Explainable Multi-Agent Systems through Blockchain Technology

Davide Calvaresi[1], Yazan Mualla[2], Amro Najjar[3], Stéphane Galland[2], and Michael Schumacher[1]

[1] University of Applied Sciences and Arts Western Switzerland, Switzerland
[2] CIAD, Univ. Bourgogne Franche-Comté, UTBM, F-90010 Belfort, France
[3] UMEA University, Sweden
{davide.calvaresi, michael.schumacher}@hevs.ch, najjar@cs.umu.se,
{yazan.mualla, stephane.galland}@utbm.fr

Abstract. Advances in Artificial Intelligence (AI) are contributing to a broad set of domains. In particular, Multi-Agent Systems (MAS) are increasingly approaching critical areas such as medicine, autonomous vehicles, criminal justice, and financial markets. Such a trend is producing a growing AI-Human society entanglement. Thus, several concerns are raised around user acceptance of AI agents. Trust issues, mainly due to their lack of explainability, are the most relevant. In recent decades, the priority has been pursuing the optimal performance at the expenses of the interpretability. It led to remarkable achievements in fields such as computer vision, natural language processing, and decision-making systems. However, the crucial questions driven by the social reluctance to accept AI-based decisions may lead to entirely new dynamics and technologies fostering explainability, authenticity, and user-centricity. This paper proposes a joint approach employing both blockchain technology (BCT) and explainability in the decision-making process of MAS. By doing so, current opaque decision-making processes can be made more transparent and secure and thereby trustworthy from the human user standpoint. Moreover, several case studies involving Unmanned Aerial Vehicles (UAV) are discussed. Finally, the paper discusses roles, balance, and trade-offs between explainability and BCT in trust-dependent systems.

Keywords: MAS · goal-based XAI · explainability · UAV · Blockchain.

1 Introduction

Human decisions are increasingly relying on Artificial Intelligence (AI) techniques implementing autonomous decision making and distributed problem solving. Human-system interaction is pervading many domains, including healthcare [7], Cyber-Physical Systems [13, 32], financial markets [39], and cloud computing [36]. Such entanglements enforced the ratification of the recent European General Data Protection Regulation (GDPR) law which underlines the right to

explanations [15] and ACM US Public Policy Council (USACM)'s algorithmic transparency and accountability [1].

Therefore, the design of transparent and intelligible technologies is an impelling necessity. However, the interaction between autonomous AI-based systems (e.g., robots and agents) and humans decision processes raises concerns about the trust, reliability, and acceptance of autonomous systems. Recent studies proved that for both humans and software agents/robots, the trust into autonomous intelligent systems is strengthened if rules, decisions, and results can be explained. Hence, in the last decade, the hype about eXplainable Artificial Intelligence (XAI) [26, 4] picked again. However, the majority of the recent studies focus on the interpretability and explanations for data-driven algorithms [5, 19, 25, 42], thus still leaving open investigations concerning explainable agents and robots [4].

Humans tend to associate rationales to understanding and actions, developing a "mental states" [27]. A missing explanation can generate understanding that does not necessarily reflect AI's internal stance (self-deception). To a certain extent, dangerous situations may arise, putting the user safety at risk. According to the recent literature [5, 38], explanations help users to increase confidence and trust, whereas misunderstanding the intentions of the intelligent system creates discomfort and confusion. Therefore, endowing these agents and robots with explainable behavior is paramount for their success. Interacting with these systems, however, there are domains and scenarios in which giving a proper explanation is **not** *(i)* possible, *(ii)* worth it, or *(iii)* enough. Therefore, the novelty proposed by this work is the following.

Contribution:
This paper proposes to combine XAI, with blockchain technologies to ensure trust in domains where, due to environmental constraints or to some characteristics of the users/agents in the system, the effectiveness of the explanation may drop dramatically.

The rest of this article is organized as follows. Section 2 presents the background of this work in the domains of trust, explainability, and blockchain technology. Section 3 identifies three key research domains in which the synergy between BCT and XAI is necessary. Section 4 highlights the major challenges, Section 5 presents the proposed solution. Section 6 presents a use-case scenario, Section 7 discusses the scope of attainable solutions in which a combination of BCT and XAI is to be successful, and finally Section 8 concludes the paper.

2 Background

This section gives an overview of *trust* (Section 2.1), *explainability* (Section 2.2), and *blockchain* (Section 2.3) which are the key elements enabling the understanding of what their combination can provide to Multi-Agent Systems (MAS).

2.1 Trust

Autonomy is considered a basic feature for intelligent agents. Although it is highly desirable, such a property raises several challenges [41]. For example, *(i)* the agent designer must take into account the *autonomy of other agents* (run-time adaptation is a must for any agent to be competitive), and *(ii)* it is unrealistic to assume that other agents adopt a same/similar conduct.

Thus, artificial societies need some sort of control mechanisms. Traditionally, computational security has been claimed to be able to address a set of well-defined threats/attacks by relying on cryptography algorithms [22]. Yet, this approach requires the existence of a Trusted Third Party (TTP) to provide public and private keys and other credentials, which, for decentralized and open application scenario, becomes unrealistic [8]. On turn, several other *soft control* techniques have been defined to provide a certain degree of control without restricting the system development. These approaches rely on *social control mechanisms* (e.g., trust and reputation) that do not prevent undesirable events but ensure some social order in the system [16]. Nevertheless, they can allow the system to evolve in a way which prevents them from appearing again.

Several definitions have been proposed to define the notion of trust. Yet, the definition proposed by Gambetta *et al.* [24] is particularly useful and adopted by the MAS community.

> "Trust is the subjective probability by which an agent $\mathcal{A}$ expects that another agent $\mathcal{B}$ performs a given action on which its welfare depends".

Therefore, trust is seen as an estimation or a prediction of the future or an expectation of an uncertain behavior, mostly based on previous behaviors [10]. A second form of trust is the *act of taking a decision* itself (e.g., relying on, counting on, or depending on the trustee). Summarizing, trust is both:

(i) a mental state about the others trustworthiness (an evaluation) and
(ii) a decision or intention based on that evaluation [41]. To evaluate the trust, an agent relies on the *image* of the other agents. An image is an evaluative belief that tells whether the target is good or bad with respect to the given behavior. Images are results of internal reasoning from different sources of information that lead the agent to create a belief about the behavior of other agents [41].

2.2 Explainability

Explaining the decisions taken by an "intelligent system" has received relevant contributions from the AI community [17, 28]. Earlier works on sought to build explainable expert systems. For this reason, after a prosperous phase, explainability received less attention in the 2000's. Recently, as AI systems are getting increasingly complex, explainable AI (XAI) reemerged to push for interpreting the "black-box" machine learning mechanisms and understanding the decisions of robots and agents. Consequently, research on XAI can be classified in two main branches:

- **Data-driven** (so-called *perceptual* [37]) XAI:
  It aims at *interpreting* the results of "black-box" machine learning mechanisms such as Deep Neural Networks (DNN) [49]. This research achieved intriguing results (e.g., understanding why a DNN mistakenly labelled a tomato as a dog [45]). Therefore, the lust to *interpret*, or *provide a meaning* for an obscure machine learning model (whose inner-workings are otherwise unknown or non-understandable by the human observer) is tickling the researchers.
- **Goal-driven** (so-called *cognitive* [37]) XAI:
  Research from cognitive science has shown that humans attribute mental states to robots and autonomous agents. This means that humans tend to attribute goals, intentions and desires to these systems. This branch of XAI aims at explaining the rationales of the decisions of intelligent agents and robots by citing their goals, beliefs, emotions, etc. [4]. Providing such explanations allows the human to understand *capabilities*, *limits*, and *risks* of the agent/robot they are interacting with, and thereby raising the user awareness and trust in the agent, facilitating *critical decisions* [4, 14].

2.3 Blockchain technology

Blockchain is a distributed technology employing cryptographic primitives that rely on a *(i)* membership mechanism, and *(ii)* a consensus protocol to maintain a shared, immutable, and transparent append-only register [10]. Observing The information (digitally signed transactions) delivered by the entities part of the network are grouped into blocks chronologically time-stamped.

The single block is identified by a unique block-identifier, which is obtained by applying a hash function to its content and it is stored in the subsequent block. Such a technique is part of a set of mechanisms considered *tamper-proof* [9] modification of the content of a block, can be easily verified by hashing it again, and comparing the results with the identifier from the subsequent block. Moreover, depending on the distribution and consensus mechanism, the blockchain can be replicated and maintained by every (or a sub-set) participant(s) (so-called peers). Thus, a malicious attempt to tamper the information stored in the registry can be immediately spotted by the participants, thus guaranteeing immutability of the ledger [9]. Several technological implementations of the blockchain can execute arbitrary tasks (so-called smart contracts) allowing the implementation of desired functionality. Alongside the blocks, such smart contracts represent the logic applied and distributed with the data [29].

Technology BCT can be distinguish between *permissionless* and *permissioned* (public and private) blockchain systems [44]:

- A blockchain is *permissionless* when the identities of participants are either pseudonymous or anonymous (every user can participate in the consensus protocol, and therefore append a new block to the ledger).

– A blockchain is permissioned if the identities of the users and rights to participate in the consensus (writing to the ledger and/or validating the transactions) are controlled by a membership service.

Moreover, on the one hand, a **permissioned** blockchain is *public* when anyone can read the ledger, **but** only predefined set of users can participate in the consensus. On the other hand, it is *private* when even the right to read the ledger is controlled by a membership/identity service.

3 Application Domains

Trust is still an outstanding open challenge in the area of intelligent systems. However, **Blockchain** technology and techniques derived from the **XAI** discipline can be tightly coupled to provide reconciling, feasible, and cost-effective solutions. On the one hand, explainable behaviors can enable the trustor to evaluate the soundness and the completeness of the actions of the trustee, and thereby it can evaluate its competences, and examine the rationale behind its behavior. On the other hand, BCT can allow the trustor to unequivocally assess the *reputation* of the trustee based on existing history knowledge about it. In this paper, we explore reconciling solutions combining both XAI and BCT. This synergy can be beneficial for several application domains involving collaborations among agents to undertake joint decisions in a decentralized manner. Below, we identify three types of applications in which such a synergy would be highly beneficial.

Cloud Computing is a distributed ecosystem involving multiple actors each concerned with accomplishing a different set of goals. Agent-based systems have been underlined as a platform capable of adding intelligence to the cloud ecosystem and allowing to undertake critical tasks such as resource management in a decentralized manner that considers the distributed and multi-partite nature of the cloud ecosystem [46]. In a typical three partite scenario, it involves: *(i)* Cloud providers who seek to offer an adequate Quality of Service (QoS) while minimizing the energy consumption and maintenance costs of its data-centers [23], *(ii)* Cloud users whose aim is to minimize the cost they pay to the provider while furnishing a satisfactory service to their end-users [35], and brokers. In exchange for a fee, a broker reserves a large pool of instances from cloud providers and serves users with price discounts. Thus, it optimally exploits both pricing benefits of long-term instance reservations and multiplexing gains [48]. In such a scenario, given the multitude of providers, brokers and offers available in the cloud market, both explainability and trust are critical to help these actors make their strategic decisions. For instance, when recommending resources from a particular cloud provider, a broker could rely on BCT technology to assess the reputation and the trustworthiness of the provider. Several important data could be inscribed on the ledger including the availability, reliability and the average response time of the virtual instances leased from this provider. When giving a recommendation, the broker might also use explainability to provide a transparent service to its

client and explain why some specific decision were made (e.g., the choice of one provider) and why some un-expected events took place (e.g., an SLA violation).

Smart Cities. The densely populated smart cities are administrated by several governmental and civil society actors, where vivid economic services involving a multitude of individual stakeholders take place. In such services, the use of agents for Unmanned Aerial Vehicles (UAVs) is gaining more interest especially in complex application scenarios where coordination and cooperation are necessary [33]. In particular, in the near future, UAVs will require access to an interoperable, affordable, responsive, and sustainable networked system capable of providing service, joint, inter-agency, and real-time information exchanges. Such systems must be distributed, scalable, and secure. The main components are human interfaces, software applications, network services, information services, and the hardware and interfaces necessary to form a complete system that delivers secured UAVs operations [29]. Recalling that BCT allows creating a peer-to-peer decentralized network with an information protection mechanism [3], such a network can provide secure communication system within the MAS [21], thus operating as distributed control and secure system to ensure the trust among UAVs and other actors.

User Satisfaction Management. Agents are autonomous entities bound to individual perspectives, for these reasons, user agents were used to represent user satisfaction [36]. However, end-user satisfaction is known to be subjective [34] and influenced by several Influence Factors (IF) [40], including Human IFs (e.g., expertise, age, personality traits, etc.), Context IF (e.g., expectations) and System IFs (i.e., the technical properties of the systems used to consume the service). Both XAI and BCT can have key contributions helping agents overcome these challenges and improve user satisfaction. On the one hand, explainability enables the agent to provide convincing recommendations to the user by showing that the agent's decisions were in line with the user preferences. On the other hand, BCT can play an important role in assuring both the user and her agent that privacy and authentication measures are integrated to protect the user preferences and private data from exploitation.

4 Challenges

The combination of MAS, BCT, and XAI can be particularly strategic in several application fields. Real-world scenarios are often characterized by a combination of limited resources such as computational capability, memory, space, and in particular *time* [12, 13, 20]. Therefore, Section 4.1 tackles the application of the proposed solution in Resource-Constrained (RC). Another relevant dimension characterizing real-world application is the *trust* in the systems or in their components [10, 11, 41]. Thus, Section 4.2 addresses the Lack of Trust (LT) as main driver.

4.1 RC Scenarios

In real-world applications, systems must cope with a bounded availability of resources. On the one hand, we can mention tangible resources such as memory, computational capability, and communication bandwidth [13]. On the other hand, we can have reputation, trust, and time. The latter is crucial especially in safety-critical scenarios, when failing to deliver a result in/on time might have catastrophic consequences [6].

A possible example can be a UAVs firefighting scenario.

Let us assume that a UAV detects a fire in a nearby woods, and that the fire has already spread to an extent unmanageable by a single UAV. The only viable option for the UAV which detected the fire is to ask for support from the firefighting center, managed by humans, to send other UAVs. This requires the UAV to explain the situation to the representative human in the firefighting center. Considering that such a situation needs an intervention as prompt as possible, the UAV requesting assistance cannot *produce* and *deliver* an "extensive" explanation for its requests, plans, and the consequences of possible inaction. Achieving a consensus on an over-detailed (for the situation) explanation would be unaffordably time-consuming, thus leading to potentially considerable losses. A possible solution is to enable the requester to rely on BCT, which can ensure its possible trustworthiness (e.g., via reputation) and authenticity, compensating a less detailed explanation leading to a faster reaction to handle the fire.

4.2 LT Scenarios

In scenarios where time is not critical, the opportunity is given to an agent with low reputation to express itself to increase the trust with other actors. For example, a swarm of UAVs can be created to perform tasks that cannot be performed by one UAV or to increase the efficiency of a specific task. In such situations, there is a need for a mechanism for UAVs to join a swarm. Yet, a UAV with a low reputation may find it difficult to join a swarm. With explainability, it is possible that swarm management gives this UAV a chance to express itself in order to increase its trust and hence its chances to be accepted in the swarm. Another example is when it is not possible to determine the reputation of a UAV due to the inability to access the blockchain. This UAV can be given the chance of explaining its goals to increase the likelihood of an agreement.

5 Proposed Solution

According to the application domains and scenarios presented in Section 4, a two-folded solution (for RC and LT scenarios) follows.

5.1 RC

In scenarios in which the operating *time* is constrained (Section 4.1) and delivering a complete and high-quality explanation is not viable, the quality and

granularity of a given explanation might be degraded to still comply with the timing constraints.

Similarly, if the *understanding capability* of the recipient of a given explanation is limited (Scenarios 3 and 4 in Table 1), the quality of the explanation can be lowered (since it might not be understood/appreciated) saving both time and effort (e.g., computational capability, memory).

Lower quality explanations are characterized by less details (coarse-grained) or unfaithful explanations. While offering brief insights on how and why a decision was taken, coarse-grained explanations do not provide a fully detailed explanation unless this is explicitly demanded by the explainee. Unfaithful explanation do not respect the actual mechanism that led to a given decision. Instead, their aim is to provide an understandable and easy explanation. A possible way of providing unfaithful explanation is relying on contrastive explanations. The latter consist of justifying one action by explaining why alternative actions were not chosen. While contrastive explanations do not necessarily describe the decision-making process of the agent, recent research has shown that they can be easily produced and easily understandable by the user [30]. Therefore, both coarse-grained and unfaithful explanations convey the message, thus accomplishing the explicative intent. Since an effective explanation might not be the most precise or faithful, it is possible to infer that precision and effectiveness of an explanation can be decorrelated. On the one hand, if the principal objective is to share the rationale behind a given decision, opting for an effective and potentially less precise explanation might be the best option [4]. On the other hand, if transparency is a mandatory requirement, a detailed and faithful explanation must be provided. For example, *time* available to produce and provide an explanation in a given context/situation is a factor influencing the agent, thus possibly impacting on the faithfulness of its explanations. In case the amount of time is too constrictive, the agent might opt for a short, simple, and unfaithful explanation (even though a detailed one would be preferred). Moreover, depending on time available, context, and explainee, the explainer may attempt at explaining the same concept employing different types of data or same data but with different granularity and complexity (e.g., images, text, or raw data).

To lower the explanation quality/granularity, without affecting the trust (information-, user-, or agent-wise), we propose to enforce the provided explanation with BCT. By doing so, we would compensate a less effective explanation with the guarantees provided by BCT technology, still keeping the system running and the trust unaffected by a time-critical scenario.

Table 1 lists four possible situations we have identified. Beside the *Time Available*, expressed in seconds, the other features are represented by adimensional numbers (useful to provide a quick and synthetic overview). *Ratio*, stands for correlation between the *quality* of a given explanation (possibly combined with the support of BCT) and how it is *understood, perceived* or *accepted* (if relying more on the BCT then on the actual explanation) by the recipient.

Table 1: Possible combinations of explanations' quality and blockchain support with the recipient's capabilities of understanding.

Scenario	Time Available (seconds)	Explanation quality	Recipient Understanding	Blockchain support	Gain
1	10	10	10	0	10/10
2	5	7	10	3	10/10
3	10	10	5	0	10/5
4	10	2	5	3	5/5

Scenario 1 the first scenario reproduces an ideal situation: having *(i)* enough time to provide a comprehensive and solid explanation and *(ii)* a recipient who has time and can process/understand the provided explanation. In this case, the support of the BCT is not necessary.

Scenario 2 short in time, and with a recipient able to fully understand and process the explanation, the agent opts for degrading the quality of the explanation relying on the contribution of BCT. In this case, the recipient's decision might not be affected by the lack of granularity of the received explanation.

Scenario 3 although the available time is enough to produce a robust explanation, the recipient is not able to entirely understand/process it. Therefore, since the explanation goes already over its purpose, it is not necessary to employ the BCT.

Scenario 4 the available time might be more than enough to produce a robust explanation, which however goes beyond the understanding capability of the recipient. Therefore, to save time and resources, the explanation can be degraded and coupled with the support of BCT, enough to match the recipient expectation and capability.

5.2 LT

In circumstance where an agent/user has a reputation lower than a given threshold, it can be labelled as not fully trustworthy. In this condition, although the user/agent might be able to provide an excellent explanation, it could not be trusted, or it could not get a chance to express it. Therefore, binding the explanation with BCT might relieve the agent explaining from the burden of a low reputation (obtained for a whatever *unfortunate* reason in a precedent point in time). Such a solution/approach can be associated to the famous dilemma "The Boy Who Cried Wolf" narrated by Aesop [2], the well-known Greek fabulist and storyteller. The fable narrates of a young shepherd who, just for fun, used to fool the gentlemen of the nearby village making fake claims of having

his flock attacked by wolves. However, when the wolves attacked the flock for real, the villagers did not respond to the boy's cries (since they considered it to be just another false alarm). Therefore, the wolves end up ravaging the entire flock. This story is used as an example of the serious consequences of spreading false claims and alarms, generating mistrust in the society and resulting in the subsequent disbelieving the true claims. To "cry wolf", a famous English idiom glossed in Oxford English Dictionary [18], was derived from this fable to reflect the meaning of *spreading false alarms*.

Such a moral, applied to *human* societies, can also be applied to *agent* societies. For example, the requests of a UAV with a low reputation might be neglected because its records on the the ledger showed that it has been issuing false alarms about fires in the woods. However, with the possibility of explaining its new alarms and supporting its claims with tangible proofs (*e.g.,* images and footage from the fire location), if its explanations were convincing enough, the UAV might be able to overcome (and improve) its low reputation.

The next section addresses the UAVs package delivery, which is a use case from the real world. In such a scenario, multiple UAVs need to coordinate in order to achieve a common goal. To do so, members of the same UAV team (i.e., swarm) should share a common understanding and maintain a trustworthy relationship. To address these concerns, potentially time-constrained, the following section studies UAVs interaction and reputation by employing explainability and BCT.

6 Explainability and BCT: the UAVs package delivery use case

In 02 Aug 2018, the U.S. Patent and Trademark Office issued a new patent for retail giant Walmart seeking to utilize BCT to perfect a smarter package delivery tracking system [43]. Walmart describes a "smart package" delivered by a UAV that includes a device to record information about a blockchain related to the content of the package, environmental conditions, location, manufacturer, model number, etc. The application states that the blockchain component will be encrypted into the device and will have "key addresses along the chain of the packages custody, including hashing with a seller private key address, a courier private key address and a buyer private key address" [47].

Typically, modeled as agents, UAVs can be organized in swarms to help them achieve more than what they could solely. A decentralized swarm management system can add or remove UAVs from the swarm. To join the swarm, a reputation threshold should be acquired by the UAV. In cases of low reputation UAV (Section 5.2), the choice is given to the UAV to explain the reasons it must join the swarm.

UAVs use voting in the swarm to decide decisions like adding/removing UAVs, tasks to perform, etc. Before each vote, the possibility is given to each UAV to explain what it considers the best for the swarm in terms of what goals to achieve and how to do them. The swarm management system has a blockchain distributed ledger that is connected to Internet through various wireless networks

(e.g., WiFi, 4G/5G, satellite). It allows the swarm to check the reputation of any UAV willing to join the swarm as well as the reputation of any outer actors that wish to communicate with the swarm.

For example, suppose that a new UAV has joined the swarm and is granted a private key. Once the UAV exists on the blockchain distributed ledger of the swarm management system, the levels of access, control, and/or authority are determined for the new UAV.

External actors (UAVs or people) may ask the swarm to perform tasks for them. Negotiation will commence between the external actor and the swarm that considers the trade-off between explainablity and reputation of the actor, the profit of performing the task (in case of commercial swarms), the general welfare (in case of governmental or non-profit organizations). If the swarm accepts to perform a given task, smart contracts can be used to transfer commands between agents in the form of data or executable code in real-time.

Let us assume that an actor (human, device, etc.) in a smart home asks the swarm to make a delivery order. Depending on the time window of the delivery transaction, different scenarios that combine reputation and explainability are considered (Section 5.1). Figure 1 shows the steps to consider as per the constraints of the scenario.

If an agreement is reached, a smart contract is generated with the order data (e.g., package characteristics, client data, location, and designated UAV) and the information is sent to the Blockchain. Then, the UAV commits a transaction to the traffic coordinator to provide an air corridor for it and a new smart contract is concluded between them.

The UAV starts the delivery to the smart home. Once near the smart home, the UAV will contact the smart window using a wireless network. The smart window is connected to the internet as any other device in the smart home. This allows it to ask the blockchain if it recognizes and verifies this UAV and its swarm, and if it is the swarm that singed the smart contract. If the UAV is trustworthy, the window will open to allow it to drop the package. When the delivery is completed, the UAV notifies the traffic coordinator that the air corridor is no longer needed.

To achieve all of that, there is a need for defining two important aspects. First, protocols for the registration, verification, peer-to-peer interaction of the UAVs. Second, smart contracts between the swarm and any other actor in the environment (UAV, device, human, etc.), that govern the services used or provided by the swarm. Moreover, the use of a blockchain infrastructure helps in identifying misbehaving UAVs by multiple parties and such activities are recorded in an immutable ledger. These misbehaving assessments may be performed by analytical algorithms or machine learning models performed off-chain and interfaced with the blockchain ledger through smart contracts. Once determined, the misbehaving UAV will be given the chance to explain its behaviour and actions in the after-action phase (Section 5.2).

Of course, the service provided by the UAV will affect the weights of importance for the reputation and explainablity. For example, in time critical situations, there is no time for long/complex explanation, and the reputation plays the more significant role.

7 Discussion

Analyzing the solutions proposed in Section 5, Figure 1 summarizes the possible outcomes eliciting the attainable solutions.

In particular, time availability is the predominant factor. If an agent is short in time, explainability might not be an option. Therefore, the agent is demanded to have a trustworthy reputation (proved by the BCT) to achieve a possible agreement. In the case no explanation can be provided and the reputation value is below an acceptable threshold there is no possible solution, and the request of the agent (as we saw in the UAV example above), is rejected.

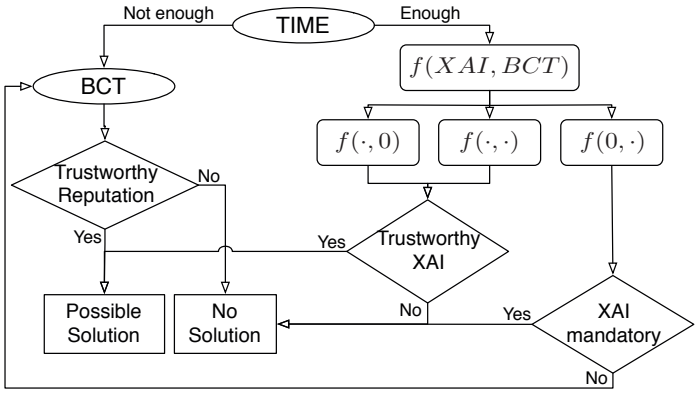

Fig. 1: Decision process wrt available time, explanation, and reputation.

If the available time to produce an explanation is enough, explainability becomes an option. The agent can rely on $f(XAI, BCT)$, a combination of explainability and BCT. The agent might rely only on explainability $f(\cdot, 0)$, only on BCT $f(0, \cdot)$, or on any given combination of both $f(\cdot, \cdot)$. In the latter case, the weights composing this combination mainly depend on the *(i)* specific context, *(ii)* nature of the problem to be explained, *(iii)* explanation capability of the agent and on *(iv)* understanding capability of the agent receiving the explanation. Moreover, on the one hand, having explainability might be necessary and enforced by law. On the other hand, low reputation/trustworthiness of an agent cannot be ignored even if it provided an adequate explanation.

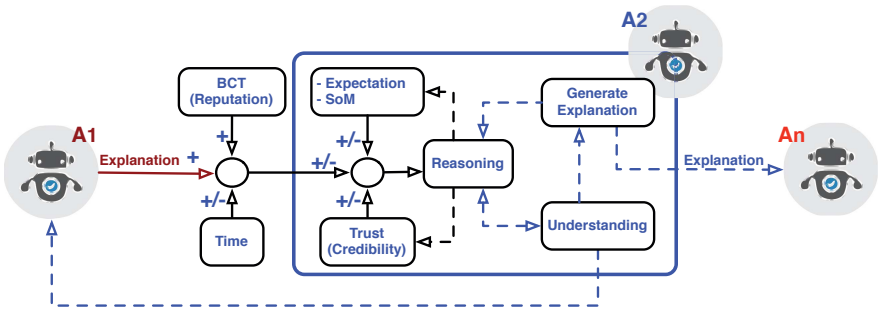

Fig. 2: Representation of the explaination life-cicle.

Figure 2 shows a sequence of interaction within a society of agents, the first agent $A1$ attempts to send an explanation to agent $A2$. Depending on the scenario, $A1$ might possibly be short in time, might possibly be able to rely on BCT for reputation. Based on the explanation/reputation submitted by $A1$ to $A2$, the latter would be able to assess the trustworthiness of $A1$, compare the behavior of $A1$ with its own expectation, and define/update a State of Mind (SoM) about $A1$ intentions. As a result of this reasoning process, $A2$ (delineated by the blue box) builds an understanding of $A1$ and its explanation. Such an understanding is then used to: *(i)* generate an explanation describing $A1$ behavior and communicate it with other agents A_n, *(ii)* refine $A2$'s SoM, reasoning, and expectations about $A1$, and *(iii)* possibly coming back to $A1$ to ask more details/clarifications about its explanation.

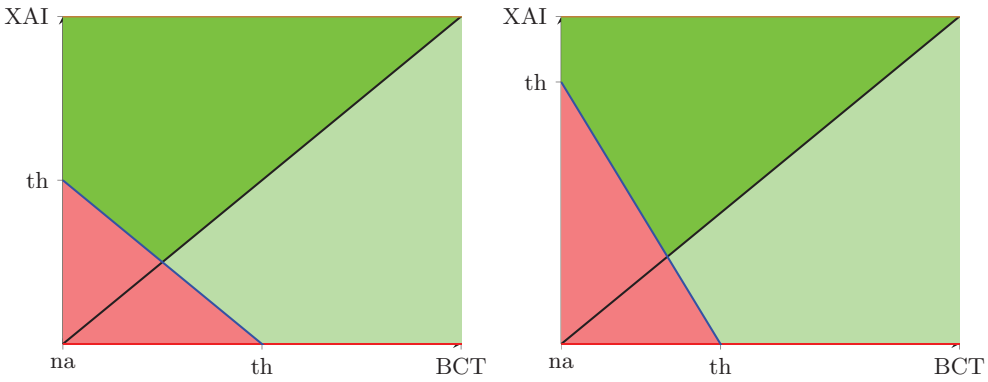

Fig. 3: Symmetric and Asymmetric XAI and BCT contributions

Figure 3 illustrates the possible synergy between XAI and BCT. The blue diagonal line represents the threshold delineating whether the combination of

XAI and BCT satisfies the minimal requirements (area in green) or not (area in red). In the left figure (symmetric case), the contributions from XAI and BCT are equal, thus symmetric. In the right figure (asymmetric case), a contribution from either XAI or BCT has higher impact (XAI in the example in the figure).

In the domain of UAVs, where the regulations are not mature enough [33], the combination of reputation and explainablity will increase the trust of clients in the use of UAVs for package delivery and other applications, while the properly tuned weights given to each factor (reputation and explainablity) will insure that various services could be provided. To acquire the mentioned tuned weights, tests should be conducted. However, some regulations restrict the use of UAVs in cities, so to perform tests with real UAVs, it is needed access to expensive hardware and field tests that usually consume a considerable amount of time and require trained and skilled people to pilot and maintain the UAV. In this context, the development of simulation frameworks that allow transferring real-world scenarios into executable models using computer simulation frameworks are welcome [31].

The design and realization of mechanisms computing trust and reputation of agents communities via blockchain are strictly dependent on the application scenarios and available technologies. Therefore, they are delegated to future studies. Nevertheless, at the current stage, it is possible to provide key research directions. For example, as mentioned in Section 2.1, to undertake the trust evaluation process, agents rely on the social *image* of other agents [41]. An agent constructs such an image relying on *(i) direct experiences* (e.g., direct interactions between the trustor and the trustee), *(ii) communicated experiences* (e.g., interactions between the trustee and another agent communicated to the trustor), and *(iii) social information* (e.g., monitoring social relations and position of the trustee in the society). The interactions and mechanism enabling the computation of trust and reputation can be stored on a blockchain. Thus, depending on the agent *image* retrieved from such a trusted technology, other agents may decide whether granting or not their trust or if demanding for more explanations might be needed to take a more appropriate decision. Yet, concerning privacy and permissions, there are several open questions to be taken into account. Moreover, since agent could communicate their experiences and opinions about other agents behaviors, a mechanism should be devised to prevent malicious agents from adding their unauthentic experiences to the ledger.

8 Conclusions

Most of today's applications are deployed in a distributed and open technical ecosystems involving multiple parties each with different goals constraints. This paper proposed an approach combining BCT and explainability supporting the decision-making process of MAS. Such an approach can remove the current opaqueness of decision-making processes making them interpretable and trustworthy from both agent and human user point of views. It is worth to recall that explanability allows collaborating parties to express their intentions and

reach common understandings and that BCT offers a decentralized authentication mechanisms capable of ensuring trust and reputation management. Then, it identified some applications where the contribution of these technologies revealed to be crucial. Three scenarios have been identified: *(i)* BCT are strictly necessary, *(ii)* explainability is mandatory, and *(iii)* a combination of them is possible and subject to a threshold function. Moreover, several practical case study involving UAVs have been discussed, analyzing roles, balance, and trade-offs between explainability and BCT in trust-dependent systems. This work is an initial step towards building synergies between explainable AI and BCT. The future work is to *(i)* investigate a MAS model suitable for XAI and BCT, *(ii)* design and develop a MAS framework to implement explainable and BCT dynamics, *(iii)* realize smart contracts supporting an efficient communication among light weight devices, *(iv)* assess a possible interdependence among explainability and BCT (in particular involving remote robots such as UAV and HGV), and *iv)* apply and study the developed solutions to UAVs swarm.

Acknowledgments

This work is partially supported by the Regional Council of Bourgogne Franche-Comté (RBFC, France) within the project UrbanFly 20174-06234/06242. This work is partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation.

References

1. ACM, U.: Public policy council: Statement on algorithmic transparency and accountability (2017)
2. Aesop: Aesop's fables. OUP Oxford (2002)
3. Ali, M., Nelson, J.C., Shea, R., Freedman, M.J.: Blockstack: A global naming and storage system secured by blockchains. In: USENIX Annual Technical Conference. pp. 181–194 (2016)
4. Anjomshoae, S., Najjar, A., Calvaresi, D., Framling, K.: Explainable agents and robots: Results from a systematic literature review. In: Proc. of the 18th International conference on Autonomous Agents and Multi-Agent Systems (AAMAS) (2019)
5. Biran, O., Cotton, C.: Explanation and justification in machine learning: A survey. In: IJCAI-17 Workshop on Explainable AI (XAI). p. 8 (2017)
6. Buttazzo, G.C.: Hard Real-Time Computing Systems: Predictable Scheduling Algorithms and Applications, Third Edition, vol. 24 (2011). https://doi.org/10.1007/978-1-4614-0676-1
7. Calvaresi, D., Cesarini, D., Sernani, P., Marinoni, M., Dragoni, A.F., Sturm, A.: Exploring the ambient assisted living domain: a systematic review. Journal of Ambient Intelligence and Humanized Computing **8**(2), 239–257 (2017)
8. Calvaresi, D., Dubovitskaya, A., Calbimonte, J., Taveter, K., Schumacher, M.: 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. pp. 110–126

(2018). https://doi.org/10.1007/978-3-319-94580-4_9, https://doi.org/10.1007/978-3-319-94580-4_9
9. Calvaresi, D., Dubovitskaya, A., Calbimonte, J.P., Taveter, K., Schumacher, M.: Multi-agent systems and blockchain: Results from a systematic literature review (2018)
10. Calvaresi, D., Dubovitskaya, A., Retaggi, D., Dragoni, A., Schumacher, M.: Trusted registration, negotiation, and service evaluation in multi-agent systems throughout the blockchain technology. In: International Conference on Web Intelligence (2018)
11. Calvaresi, D., Leis, M., Dubovitskaya, A., Schegg, R., Schumacher, M.: Trust in tourism via blockchain technology: Results from a systematic review. In: Information and Communication Technologies in Tourism 2019, pp. 304–317. Springer (2019)
12. Calvaresi, D., Marinoni, M., Dragoni, A.F., Hilfiker, R., Schumacher, M.: Real-time multi-agent systems for telerehabilitation scenarios. Artificial intelligence in medicine (2019)
13. Calvaresi, D., Marinoni, M., Sturm, A., Schumacher, M., Buttazzo, G.C.: The challenge of real-time multi-agent systems for enabling iot and CPS. In: Proceedings of the International Conference on Web Intelligence, Leipzig, Germany, August 23-26, 2017. pp. 356–364 (2017). https://doi.org/10.1145/3106426.3106518
14. Calvaresi, D., Mattioli, V., Dubovitskaya, A., Dragoni, A.F., Schumacher, M.: Reputation management in multi-agent systems using permissioned blockchain technology (2018)
15. Carey, P.: Data protection: a practical guide to UK and EU law. Oxford University Press, Inc. (2018)
16. Castelfranchi, C.: Engineering social order. In: International Workshop on Engineering Societies in the Agents World. pp. 1–18. Springer (2000)
17. Chandrasekaran, B., Tanner, M.C., Josephson, J.R.: Explaining control strategies in problem solving. IEEE Intelligent Systems (1), 9–15 (1989)
18. Dictionary, O.E.: Compact oxford english dictionary (1991)
19. Doran, D., Schulz, S., Besold, T.R.: What does explainable ai really mean? a new conceptualization of perspectives. arXiv preprint arXiv:1710.00794 (2017)
20. Dragoni, A.F., Sernani, P., Calvaresi, D.: When rationality entered time and became a real agent in a cyber-society. In: Proceedings of the 3rd International Conference on Recent Trends and Applications in Computer Science and Information Technology, RTA-CSIT 2018, Tirana, Albania, November 23rd - 24th, 2018. pp. 167–171 (2018), http://ceur-ws.org/Vol-2280/paper-24.pdf
21. Ferrer, E.C.: The blockchain: a new framework for robotic swarm systems. In: Proceedings of the Future Technologies Conference. pp. 1037–1058. Springer (2018)
22. Forouzan, B.A.: Cryptography &Amp; Network Security. McGraw-Hill, Inc., New York, NY, USA, 1 edn. (2008)
23. Fox, A., Griffith, R., Joseph, A., Katz, R., Konwinski, A., Lee, G., Patterson, D., Rabkin, A., Stoica, I.: Above the clouds: A berkeley view of cloud computing. Dept. Electrical Eng. and Comput. Sciences, University of California, Berkeley, Rep. UCB/EECS **28**(13), 2009 (2009)
24. Gambetta, D., et al.: Can we trust trust. Trust: Making and breaking cooperative relations **13**, 213–237 (2000)
25. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., Pedreschi, D.: A survey of methods for explaining black box models. ACM Computing Surveys (CSUR) **51**(5), 93 (2018)
26. Gunning, D.: Explainable artificial intelligence (xai). Defense Advanced Research Projects Agency (DARPA), nd Web (2017)

27. Hellström, T., Bensch, S.: Understandable robots-what, why, and how. Paladyn, Journal of Behavioral Robotics **9**(1), 110–123 (2018)
28. Kass, R., Finin, T., et al.: The need for user models in generating expert system explanations. International Journal of Expert Systems **1**(4) (1988)
29. Kuzmin, A., Znak, E.: Blockchain-base structures for a secure and operate network of semi-autonomous unmanned aerial vehicles. In: 2018 IEEE International Conference on Service Operations and Logistics, and Informatics (SOLI). pp. 32–37. IEEE (2018)
30. Miller, T.: Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence (2018)
31. Mualla, Y., Bai, W., Galland, S., Nicolle, C.: Comparison of agent-based simulation frameworks for unmanned aerial transportation applications. Procedia Computer Science **130**(C), 791–796 (2018)
32. Mualla, Y., Najjar, A., Boissier, O., Galland, S., Tchappi, I., Vanet, R.: A cyber-physical system for semi-autonomous oil&gas drilling operations. In: 5th Workshop on Collaboration of Humans, Agents, Robots, Machines and Sensors. Third IEEE International Conference on Robotic Computing (2019)
33. Mualla, Y., Najjar, A., Galland, S., Nicolle, C., Tchappi, I., Yasar, A.U.H., Frmling, K.: Between the megalopolis and the deep blue sky: Challenges of transport with uavs in future smart cities. In: Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems. International Foundation for Autonomous Agents and Multiagent Systems (2019)
34. Najjar, A.: Multi-agent negotiation for qoe-aware cloud elasticity management. Ph.D. thesis, PhD thesis, École nationale supérieure des mines de Saint-Étienne (2015)
35. Najjar, A., Serpaggi, X., Gravier, C., Boissier, O.: Survey of elasticity management solutions in cloud computing. In: Continued Rise of the Cloud, pp. 235–263. Springer (2014)
36. Najjar, A., Serpaggi, X., Gravier, C., Boissier, O.: Multi-agent systems for personalized qoe-management. In: Teletraffic Congress (ITC 28), 2016 28th International. vol. 3, pp. 1–6. IEEE (2016)
37. Neerincx, M.A., van der Waa, J., Kaptein, F., van Diggelen, J.: Using perceptual and cognitive explanations for enhanced human-agent team performance. In: International Conference on Engineering Psychology and Cognitive Ergonomics. pp. 204–214. Springer (2018)
38. Nomura, T., Kawakami, K.: Relationships between robot's self-disclosures and human's anxiety toward robots. In: Proceedings of the 2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology-Volume 03. pp. 66–69. IEEE Computer Society (2011)
39. Parkes, D.C., Wellman, M.P.: Economic reasoning and artificial intelligence. Science **349**(6245), 267–272 (2015)
40. Reiter, U., Brunnström, K., De Moor, K., Larabi, M.C., