An Abstract Framework for Agent-Based Explanations in AI

Extended Abstract

Giovanni Ciatto
University of Bologna
Cesena, Italy
giovanni.ciatto@unibo.it

Davide Calvaresi
HES-SO Valais
Sierre, Switzerland
davide.calvaresi@hevs.ch

Michael I. Schumacher
HES-SO Valais
Sierre, Switzerland
michael.schumacher@hevs.ch

Andrea Omicini
University of Bologna
Cesena, Italy
andrea.omiciniunibo.it

ABSTRACT

We propose an abstract framework for XAI based on MAS encompassing the main definitions and results from the literature, focussing on the key notions of *interpretation* and *explanation*.

KEYWORDS

XAI; Multi-Agent Systems; Abstract Framework

ACM Reference Format:
Giovanni Ciatto, Davide Calvaresi, Michael I. Schumacher, and Andrea Omicini. 2020. An Abstract Framework for Agent-Based Explanations in AI. In *Proc. of the 19th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2020), B. An, N. Yorke-Smith, A. El Fallah Seghrouchni, G. Sukthankar (eds.), Auckland, New Zealand, May 2020,* IFAAMAS, 3 pages.

1 INTRODUCTION

The adoption of intelligent systems (IS) in modern society is booming, mostly due to the recent momentum gained by Machine Learning (ML). As in the previous AI springs, expectations are being inflated by the promising predictive capabilities showed by ML-based IS. However, researchers and stakeholders are experiencing problems stemming from the *opacity* of ML-based solutions [1, 2, 8, 13].

The opacity of numeric, ML-powered, predictors is a broadly acknowledged issue [8, 10]. Nowadays, however, mostly due to the unprecedented pace and extent of ML adoption in many critical domains, addressing the issue of opacity is more needed than ever.

The opaqueness of ML-based solutions is an unacceptable condition in a world where ML is involved in many (safety-)critical activities. Indeed, performing good automatic predictions is as essential as letting humans involved in those contexts *understand* the rationale behind such predictions [3]. This is required, for instance, by modern regulations, which "start to" recognise citizens' right to receive meaningful explanations when automated decisions may affect their lives [4, 7]. Thus, the problem of understanding ML results is rapidly gaining momentum in recent AI research [5].

The topic of understandability in AI is nowadays the primary concern of the *eXplainable AI* (XAI) community—whose name is due to a successful project of DARPA [9]. In this paper, we argue that a fundamental issue flaws most studies in this field: namely, the lack of an unambiguous definition for the concept of *explanation*—and, consequently, of a clear understanding of what X in XAI actually means. Indeed, the notion of explanation is not explicitly established into the literature, nor is there a consensus on what the property named "explainability" should imply. Similar issues exist as far as the notion of *interpretation* is concerned. The two terms are sometimes used interchangeably into the literature, whereas other times they carry different meanings [10]. 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 [11] – and of socio-technical systems in particular –, they can be exploited to define a sound and unambiguous reference framework for XAI. Accordingly, we propose an abstract framework for XAI relying on notions and results from the MAS literature.

Proc. of the 19th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2020), B. An, N. Yorke-Smith, A. El Fallah Seghrouchni, G. Sukthankar (eds.), , May 2020, Auckland, New Zealand.

2 EXPLANATION *VS.* INTERPRETATION

Inspired by the field of Logic, 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 [6]: that 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 by "easy" we mean A requires a low *computational* (or *cognitive*) effort.

This is modelled by a function $I_A(X) \mapsto [0, 1]$ providing (a degree of) *interpretability* 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 is inaccessible in most cases. This is far from being an issue, since we are interested not in the absolute value of $I_A(X)$, for some object X, but rather 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. We stress the subjective nature of interpretations, as agents assign them to objects according to their State of Mind (SoM) [12] and background knowledge, and they need to not be formally defined any further.

Conversely, 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, in such a way that $I_A(X') > I_A(X)$, for some agent A.

