Towards XMAS: eXplainability through Multi-Agent Systems

Giovanni Ciatto[1][0000−0002−1841−8996], Roberta Calegari[1][0000−0002−6655−3869], Andrea Omicini[1][0000−0002−6655−3869], and Davide Calvaresi[2][0000−0001−9816−7439]

[1] University of Bologna, Italy
{giovanni.ciatto, roberta.calegari, andrea.omicini}@unibo.it
[2] University of Applied Sciences and Arts Western Switzerland, Switzerland
davide.calvaresi@hevs.ch

Abstract. In the context of the Internet of Things (IoT), intelligent systems (IS) are increasingly relying on Machine Learning (ML) techniques. Given the opaqueness of most ML techniques, however, humans have to rely on their intuition to fully understand the IS outcomes: helping them is the target of eXplainable Artificial Intelligence (XAI). Current solutions – mostly too specific, and simply aimed at making ML easier to interpret – cannot satisfy the needs of IoT, characterised by heterogeneous stimuli, devices, and data-types concurring in the composition of complex information structures. Moreover, Multi-Agent Systems (MAS) achievements and advancements are most often ignored, even when they could bring about key features like explainability and trustworthiness. Accordingly, in this paper we *(i)* elicit and discuss the most significant issues affecting modern IS, and *(ii)* devise the main elements and related interconnections paving the way towards reconciling interpretable and explainable IS using MAS.

Keywords: MAS · XMAS · XAI · explainability · road map

1 Introduction

In the current decade, Internet of Things (IoT) systems, devices, and frameworks boomed, demanding contributions from both industry and academia. People's daily lives are getting entangled with uncountable cyber-physical devices capable of sensing, acting, and reasoning about the surrounding environment and who populates it. This leads to an intriguing set of socio-technical challenges for the Artificial Intelligence (AI) researchers and practitioners. The complexity of the IoT systems is increasing at a fast pace, employing underlying AI techniques such as Machine Learning (ML) in the system core mechanisms to analyse, combine, and profile heterogeneous sets of data. For instance, virtual assistants such as Alexa, Cortana, Siri, or Google Home [19] exploit ML to improve a seamless vocal interaction and refined recommendation systems; Nest, the smart thermometer from Google, uses ML to learns from the user's habits. Moreover, such devices can interact with each other, thus increasing the input data-types and the overall complexity that the system has to deal with. The effect of such a deep and pervasive adoption of IoT within the productive and service fabrics of human societies is that AI and ML are going to control – or at least affect – an ever increasing number of aspects of people's lives.

The benefits of such AI-powered evolution are remarkable. Industries can now reach a novel degree of automation, whereas customers can now enjoy a plethora of new services mediated by their devices, as data and services now can fuel unprecedented business opportunities and markets. However, such a transition is unlikely to occur without costs. Besides ethical and sociological issues, the current usage of AI is far from being socially acceptable. In particular, the recent hype on ML, Deep Learning (DL), and other numeric AI methods – commonly referred as "third AI spring" – has led to a situation where several decisions are delegated to subsymbolic predictors out of human control and understanding—as demonstrated by the many cases where they blatantly misbehaved [11,15,9].

Furthermore, as broadly acknowledged by many research communities, we argue that the development process of current intelligent systems is flawed by the following issues:

Lack of generalisation — Most tasks in AI require very specific modelling, design, and development/training process. As a result, the integration, combination, and comparison of different – yet similar – methods in AI is either impossible or achieved through highly human-intense *ad hoc* (thus not scalable/extendable) system design.

Opaqueness — When numeric (data driven) methods are exploited, predictions come without an understandable motivation—or, more generally, without a model. Unfortunately, in most applications, data scientists only care about predictive performance and generalisation capability. However, the adoption of opaque methods or predictors reduces the scope of application of intelligent systems—possibly due to either *practical* or *legal* constraints, and, more concerning, to the lack of *trust* manifested by people and organisations.

Lack of automation (in the development process) — Despite AI is a tool ultimately adopted to seek automation, the training process of most numeric predictors is far from being automatic. The experience and the background of the data scientist are still heavy discriminants for the overall predictive performance. Methodologies and guidelines do not ensure any success in the general case, and a lot of *trials-and-errors* are typically unavoidable.

Centralisation of both data and computation: it is often required or preferred given the centralised nature of most algorithms or the legal constrains data is subject to. Centralisation poses severe technical limitations to the way data is managed, and to the design of computing systems.

