XAI-DAOs: Decentralised Autonomous Organisations with Explainable Intelligence

Sante Dino Facchini[1,*]

[1]Università degli Studi dell'Aquila, L'Aquila IT 67100, Italy

Abstract

One of the most important challenges in Artificial Intelligence is whether or not to trust the behaviour of intelligent agents and systems. Problems such as bias elimination, transparency, and accountability of system decisions are some of the most relevant factors in this process. When such agents and systems interact with structured organizations, like business companies, the complexity of explainability tasks increases and problems of trustworthy orchestration among heterogeneous entities also arise. Therefore, a solution that brings transparency, accountability, and immutability can be a very valuable one when artificial intelligence interfaces with human beings and complex environments where they may interact. This paper presents the activities carried out during the first two years of PhD and results obtained in integrating Intelligent Controls, Multi-agent systems, Decentralized Autonomous Organizations and Blockchain Technology to implement trustworthy and transparent Artificial Intelligence applications in complex organisations. A state-of-the art of the sector and a perspective of the goals and results expected for the last year is also presented.

Keywords

Explainable Intelligent DAOs, Multi-agents systems, Trustworthy orchestration,

1. Introduction

1.1. Explainability and Trust in Intelligent Applications on Decentralised Organizations

In recent years, Artificial Intelligence (AI) and intelligent technologies applications have seen a significant increase in interest in both commercial and research areas. The more complex the applications, the more complex the tasks and functionalities they can perform. However, the ethical and moral implications also become more complex, especially when the performance of the systems approaches that of humans in that field. The introduction of Decentralized Autonomous Organizations (DAOs) has made it possible to transfer the benefits of Blockchain Technologies (BTs) to complex entities, such as companies or associations, potentially composed of many users. Here we intend both human users and Multi-agent systems (MAS). In this environment, it is of paramount importance to "explain" how decisions are taken by the agents. This is valid for developers of the application, who may benefit from clarity in debugging applications, for end users, who can get a better understanding of the system, but ultimately also

The AIxIA Doctoral Consortium (DC) - 22nd Conference of the Italian Association for Artificial Intelligence (AIxIA), November 6-9, 2023, Rome, Italy

*Corresponding author.

santedino.facchini@graduate.univaq.it (S. D. Facchini)

0000-0002-2009-5209 (S. D. Facchini)

for people affected by system decisions [1]. The generic research question we want to answer is the following: "Is it possible to apply classic Explainable Artificial Intelligence (XAI) paradigms and techniques to real-world applications represented as MAS interacting and integrating with DAOs?" The research questions (RQs) derived from the free-form question are the following: **RQ1**. "How has the application of AI to Blockchain systems and DAOs evolved over the years? How XAI has adapted to new paradigms in intelligent systems" **RQ2**. "What are the reasons for using MAS, BTs, and DAOs to model complex entities governed by Intelligent Controls?" **RQ3**. "What are pros and cons of using such technologies?" **RQ4**. "What could be the future research directions and application challenges?"

1.2. History and State of the Art

Answer to RQ1 Explainability in the area of Artificial Intelligence (AI) has its roots in the late 1970s, in the field of expert systems [2]. The idea at that time was to explain behavior through applied rules. It is with the boom of Deep Neural networks (DNNs) in the 2020s that this research area has assumed a relevant role. DNNs are considered to be opaque black boxes [3], so new explanatory tools had to be introduced to illustrate how inference decisions were made. These tools include post-hoc explanations (or how the inferred result was reached) and transparency design (how the model works). These innovative paradigms were introduced in DARPA's [1] Explainable Artificial Intelligence (XAI) Program in 2017 [4], which represents the modern reference in this field. The program also introduced the idea that prediction accuracy is inversely proportional to the explainability of the model. Main branches of research in XAI are Data-driven or Perceptual XAI focused on interpreting results of black-box and Goal-driven or Cognitive XAI that explain systems presenting their goal, intentions, beliefs, etc. Other important documents on AI, this time on the regulatory side, were the publication of the Chinese government's 2017 "The Development Plan for New Generation of Artificial Intelligence" [5], the "Statement on algorithmic transparency and accountability" of US ACM Public Policy Council [2] and the European Union's 2018 "General Data Protection Regulation" (GDPR). The GDPR particularly defines the right to explanation for citizens affected by any algorithmic decision [6]. By the end of 2023, the approval of the Artificial Intelligence Act by the European Union [3] should represent the first attempt to regulate the aspects of artificial intelligence in everyday life applications. Coming to application of BTs to MAS, the motivation for merging these to paradigms comes mainly from the need to extend the properties of blockchains to the internal mechanisms of agents and their intercommunication. Distributed Ledgers Technologies (DLTs) are very useful for keeping track of events, as they can timestamp information, keep it immutable, and register actions. In this field, the literature is mainly focused on the conceptual aspects of the problem in the areas of trust, collaborative governance of systems, and reputation management. Aspects such as agent coordination applications, integration of smart contracts into agents, and scalability of distributed score-based systems are still in their early stages. [7]. Furthermore, the discussion of the benefits of integrating MAS and BTs is still too theoretical