2.1 A conceptual framework for XAI

Several AI 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 $\mathcal{Y}$.

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 in 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.

Let $M, M' \in \mathcal{M}$ be two analogous models. We 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, $\Delta f : 2^{\mathcal{Y}} \times 2^{\mathcal{Y}} \rightarrow \mathbb{R}_{\geq 0}$ measures the performance *difference* between two analogous models.

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 case M is a classifier, R may 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. So, providing an 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.

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 the M' is produced through the explanation strategy E. Furthermore, we say an explanation M' is *admissible* if it has a good fidelity w.r.t. the original model M and 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 $\Delta f(M(Z), M'(Z)) < \delta$. Finally, we say an explanation M' is *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 be actually 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 make the search for more interpretable representations easy. 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.

The theoretical framework described so far aims at modelling local interpretations and explanations—as the two means an explanator agent may use in order to make AI tasks' *outcomes* more understandable to some explanee. However, when the goal is not to understand some model outcome but the model itself, from a *global* perspective, the framework is simplified by considering $Z \equiv X$.

2.2 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. Furthermore, our framework explicitly recognises the *subjective* nature of interpretation, as well as the *objective* nature of explanation. Indeed, interpretation is a subjective activity directly related to agents' perception and SoM, whereas explanation is an epistemic, computational action aimed at producing a high-fidelity model. Our framework also captures the importance of representations. This is yet another degree of freedom that agents may exploit in their search for a wider understandability of a given model. Finally, our framework acknowledges the global/local duality of both explanation and interpretation, thus enabling AI models to be understood either in general or with respect to a particular input/output pair.

2.3 Practical remarks

According to our conceptual framework, a *rational* agent trying 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).

The nature of the model actually constrains the set of admissible representations. We argue that each family of AI models comes with just a few *natural* representations. In real-world scenarios, then, an agent looking for understandability could be expected to "work" on both the interpretation and the explanation axes.

Another features of our framework concerns the semantics of clear explanations. The current definition simply requires explanation strategies to consume a model M with a given representation R and to produce an 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. Similarly, in some cases, it may be enough – other than more feasible – to find an *admissible* explanation—that is, an high-fidelity model for which *some* representation exists that is more interpretable than *some* representation of the original model.

3 CONCLUSION

Intelligent systems adopting machine learning techniques are increasingly pervading our everyday lives. The ever-increasing (and sometimes indiscriminate) adoption of ML-based approaches generates the impelling need to understand the no longer acceptable *opacity* of systems. Besides the efforts of the XAI community in addressing such issues, most works in this area tend to rely on natural-language-based definitions of fundamental concepts such as *explanation* and *interpretation*. In this work, we first explore the inconsistencies still affecting the definitions of interpretability and explainability in some recent impactful papers. Then, in order to overcome the classical limitations of natural language definitions, we propose an abstract and formal framework for XAI deeply rooted in the MAS mindset.

REFERENCES

[1] Sule Anjomshoae, Amro Najjar, Davide Calvaresi, and Kary Främling. 2019. Explainable agents and robots: Results from a systematic literature review. In *Proceedings of the 18th International Conference on Autonomous Agents and Multi-Agent Systems*. International Foundation for Autonomous Agents and Multiagent Systems, ACM, San Francisco, CA, USA, 1078–1088.