Both the industry and the academia tend to tackle such problems individually, without looking at the whole picture. As a result, most of research activities focus on: *(i)* creating ad-hoc integration of AI sub-systems tailored on specific problems; *(ii)* developing techniques easing the interpretation of specific numeric predictors/predictions, exploiting results from the eXplainable AI (XAI) research area [18]; *(iii)* improving AI/ML performances in specific problems; *(iv)* setting up custom parallel or distributed implementations of specific AI methods—which may easily result in overly complicated solutions if legal constrains on data location have to be enacted. Accordingly, we believe that such trends actually slow down the identification of a general and comprehensive solution.

In this paper we claim that Multi-Agent Systems (MAS) [14] have the potential to provide a general – both conceptual and technical – framework to address most of the aforementioned issues. MAS are composed of several (possibly distributed) intelligent software (or cyber-physical) agents prone to automatic reasoning and symbolic manipulation. We argue that such agents can be employed to: *(i)* dynamically provide *interpretations* and *explanations* for opaque systems, *(ii)* ease the integration among different solutions/components for similar tasks or predictors, *(iii)* increase the degree of automation characterising the development of intelligent systems, and *(iv)* support the provisioning of machine intelligence even in those (possibly distributed and decentralised) contexts where data cannot be moved due to technical or legal limitations. Moreover, MAS can be fruitfully combined with methods for symbolic knowledge extraction (out of numeric predictors) [2,12], and Distributed Ledger Technologies (DLT, a.k.a. blockchains [4,10]).

Contribution. This paper drafts a long-term research line supporting our claim. In particular, we provide a i^* [27] formalisation of the foreseeable research activities and the dependencies among them.

The reminder of this paper is paper is organised as follows. Section 2 recalls the main research topics to be involved and combined in our research activity. Section 3 presents and discusses the aforementioned i^* modelling. Finally, Section 4 provides some remarks and concludes the paper.

2 State of the Art

2.1 eXplainable Artificial Intellingence (XAI)

Modern intelligent systems (IS) often leverage on *black-box* predictive models which are trained through ML or DL, or other *numeric* approaches. The "black-box" expression refers to models where knowledge is not explicitly represented [21]. As a consequence, humans have difficulties (or, have no way) to understand that knowledge, and why it led to suggest or undertake a given decision. Obviously, troubles in understanding black-boxes content and functioning prevents people from fully trusting – therefore accepting – them. In contexts such as medical or financial, having IS merely suggesting / taking decisions is not acceptable any more—e.g., due to ethical and legal concerns. For example, current regulations such as the GDPR [26], ACM Statement on Algorithmic Transparency and Accountability [1], and the European Union's "Right to Explanation" [16], demand IS to become *understandable* in the near future. In fact, 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. However, the notion of understandability is neither standardised nor systematically assessed, yet. The recently-emerged XAI research field is targeting such issues—e.g., DARPA has proposed a comprehensive research road map [18]. Research efforts in XAI focus on achieving key properties in AI such as *interpretability*, *algorithmic transparency*, *explainability*, *accountability*, and *trustworthiness* [3]. Unfortunately, such goals are still far from reach. One of the main reasons is the lack of a formal and agreed-upon definition of such concepts [21]. Moreover, most works only target classification problems, and they rarely take wider properties – such as accountability and trustworthiness – into account [17].

Interpretability vs. Explainability. In the context of XAI, the terms "explainability" and "interpretability" are too often used carelessly [17]. In particular, they are interchanged or just conveniently associated with in-house – misleading, often erroneous – definitions. Although they are closely related and both contributing to the ultimate goal of understandability, it is worth pointing out the differences in order to better comprehend our XMAS vision—where XMAS stands for *eXplainability through Multi-Agent Systems*.

On the one hand, we borrow the definition of "interpretation" from logic, where the word essentially describes the operation of binding objects to their actual meaning in some context. Thus, as far as numeric models are concerned, the goal of interpretability is to convey to humans the meaning hidden into the data and the mechanisms/decisions characterising the predictors.

On the other hand, we define "explanation" as the *act* making someone understand the information conveyed in a given discourse. It worth to highlight that the act of explaining is an *activity* involving at least two *interacting* parties, one *explaining* (explainer) and the other(s) willing to *understand* (explainee).

In the context of IS, the goal of explainability is to transfer to the receiver (possibly humans) given information (e.g., awareness of the *reasons* leading the system to act in a certain way) on a semantic level, aligning the State of Mind (SoM) [23] of the explainer and the explainee. The practice of explaining involves unveiling some background knowledge, or some latent information, that the explainee may not have "noticed" or explicitly required.