[1] https://www.darpa.mil/program/explainable-artificial-intelligence, updated Feb 24
[2] https://www.acm.org/binaries/content/assets/public-policy/2017_usacm_statement_algorithms.pdf, updated Feb 24
[3] https://www.europarl.europa.eu/news/en/headlines/society/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence, updated Feb 24

and lacks real applications and demonstrators.(see [8]). State-of-the-art research on explainable AI (XAI) applied to MAS on BT has been somewhat lacking in excitement, as researchers have primarily focused on explaining data-driven algorithms rather than agent and robot systems [9]. However, recent studies have highlighted the need to increase trust in autonomous intelligent environments as well, by explaining the rules that are applied, the way decisions are made, and the meaning of the results obtained [10]. The main areas of application of this research field have been cloud computing, smart cities, and user satisfaction management systems. The objective has been to improve real-world scenarios by addressing the problems of resource-constrained environments and lack of trust in the systems. The DAOs paradigm is a relatively new one, first defined by Vitalik Buterin in 2014 [11]. In a very simple but effective way, they can be described as a blockchain-based entity where participants interact on a peer-to-peer basis and governance is achieved through automated rules written in smart contracts. DAOs are gaining more interest due to the evolution of increasingly powerful digital platforms and new technologies that enable the entanglement of social behaviour in distributed systems [12]. Murray et al. [13] propose an interesting field of study that postulates a "conjoined agency" [4] between agents and humans thanks to such advances. However, the literature is still not mature, lacking empirical and real-world applications, and has not yet reached a common understanding of DAOs.

2. Innovative Approach

The aim of the doctoral program is to define a new approach to DAOs architecture that integrates MAS and intelligent controls to extend the good practices of XAI to complex entities. This innovative approach will also be tested on real-world case studies through software demonstrators and applications, as this is one of the main gap in the research area (see Section 1.2).

2.1. How to Define Trust in XAI-DAOs

Many are the definitions of trust when coming to explainable AI, but the most accepted in the MAS sector is the one of Gambetta et al.: [14] "Trust is the subjective probability by which an agent A expects that another agent B performs a given action on which its welfare depends"; the idea is that trust has a binary nature. It is firstly an evaluation about others reliability, agent A predicts the future behaviour of an agent B relying on its mental state; secondly is an actuation taken on such measure of trust. We may try to formalize and generalize Gambetta's paradigm using logic proposition, let's say $T_i(j) \rightarrow [0, 1]$ is the trust evaluation agent i makes of j assigning a degree of trustfulness. Agent i will takes an action a based on trust of j if such function T is greater than a certain reputation threshold r_{i_j}. The proposition that describe it is $a \Longleftarrow T_i(j) \geq r_{i_j}$.
Extention to DAOs: Consider that the main actions taken by agents in DAOs are voting on resolutions for organisational governance. We can think of a trust estimate of each vote by a smart contract that regulates the DAO. It could self-estimate the trustfulness of a decision by calculating the weighted average trust of each vote. If $T_i(j) \rightarrow [0, 1]$ is the trust evaluation that agent i has of agent j and $Rj \in \mathbb{R}$ is the reputation agent j has in the DAO environment, then level of confidence C of the vote V cast by n agents could be $C(V) = \frac{\sum_i \sum_j Rj * T_i(j)}{n}$. This

[4] Also referred to as Human AI, is the process of sharing a decision process between human and non-human entities

is a possible interpretation of trust in DAOs that relies on an agent's reputation. However, other features of the agents could be considered as substitutes or in combination (e.g., level of experience, age in the system, speed of reply, wealth, etc.). In this way, we can configure a "reputational" system based on a multidimensional array to represent the state of mind of an agent in relation to a DAO activity.