[2] Tarek R. Besold and Sara L. Uckelman. 2018. The What, the Why, and the How of Artificial Explanations in Automated Decision-Making. *CoRR* abs/1808.07074 (2018), 1–20. arXiv:1808.07074 http://arxiv.org/abs/1808.07074

[3] Roberta Calegari, Giovanni Ciatto, Jason Dellaluce, and Andrea Omicini. 2019. Interpretable Narrative Explanation for ML Predictors with LP: A Case Study for XAI. In *WOA 2019 – 20th Workshop "From Objects to Agents"*, Federico Bergenti and Stefania Monica (Eds.). CEUR Workshop Proceedings, Vol. 2404. Sun SITE Central Europe, RWTH Aachen University, 105–112. http://ceur-ws.org/Vol-2404/paper16.pdf

[4] Davide Calvaresi, Yazan Mualla, Amro Najjar, Stéphane Galland, and Michael Schumacher. 2019. Explainable Multi-Agent Systems Through Blockchain Technology. In *Explainable, Transparent Autonomous Agents and Multi-Agent Systems - First International Workshop, EXTRAAMAS 2019, Montreal, QC, Canada, May 13-14, 2019, Revised Selected Papers*. Springer, Berlin Heidelberg, 41–58. https://doi.org/10.1007/978-3-030-30391-4_3

[5] Giovanni Ciatto, Roberta Calegari, Andrea Omicini, and Davide Calvaresi. 2019. Towards XMAS: eXplainability through Multi-Agent Systems. In *Proceedings of the 1st Workshop on Artificial Intelligence and Internet of Things co-located with the 18th International Conference of the Italian Association for Artificial Intelligence (AI*IA 2019), Rende (CS), Italy, November 22, 2019 (CEUR Worhsop Proceedings)*. Sun SITE Central Europe, RWTH Aachen University, 40–53. http://ceur-ws.org/Vol-2502/paper3.pdf

[6] Giovanni Ciatto, Roberta Calegari, Andrea Omicini, and Davide Calvaresi. 2019. Towards XMAS: eXplainability through Multi-Agent Systems. In *AI&IoT 2019 – Artificial Intelligence and Internet of Things 2019*, Claudio Savaglio, Giancarlo Fortino, Giovanni Ciatto, and Andrea Omicini (Eds.). CEUR Workshop Proceedings, Vol. 2502. Sun SITE Central Europe, RWTH Aachen University, 40–53.

[7] Bryce Goodman and Seth Flaxman. 2017. European Union Regulations on Algorithmic Decision-Making and a "Right to Explanation". *AI Magazine* 38, 3 (2017), 50–57. https://doi.org/10.1609/aimag.v38i3.2741

[8] Riccardo Guidotti, Anna Monreale, Franco Turini, Dino Pedreschi, and Fosca Giannotti. 2019. A Survey of Methods for Explaining Black Box Models. *ACM Computing Surveys (CSUR)* 51, 5, Article 93 (Jan. 2019), 42 pages. https://doi.org/10.1145/3236009

[9] David Gunning. 2016. *Explainable artificial intelligence (XAI)*. Funding Program DARPA-BAA-16-53. Defense Advanced Research Projects Agency (DARPA). http://www.darpa.mil/program/explainable-artificial-intelligence

[10] Zachary Chase Lipton. 2018. The Mythos of Model Interpretability. *ACM Queue* 16, 3, Article 2 (May–June 2018), 27 pages. https://dl.acm.org/citation.cfm?id=3241340

[11] Andrea Omicini and Franco Zambonelli. 2004. MAS as Complex Systems: A View on the Role of Declarative Approaches. In *Declarative Agent Languages and Technologies*, João Alexandre Leite, Andrea Omicini, Leon Sterling, and Paolo Torroni (Eds.). Lecture Notes in Computer Science, Vol. 2990. Springer, Berlin Heidelberg, 1–17. https://doi.org/10.1007/978-3-540-25932-9_1 1st International Workshop (DALT 2003), Melbourne, Australia, 15 July 2003. Revised Selected and Invited Papers.

[12] David Premack and Guy Woodruff. 1978. Does the chimpanzee have a theory of mind? *Behavioral and brain sciences* 1, 4 (1978), 515–526. https://doi.org/10.1017/S0140525X00076512

[13] Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'16)*, Vol. abs/1602.04938. ACM, San Francisco, CA, USA, 1135–1144. https://doi.org/10.1145/2939672.2939778