Such a distinction between interpretability and explainability is crucial since it shows how most XAI approaches proposed into the recent literature mostly focus on interpretability. Thus, while research into interpretable ML is widely recognised as important, a joint understanding of the concept of explainability still needs to evolve.

Symbolic AI for Explainability. XMAS targets both the aspects characterizing understandability. However, differently from other research lines branching from XAI, our vision poses a remarkable emphasis on explainability. In particular, in XMAS, we commit to *symbolic AI* as the main means for explainability.

By “symbolic AI” we mean the branch of AI focusing on symbolic knowledge representation, automatic reasoning, constraint satisfaction, and logic programming [24]. Such areas are deeply entangled with the results from computational logics [5], making their applications either inherently interpretable or easy to explain—given their lack of ambiguity and their underlying sound reasoning.

The reasons behind this commitment are threefold. Firstly, we let XMAS support the wide gamma of results, methods, algorithms, and toolkits developed under the umbrella of symbolic AI. Secondly, as further discussed in the following sections, we believe that the adoption of symbolic AI to be an enabling choice for the full exploitation of MAS. Finally, we argue that symbolic representations (e.g., the language of 1OL formulas), may act as a *lingua franca* for knowledge representation and exchange among heterogeneous IS.

In particular, the potential of logic-based models and their extensions is mainly due to their *declarativeness* and *explicit knowledge representation* – enabling knowledge sharing at an adequate level of abstraction – modularity, and separation of concerns [22]—which are especially valuable in open and dynamic distributed systems.

Symbolic knowledge extraction. The generality of symbolic approaches is also due to the many research works recently pointing out that the explanation capability of numeric predictors can be achieved via *symbolic knowledge extraction* (SKE) [12,6].

SKE groups methods and techniques for extracting symbolic representations of the numeric knowledge buried in data and captured through ML during predictors training. Indeed, one of the main issues in symbolic AI is that human experts must often handcraft symbolic knowledge relying on their background and experience. This is not what happens in ML, where useful – yet hard to interpret – numeric knowledge is mined from data. Therefore, SKE can enable the exploitation of both symbolic and numeric AI without their respective shortcomings. In turn, XMAS aims at leveraging on the symbolic knowledge extracted from ML-powered predictors as a basis for providing explanations of its predictions and functioning.

Many SKE techniques have been proposed in the literature. Some of them focus on specific sorts of ML predictors, such as neural networks – and they are therefore called “decompositional” –, whereas others are more general any may virtually target any predictor—and they are thus called “pedagogical”. Several relevant contributions to the topic are outlined in surveys such as [28,2].

2.2 Multi-agent systems (MAS)

Multi-agent systems (MAS) represent an extensive research area placed at the intersection between computer science, AI, and psychology, studying systems composed by interactive, autonomous, and usually intelligent entities called agents [14].

The agent abstraction has been described in many ways. However, most definitions agree on the following traits. *(i)* Agents are entities operating into an *environment* possibly perceived through sensors and affect through actuators. *(ii)* Agents are *autonomous* in the sense that have the capability of deciding on their own which actions are to be performed in order to achieve (or maintain) the goals they have been provided with—which in most cases are explicitly represented through some symbolic language. *(iii)* Agents are *social*, meaning that they can (and usually need to) interact (e.g., communicate, cooperate, and/or compete) with each other or with human users in their attempt to achieve/maintain their goals. *(iv)* Agents have a *mind* consisting of a belief base (BB)—storing symbolic data representing the (possibly wrong, steady, or biased) information each agent believes to know about the environment or other agents. The content of a given agent's BB is affected by its perceptions and interactions, and it may influence the its actions. The general notion of agent is so wide that both software entities and human beings may fit it. Such formal laxity is deliberate and useful. In fact, it allows human-machine and machine-machine interactions to be captured at the same level of abstraction and to be described through a coherent framework.

Dialogical argumentation. A central role in agent sociality is played by *argumentation* [13]: there, the emphasis is on the exchange of arguments and counter-arguments between agents, commonly aimed at making them able of reason and act intelligently even in presence of incomplete, inconsistent, biased, or partially wrong belief bases. Of course, the activity of argumentation involves a number capabilities – ranging from arguments mining or building to argument exchange in multi-party dialogues, and stepping through acceptability semantics –, and as many research lines.