2.2. How to Define Explainability in XAI-DAOs

An interesting framework for XAI, which will be taken as reference in the doctoral program has been given by Ciatto et al. [15]. The idea at the base is to model explainability as two related activities: the objective explanation of the system and the subjective interpretation of its behaviour. Interpretation is the assignment of a subjective meaning by an agent i to a system X, in this way a function $I_i(X) \rightarrow [0, 1]$ can be defined such as a degree on interpretability of the system. If $I_i(X') \geq I_i(X)$ then system X' is more interpretable of X. Explanation is defined instead, as the activity taken by an agent i of increasing the interpretability of a system X to get a second system X' such as it is more interpretable of the former one. Explanation is thus a function $E_i(X) \rightarrow X'$ such as $I_i(X') \geq I_i(X)$. The cited framework is also extensible to a Multi-agent environment with heterogeneous behaviours and different knowledge bases. It models an explanator-explanee paradigm where a "mutual understanding" relationship between agents (either human or sofwtare) on a 1-to-1 or n-to-m base. The idea is that if an agent wants to explain to another agent a model, an agreement involving a shared taxonomy and a mechanism to reconciliate mutual knowledge base must be reached. **Extention to DAOs:** This framework could be extended to DAOs where an agent A (the explanator) may have to explain to other agents or people (the explanee) afflicted by DAO's voting poll V, why he has taken such decision on resolution R. In this case the explanation E could be expressed as $E = e(A, V, R)$ mapped for example on a minimal set of features $f_1..f_n$ that agent A consider sufficient to take decision V. The explanee at this point can: (i) be satisfied or start dialogue where may (ii) request a more granular explanation E' which could be calculated on more features $f_1..f_n, n > m$ or an historical sequence of them. It could ultimately (iii) ask to another voting agent B ti have a new explaination $E' = e(B, V, R)$.

2.3. Increase Trust and Explainability in Distributed Organizations

Answer to RQ2: Modelling complex organizations of interacting entities is a task of great importance, as distributed AI systems are becoming increasingly important. Introducing explainability and trust mechanisms, especially in sensitive applications (e.g., medical expert systems, financial applications, etc.), can help create trustworthy organizations where human-non-human interaction is more productive and secure. Extensions and integrations proposed in previous sections would allow us to implement such paradigms required by human AI needs. DAOs, on the other hand, can enable transparent and accountable orchestration of all players involved in the explanation and interpretation process. **Answer to RQ3:** The main problems that could arise from such integration may regard the performance of the system, in terms of the speed of transactions and operations. Integrating blockchain can slow down the overall system performance, as approval and validation of transactions can take time (and often involve a cost).

Additionally, DAO mechanisms can introduce complexity to the decision-making process (e.g., waiting for a poll to be taken). As a result, real-time systems or applications where time is a critical factor may be severely impacted by these issues and may not be compatible.

3. Innovative Results and Doctoral Program Activity

The first year focused on identifying and evaluating technologies for integrating MAS with DAOs and blockchain. This involved surveying the state-of-the-art in MAS research, identifying gaps in the research, and testing and selecting the best frameworks for developing demonstrators [16]. As a result of this activity, a demonstrator was implemented that integrates DAOs and blockchain to model the redemption of fiscal credits [17]. In the second year, the focus is on implementing test systems on more realistic scenarios based on the extended frameworks. An extended demonstrator of tax credit redemption, with agent orchestration based on the MESA Framework [5], and a simple explanation system, has been implemented. The article describing this work is currently under review by a journal.

4. Future Applications and Research Direction

Answer to RQ4: Third year of PhD will aim to fill the research gap in the lack of real-world applications. This will involve two main areas. The first area will focus on practical aspects and will involve enriching the demonstrator with a more realistic DAO implementation and with applications to other real-world cases (tracking system of honey production and selling, tracking of seismic works on buildings). The second area will focus on defining an innovative framework for configuring explainable AI (XAI) modules for DAOs. This will decouple configuration parameters (such as nodes and decision features) from the code, making it easier to develop and deploy modules.

