], agents support humans' understanding via *model reconciliation*, that is, a corpus of methods aimed at letting a human receive explanations w.r.t. the sequence of actions computed by a planning agent. In particular, such methods *(i)* define explanation in a planning-specific way, and *(ii)* involve interaction among the human (explanee) and the agent (explanator). However, despite some common insights, we argue our framework is original w.r.t the area of explainable planning. Indeed, whereas works in this area mostly focus on planning – which is an important subset of *symbolic* AI – our work mostly focuses on *sub-symbolic* AI—a difference which heavily affects how understandability is defined and pursued. Furthermore, while other works target scenarios involving both humans and software agents, we explicitly target both this case and the agents-only one.

6 Conclusion

Despite the many efforts of the XAI community in addressing *opacity* issues in ML-based intelligent systems, most works in this area still rely on natural-language-based definitions of fundamental concepts such as *explanation* and *interpretation*. Accordingly, in this work, we firstly explore the inconsistencies still affecting the definitions of interpretability and explainability in some recent impactful papers. Then, to overcome the limitations of natural language definitions, we propose an abstract framework for XAI deeply rooted in the MAS mindset—which is the main contribution of this paper. To assess the proposed framework, we compare it against existing studies in the field of XAI, showing how it can naturally and unambiguously provide clear definitions for the main sorts of tasks laying under the XAI umbrella. Finally, we propose some ways to scale the intra-agent to the inter-agent explainability and elaborate on the potential social implications characterising the dynamics among the agents.

References

1. Andrews, R., Diederich, J., Tickle, A.B.: Survey and critique of techniques for extracting rules from trained artificial neural networks. Knowl.-Based Syst. **8**(6), 373–389 (1995). https://doi.org/10.1016/0950-7051(96)81920-4
2. Anjomshoae, S., Najjar, A., Calvaresi, D., Främling, K.: Explainable agents and robots: results from a systematic literature review. In: Proceedings of the 18th International Conference on Autonomous Agents and Multi-Agent Systems, pp. 1078–1088. International Foundation for Autonomous Agents and Multiagent Systems (2019)
3. Besold, T.R., Uckelman, S.L.: The what, the why, and the how of artificial explanations in automated decision-making, pp. 1–20. CoRR abs/1808.07074 (2018)
4. Calegari, R., Ciatto, G., Dellaluce, J., Omicini, A.: Interpretable narrative explanation for ML predictors with LP: a case study for XAI. In: Bergenti, F., Monica, S. (eds.) WOA 2019–20th Workshop "From Objects to Agents", CEUR Workshop Proceedings, vol. 2404, pp. 105–112. Sun SITE Central Europe, RWTH Aachen University, Parma, 26–28 June 2019. http://ceur-ws.org/Vol-2404/paper16.pdf
5. Calvaresi, D., Najjar, A., Schumacher, M., Främling, K. (eds.): EXTRAAMAS 2019. LNCS (LNAI), vol. 11763. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-30391-4
6. Chakraborti, T., Sreedharan, S., Kambhampati, S.: Balancing explicability and explanation in human-aware planning (2017). https://arxiv.org/abs/1708.00543
7. Chakraborti, T., Sreedharan, S., Zhang, Y., Kambhampati, S.: Plan explanations as model reconciliation: moving beyond explanation as soliloquy. In: 26th International Joint Conference on Artificial Intelligence (IJCAI 2017), pp. 156–163. AAAI Press, Melbourne (2017). https://doi.org/10.24963/ijcai.2017/23
8. Ciatto, G., Calegari, R., Omicini, A., Calvaresi, D.: Towards XMAS: eXplainability through multi-agent systems. In: Savaglio, C., Fortino, G., Ciatto, G., Omicini, A. (eds.) AI&IoT 2019 - Artificial Intelligence and Internet of Things 2019. CEUR Workshop Proceedings, vol. 2502, pp. 40–53. Sun SITE Central Europe, RWTH Aachen University, November 2019
9. Ciatto, G., Calvaresi, D., Schumacher, M.I., Omicini, A.: An abstract framework for agent-based explanations in AI. In: 19th Interational Conference on Autonomous Agents and Multi-Agent Systems (AAMAS 2020). IFAAMAS, Auckland (2020)
10. Doshi-Velez, F., Kim, B.: Towards a rigorous science of interpretable machine learning. CoRR abs/1702.08608 (2017)
11. García, A.J., Simari, G.R.: Defeasible logic programming: an argumentative approach. Theor. Pract. Log. Prog. **4**(2), 95–138 (2004). https://doi.org/10.1017/S1471068403001674