Dialogical argumentation, in particular, is the activity performed by a number of agents dynamically *discussing* about some topic they are concerned with from different perspectives, in an attempt to agree on some shared truth about that topic. Thus, dialogical argumentation accounts for how arguments and counter-arguments are generated and evaluated, how the agents interact – i.e. what kinds of dialogical moves they can make –, and how agents can retract arguments, update beliefs, etc. Usually, it is set against *monological* argumentation [20], where the goal is two provide algorithms for computing which arguments are winning in a given setting, and which conclusions can be therefore drawn.

2.3 The i^* modelling language

Modelling a domain is a human-intensive, non-automated task. The domain IoT-powered IS is characterised, by itself, by complex requirements, theories, and methods converging from several scientific fields. Also, the XMAS vision intertwines results from disparate research areas, which historically has been kept mostly disjoint.

To generate a clear and structured understanding of our vision, we adopt the Goal-Oriented Requirement Engineering (GORE) approach [25], and in particular we exploit i^* as a modelling language [27]. i^* is a graphical language usually employed to model requirements for a single system. Nevertheless, it has been successfully employed to explore and map user needs and requirements for extensive application domains [7].

Here, an i^* model consists of a graph whose vertices are elements of four kinds: *Goals* (ranged by $\mathbf{G_i}$), *Soft Goals* (ranged by $\mathbf{SG_j}$), *Tasks* (ranged by $\mathbf{T_k}$), *Resources* (ranged by $\mathbf{R_l}$); edges (a.k.a. *links*) represent relations of various sorts among the aforementioned elements.

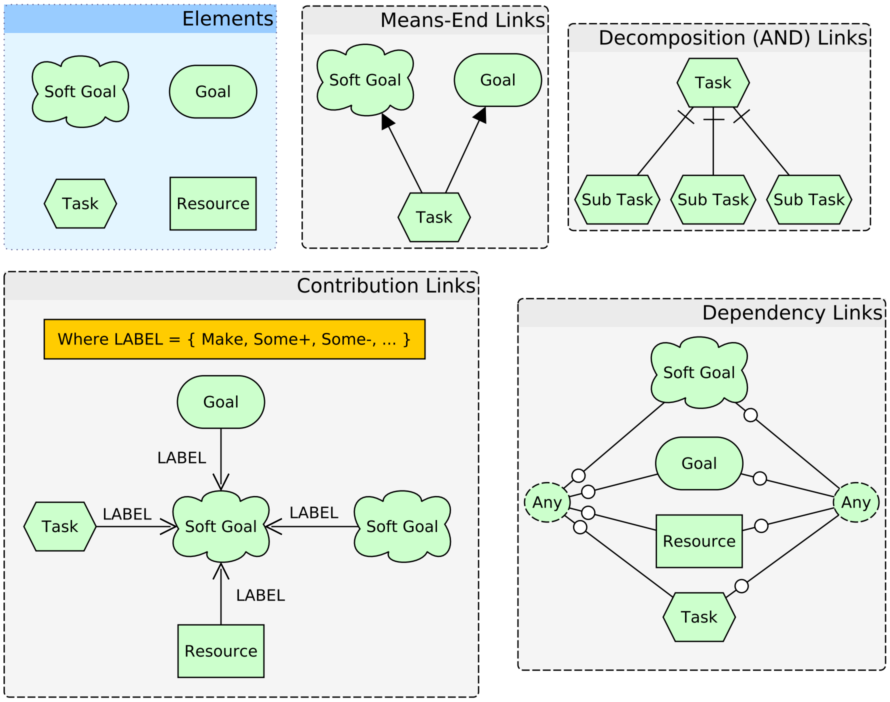

Fig. 1. i^* meta-model overview: elements and links

Figure 1 depicts how elements and links are graphically represented. A more complete description of the i^* graphical formalism can be found on the i^* wiki web page[3]. Informally, goals are desired properties or objectives whose achievement can either be satisfied or not, in a discrete fashion. Conversely, soft goals are (non-necessarily measurable) desirable properties or objectives which can be satisfied qualitatively or up to some degree. Tasks represent activities to be performed as an attempt to satisfy (resp. positively affect) one or more goals (resp. soft goals). Resources represent entities to be produced or consumed by tasks, and whose availability may favor or hinder the satisfaction of goals.

[3] `http://istar.rwth-aachen.de/tiki-index.php?page=iStarQuickGuide`

As far as links are concerned, soft goals are usually connected to each other via “contribution links”, which specify their contribution in fulfilling the needs—e.g., positive: `Some+`, or negative: `Some-`; goals are connected via “means-end” arrows; tasks are connected via “decomposition” links. Other sorts of links mostly define generic “dependencies”.