5. Acknowledgments and Declarations

This research project was outlined and supervised by my tutor at DISIM - Università degli Studi dell'Aquila, Professor Giovanni De Gasperis. The demonstrators and applications described were tested on real-world business cases provided by the innovative startup BCC Studio and the engineering company Studio Berkana. The author used occasionally the Bard AI tool[6] in order to improve the readability of some parts of the text. After using the tool, the author reviewed and edited the content as needed, and took full responsibility for the content of the publication.

References

[1] F. Xu, H. Uszkoreit, Y. Du, W. Fan, D. Zhao, J. Zhu, Explainable ai: A brief survey on history, research areas, approaches and challenges, in: Natural Language Processing and

[5] https://mesa.readthedocs.io/en/main/index.html, updated Feb 24
[6] https://bard.google.com/

Chinese Computing: 8th CCF International Conference, NLPCC 2019, Dunhuang, China, October 9–14, 2019, Proceedings, Part II 8, Springer, 2019, pp. 563–574.
[2] A. C. Scott, W. J. Clancey, R. Davis, E. H. Shortliffe, Explanation capabilities of production-based consultation systems, American Journal of Computational Linguistics (1977) 1–50.
[3] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, D. Pedreschi, A survey of methods for explaining black box models, ACM computing surveys (CSUR) 51 (2018) 1–42.
[4] D. Gunning, Explainable artificial intelligence (xai), Defense advanced research projects agency (DARPA), nd Web 2 (2017) 1.
[5] H. Roberts, J. Cowls, J. Morley, M. Taddeo, V. Wang, L. Floridi, The chinese approach to artificial intelligence: an analysis of policy, ethics, and regulation, AI & society 36 (2021) 59–77.
[6] B. Goodman, S. Flaxman, European union regulations on algorithmic decision-making and a "right to explanation", AI magazine 38 (2017) 50–57.
[7] J. Gattermayer, P. Tvrdik, Blockchain-based multi-level scoring system for p2p clusters, in: 2017 46th International Conference on Parallel Processing Workshops (ICPPW), IEEE, 2017, pp. 301–308.
[8] D. Calvaresi, A. Dubovitskaya, J. P. Calbimonte, K. Taveter, M. Schumacher, Multi-agent systems and blockchain: Results from a systematic literature review, in: Advances in Practical Applications of Agents, Multi-Agent Systems, and Complexity: The PAAMS Collection: 16th International Conference, PAAMS 2018, Toledo, Spain, June 20–22, 2018, Proceedings 16, Springer, 2018, pp. 110–126.
[9] D. Calvaresi, Y. Mualla, A. Najjar, S. Galland, M. Schumacher, 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 1, Springer, 2019, pp. 41–58.
[10] T. Hellström, S. Bensch, Understandable robots-what, why, and how, Paladyn, Journal of Behavioral Robotics 9 (2018) 110–123.
[11] V. Buterin, Daos, dacs, das and more: An incomplete terminology guide, Ethereum Blog 6 (2014) 2014.
[12] C. Santana, L. Albareda, Blockchain and the emergence of decentralized autonomous organizations (daos): An integrative model and research agenda, Technological Forecasting and Social Change 182 (2022) 121806.
[13] A. Murray, S. Kuban, M. Josefy, J. Anderson, Contracting in the smart era: The implications of blockchain and decentralized autonomous organizations for contracting and corporate governance, Academy of Management Perspectives 35 (2021) 622–641.
[14] D. Gambetta, Trust: Making and breaking cooperative relations (1988).
[15] G. Ciatto, M. I. Schumacher, A. Omicini, D. Calvaresi, Agent-based explanations in ai: Towards an abstract framework, in: International workshop on explainable, transparent autonomous agents and multi-agent systems, Springer, 2020, pp. 3–20.
[16] S. Facchini, Decentralized autonomous organizations and multi-agent systems for artificial intelligence applications and data analysis, in: Doctoral Consortium, 2022, pp. 5851–5852.
[17] G. De Gasperis, S. D. Facchini, A. Susco, Demonstrator of decentralized autonomous organizations for tax credit's tracking, in: International Conference on Practical Applications of Agents and Multi-Agent Systems, Springer, 2022, pp. 480–486.