12. Goodman, B., Flaxman, S.: European Union regulations on algorithmic decision-making and a “right to explanation”. AI Mag. **38**(3), 50–57 (2017). https://doi.org/10.1609/aimag.v38i3.2741
13. Guidotti, R., Monreale, A., Turini, F., Pedreschi, D., Giannotti, F.: A survey of methods for explaining black box models. ACM Comput. Surv. **51**(5), 1–42 (2019). https://doi.org/10.1145/3236009
14. Katarzyniak, R.P., Nguyen, N.T.: Reconciling inconsistent profiles of agents' knowledge states in distributed multiagent systems using consensus methods. Syst. Sci. **26**(4), 93–119 (2000)
15. Lipton, Z.C.: The mythos of model interpretability. Commun. ACM **61**(10), 36–43 (2018). https://doi.org/10.1145/3233231
16. Miller, G.A.: The magical number seven, plus or minus two: some limits on our capacity for processing information. Psychol. Rev. **63**(2), 81–97 (1956). https://doi.org/10.1037/h0043158
17. Miller, T.: Explanation in artificial intelligence: insights from the social sciences. Artif. Intell. **267**, 1–38 (2019). https://doi.org/10.1016/j.artint.2018.07.007
18. Omicini, A., Zambonelli, F.: MAS as complex systems: a view on the role of declarative approaches. In: Leite, J., Omicini, A., Sterling, L., Torroni, P. (eds.) DALT 2003. LNCS (LNAI), vol. 2990, pp. 1–16. Springer, Heidelberg (2004). https://doi.org/10.1007/978-3-540-25932-9_1
19. Premack, D., Woodruff, G.: Does the chimpanzee have a theory of mind? Behav. Brain Sci. **1**(4), 515–526 (1978). https://doi.org/10.1017/S0140525X00076512
20. Ribeiro, M.T., Singh, S., Guestrin, C.: “Why should I trust you?”: explaining the predictions of any classifier. In: 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2016), pp. 1135–1144. ACM Press, San Francisco, 22–26 August 2016. https://doi.org/10.1145/2939672.2939778
21. Rosenfeld, A., Richardson, A.: Explainability in human–agent systems. Auton. Agent. Multi-Agent Syst. **33**(6), 673–705 (2019). https://doi.org/10.1007/s10458-019-09408-y
22. Rudin, C.: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat. Mach. Intell. **1**(5), 206–215 (2019). https://doi.org/10.1038/s42256-019-0048-x
23. Tamma, V., Bench-Capon, T.: A conceptual model to facilitate knowledge sharing in multi-agent systems. In: Ontologies in Agent Systems (OAS 2001). CEUR Workshop Proceedings, vol. 52, pp. 69–76 (2001). http://ceur-ws.org/Vol-52/oas01-tamma.pdf
24. Turek, M.: Explainable artificial intelligence (XAI). Funding Program DARPA-BAA-16-53, Defense Advanced Research Projects Agency (DARPA) (2016). http://www.darpa.mil/program/explainable-artificial-intelligence
25. Voigt, P., von dem Bussche, A.: The EU General Data Protection Regulation (GDPR). LNCS (LNAI). Springer, Cham (2017). https://doi.org/10.1007/978-3-319-57959-7