3 XMAS Vision: eXplainability through MAS

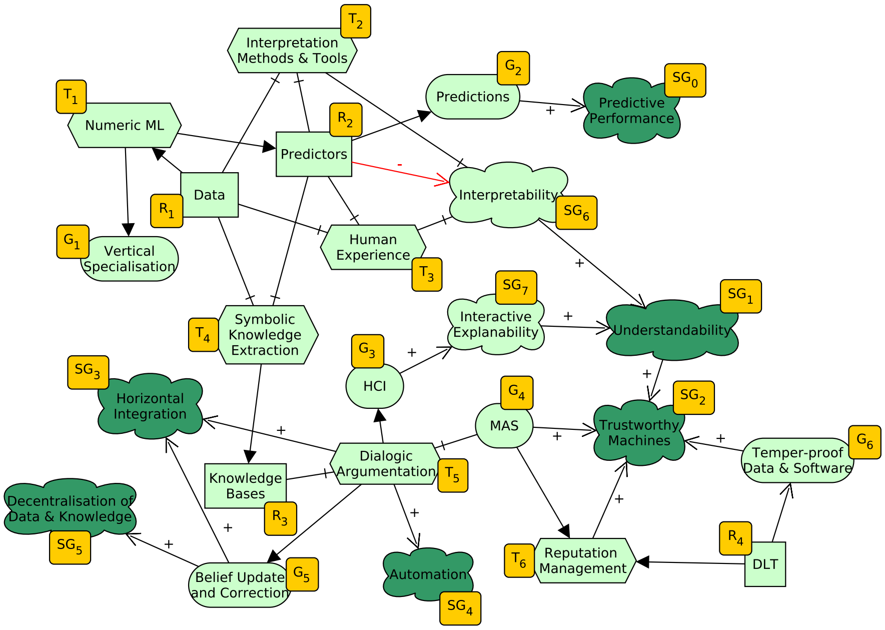

Fig. 2. XAI in IS: road-map and implications

Figure 2 graphically represents the *i** modelling of the XMAS vision, and highlights its main aspects. The representation aims at providing an intuitive graphical assessment of the various elements and their interconnections concurring in advancing state of the art of XAI in IS. We argue that having a overall mapping of XMAS requirements, objectives, as well as of their mutual interdependencies, can also facilitate the assessment, presentation, design, and implementation of a coherent research activity aimed at supporting our vision.

Figure 2 is composed of elements with diverse levels of abstraction. The most abstract elements are modelled as *soft goals*, representing most long-term objectives of the XMAS vision, which satisfaction is *not* measurable (e.g., the satisfaction of a soft-goal is not just binary, but has partial non-fully-assertable degrees). Conversely, goals are used to represent both the achievement of *measurable* results and the adoption of well-established frameworks, methods, and result from the literature. Finally, tasks are used to represent human or machine activities involving some work, whereas resources/model/entities having a physical form or a digital representation.

About the main soft goals. Our vision stems from the recognition that the success of IoT-powered IS is due to the general, increasing demand of analytical and *predictive performance* – corresponding to the $\mathbf{SG_0}$ weak goal in Figure 2 – transversally pervading the productive fabric of most developed societies. However, we also acknowledge several other *desiderata* – mostly targeting the issues highlighted in Section 1, and corresponding to as many weak goals into our i^* model – which are required to some extent by modern AI, mostly due to the increasingly pervasive adoption of AI techniques.

For instance:

- IS need to be *understandable* ($\mathbf{SG_1}$), meaning both interpretable and (above all) explainable, in the sense outlined in Section 2.
- Understandability, in turn, is only one of the aspects concurring in making modern cyber-physical systems perceived as *trustworthy* ($\mathbf{SG_2}$). Other aspects are important as well, like, egg,, having some degree of control on the behaviour of autonomous agents, and on the data and knowledge they are relying upon. In particular, when it comes to data and software, integrity and tampering-resistance are properties of paramount importance, strongly affecting the trustworthiness of IS.
- The IoT landscape also stresses the need to widen researchers' focus towards the "system of systems" dimension. The *vertical specialisation* ($\mathbf{G_1}$) of ML-based solution w.r.t. specific tasks is not the only concern any longer. Indeed, the *horizontal integration* ($\mathbf{SG_3}$) of heterogeneous systems is important as well. There, we expect IS to acquire the capability of dynamically and autonomously integrate, complement, and extend their respective knowledge, in a similar way to what human beings do when talking to each other.
- At the same time, the need for a higher *degree of automation* ($\mathbf{SG_4}$) in IS development is impelling, as the current bottleneck in the development of IS is due to the deep dependence of the process on human intuition.
- Finally, in spite of the many legal and ethical constraints affecting data and their usage, IoT-powered IS eventually need to overcome the current tendency to *data centralization* ($\mathbf{SG_5}$), as it imposes severe limitations over their effectiveness, efficiency, and adoption.

The XMAS vision pursues the goal of tackling – or at least improving – all such issues, and, ultimately, of making intelligent systems more trustworthy. To do so, we analysed the current trends in this area and understood that the combination of numeric methods with more classical symbolic AI approaches, possibly mediated by MAS, may provide beneficial effects at several levels.

On predictive capability. However, before describing how and why our proposal may provide an advantage, we need to briefly recall the most relevant aspects in IS development, as well as their mutual interdependencies.

In their quest for higher predictive performances ($\mathbf{SG_0}$), IS designers simply apply *numeric ML* ($\mathbf{T_1}$) methods to the *data* ($\mathbf{R_1}$) they have collected through IoT devices. Such a process is far from being automatic, as it requires the experience and the trial-and-error work of well-prepared data-scientists. It is aimed at the creation of *predictors* ($\mathbf{R_2}$), which are mathematical models capable of providing numerical *predictions* ($\mathbf{G_2}$) in situations analogous to the ones described by the data they have been trained with.

Nevertheless, even after the deployment of the IS, its *vertical specialisation* ($\mathbf{G_1}$) on the specific problem described by data remains an ever-lasting process, as new data keeps to be produced/captured by the systems in production. Most researchers or data scientists prefer to focus on such sorts of tasks as they are very valuable and numerically quantifiable—in terms of predictive performance.

On understandability. As far as the soft goal of understandability ($\mathbf{SG_1}$) is concerned, we argue that it can be tackled either by easing predictors/data *interpretability* ($\mathbf{SG_6}$) or by providing means for their *explanability* ($\mathbf{SG_7}$).

When it comes to let humans *interpret* the predictions provided by ML-powered predictors, the most common way to do so is to employ the most adequate technique for the problem at hand, possibly mediated by some *analytical* or *visualisation toolkit* ($\mathbf{T_2}$) and let the *human intuition* of experts ($\mathbf{T_3}$) do the magic. Thus, despite being very effective in specific cases, such an approach lacks generality and hinders automation.

Conversely, when it comes to providing *explanations* to the users, the XMAS vision recognises the prominent role of *interaction* in letting knowledge be transferred from IS to humans (and possibly vice versa). In particular, we envision a scenario where intelligent agents exchange symbolic knowledge with humans through various channels, interfaces, and languages—i.e., we envision several possible means for *Human-Computer Interaction* (HCI, $\mathbf{G_3}$).

As a first step in this direction, XMAS leverages on *symbolic knowledge extraction* (SKE, $\mathbf{T_4}$) as a means for attaining logic-based, *symbolic rules* and *facts* ($\mathbf{R_3}$) out of numeric predictors ($\mathbf{R_2}$) and raw data ($\mathbf{R_1}$).

The next step consists of employing such symbolic rules and facts as knowledge bases for cognitive, distributed *agents* ($\mathbf{G_4}$). More precisely, we state a one-to-one correspondence among numeric predictors and the agents to be deployed. Thus, we say that each predictor is *wrapped* by an agent.

Such a wrapping is an enabling step in several directions. For instance, we expect that by employing *dialogical argumentation* ($\mathbf{T_5}$), cognitive agents may become able to compare and complement the knowledge they have extracted from numeric predictors.

The capability of knowledge revision is particularly interesting, especially if one of the agents involved in the argumentation process is a human being. Indeed, if adequately constrained, dialogical argumentation may act as a means for providing *interactive explanations* ($\mathbf{SG_7}$) to the users, concerning the symbolic knowledge wrapped by agents.

In particular, such explanations can be even more effective if the interaction among users and software is mediated by some textual, vocal, or avatar-based user interface aimed at easing *human-computer interaction* ($\mathbf{G_3}$).

On the benefits of argumentation. However, the adoption of extracted symbolic knowledge and dialogic argumentation is not merely aimed at supporting the explanations.

Instead, it may also positively affect what we call the *horizontal integration* ($\mathbf{SG_3}$) of heterogeneous IS attained by different – yet related – data. This, in turn, enables the integration and exploitation of different perspectives on the information carried by data—which implies that different points of view can be merged to more precise predictions, as well as alternative predictive scenarios can be produced. Horizontal integration could thus make (more) valuable the many degrees of freedom and the inherent randomness characterising the processes of data retrieval, selection, engineering, partitioning, and analysis.

At the same time, the agents' capability of mutually *updating* and *correcting* their belief bases ($\mathbf{G_5}$) may pave the way towards the development of IS where predictions can be attained without relying on the centralization of data on a specific computational facility, nor on its transfer outside the organizational domain it belongs to. In other words, XMAS enables the decentralization of knowledge and computation ($\mathbf{SG_5}$). Despite data being usually subject to strict regulations limiting – among the others – its transfer, this is possible because aggregated – thus anonymous – data, such as the high-level rules extracted from data or predictors, are subject to less limiting regulations.

Similarly, argumentation may be conceived as a means for supporting a higher degree of automation ($\mathbf{SG_4}$) in the development of IS. In particular, protocols could be defined, letting new agents query other agents for symbolic knowledge they do not have. By doing so, cognitive agents can learn predictive or explanatory rules *autonomously*, even without needing direct access to the data.

On trustworthiness. If the XMAS vision will be accomplished, the effect of a handcrafted malicious (or buggy) agent, deliberately or mistakenly attempting to inject wrong knowledge into an agent society could be nefarious—and, by assuming an open and distributed society such as the IoT, this contingency cannot be excluded. This is another critical issue preventing people from fully trusting IS nowadays.

To mitigate such concerns, *DLT* ($\mathbf{R_4}$) could be exploited to prevent the *tampering of data or software* ($\mathbf{G_6}$), or, to keep track of agents' reputation—assuming some *reputation-enforcing protocol* ($\mathbf{T_6}$) [8] is enacted by the agent society.

4 Conclusion

In this paper we point out a number of issues affecting modern IoT, and in general distributed IS whose intelligence leverages on ML. In particular, focusing on the data analytics layer of most IoT-based applications, we argue that a number of issues are still far from being completely closed. For instance, we discuss why most ML-powered IS lack transparency, automation (in the development process), and decentralisation (of both data and computation).

Elaborating on such open issues, we discuss a research line – called *eXplainability through Multi-Agent Systems* (XMAS) – aimed at addressing them altogether in a coherent and effective way. In the XMAS vision, we plan to integrate a number of contributions from the symbolic AI, MAS, and XAI research areas.

Accordingly, in this paper we provide an overview of the state of the art of the aforementioned areas, shortly discuss their main achievement and limitations in the XMAS perspective, and present a formal model of the XMAS vision using the i^* modelling language.

References

1. ACM US Public Policy Council: Statement on algorithmic transparency and accountability (Jan 2017), `https://www.acm.org/binaries/content/assets/public-policy/2017_usacm_statement_algorithms.pdf`
2. Andrews, R., Diederich, J., Tickle, A.B.: Survey and critique of techniques for extracting rules from trained artificial neural networks. Knowledge-Based Systems **8**(6), 373–389 (Dec 1995). https://doi.org/10.1016/0950-7051(96)81920-4
3. Anjomshoae, S., Najjar, A., Calvaresi, D., Främling, K.: Explainable agents and robots: Results from a systematic literature review. In: 18th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS 2019). pp. 1078–1088. International Foundation for Autonomous Agents and Multiagent Systems (2019), `https://dl.acm.org/citation.cfm?id=3331806`
4. Bashir, I.: Mastering Blockchain: Distributed ledger technology, decentralization, and smart contracts explained. Packt Publishing Ltd (2018)
5. Boyer, R.S., Moore, J.S.: A Computational Logic Handbook, Perspectives in Computing, vol. 23. Academic Press (1988)
6. 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 (26–28 Jun 2019), `http://ceur-ws.org/Vol-2404/paper16.pdf`
7. Calvaresi, D., Claudi, A., Dragoni, A.F., Yu, E., Accattoli, D., Sernani, P.: A goal-oriented requirements engineering approach for the ambient assisted living domain. In: 7th International Conference on PErvasive Technologies Related to Assistive Environments (PETRA 2014) (2014). https://doi.org/10.1145/2674396.2674416
8. Calvaresi, D., Mattioli, V., Dubovitskaya, A., Dragoni, A.F., Schumacher, M.: Reputation management in multi-agent systems using permissioned blockchain technology. In: 2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI 2018) (2018). https://doi.org/10.1109/WI.2018.000-5

9. Calvaresi, D., Mualla, Y., Najjar, A., Galland, S., Schumacher, M.: 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. pp. 41–58 (2019). https://doi.org/10.1007/978-3-030-30391-4_3, https://doi.org/10.1007/978-3-030-30391-4_3
10. Ciatto, G., Bosello, M., Mariani, S., Omicini, A.: Comparative analysis of blockchain technologies under a coordination perspective. In: De La Prieta, F., González-Briones, A., Pawleski, P., Calvaresi, D., Del Val, E., Lopes, F., Julian, V., Osaba, E., Sánchez-Iborra, R. (eds.) Highlights of Practical Applications of Survivable Agents and Multi-Agent Systems. The PAAMS Collection, Communications in Computer and Information Science, vol. 1047, chap. 7, pp. 80–91. Springer (Jun 2019). https://doi.org/10.1007/978-3-030-24299-2_7
11. Crawford, K.: Artificial intelligence's white guy problem. The New York Times (2016), http://www.nytimes.com/2016/06/26/opinion/sunday/artificial-intelligences-white-guy-problem.html
12. d'Avila Garcez, A.S., Broda, K., Gabbay, D.M.: Symbolic knowledge extraction from trained neural networks: A sound approach. Artificial Intelligence **125**(1-2), 155–207 (Jan 2001). https://doi.org/10.1016/S0004-3702(00)00077-1
13. Dung, P.M.: On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. Artificial Intelligence **77**(2), 321–357 (1995). https://doi.org/10.1016/0004-3702(94)00041-X
14. Ferber, J.: Multi-Agent Systems: An Introduction to Distributed Artificial Intelligence. Addison-Wesley (1999)
15. Fourcade, M., Healy, K.: Categories all the way down. Historical Social Research / Historische Sozialforschung **42**(1), 286–296 (2017), http://www.jstor.org/stable/44176033
16. Goodman, B., Flaxman, S.: European union regulations on algorithmic decision-making and a "right to explanation". AI Magazine **38**(3), 50–57 (2017). https://doi.org/10.1609/aimag.v38i3.2741
17. Guidotti, R., Monreale, A., Turini, F., Pedreschi, D., Giannotti, F.: A survey of methods for explaining black box models. ACM Computing Surveys (CSUR) **51**(5) (Jan 2019). https://doi.org/10.1145/3236009
18. Gunning, D.: 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
19. Kepuska, V., Bohouta, G.: Next-generation of virtual personal assistants (Microsoft Cortana, Apple Siri, Amazon Alexa and Google Home). In: 2018 IEEE 8th Annual Computing and Communication Workshop and Conference (CCWC 2018). pp. 99–103. IEEE (2018). https://doi.org/10.1109/CCWC.2018.8301638
20. Kontarinis, D.: Debate in a multi-agent system: multiparty argumentation protocols. Ph.D. thesis, Université René Descartes – Paris V (Nov 2014), https://tel.archives-ouvertes.fr/tel-01345797
21. Lipton, Z.C.: The mythos of model interpretability. ACM Queue **16**(3) (May–Jun 2018). https://doi.org/10.1145/3236386.3241340
22. Oliya, M., Pung, H.K.: Towards incremental reasoning for context aware systems. In: Advances in Computing and Communications, Communications in Computer and Information Science, vol. 190, pp. 232–241. Springer (2011). https://doi.org/10.1007/978-3-642-22709-7_24

23. Premack, D., Woodruff, G.: Does the chimpanzee have a theory of mind? Behavioral and brain sciences **1**(4), 515–526 (1978). https://doi.org/10.1017/S0140525X00076512
24. Sun, R.: Artificial intelligence: Connectionist and symbolic approaches. `http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.34.9688` (Dec 1999)
25. Van Lamsweerde, A.: Goal-oriented requirements enginering: a roundtrip from research to practice [enginering read engineering]. In: 12th IEEE International Requirements Engineering Conference (ICRE 2004). pp. 4–7. IEEE (2004). https://doi.org/10.1109/ICRE.2004.1335648
26. Voigt, P., von dem Bussche, A.: The EU General Data Protection Regulation (GDPR). A Practical Guide. Springer (2017). https://doi.org/10.1007/978-3-319-57959-7
27. Yu, E.S.K.: Modelling Strategic Relationships for Process Reengineering. Ph.D. thesis, University of Toronto, Toronto, Ontario, Canada (2014)
28. Zilke, J.R., Loza Mencía, E., Janssen, F.: DeepRED – rule extraction from deep neural networks. In: Calders, T., Ceci, M., Malerba, D. (eds.) Discovery Science (DS 2016). Lecture Notes in Computer Science, vol. 9956, pp. 457–473. Springer (2016). https://doi.org/10.1007/978-3-319-46307-0_29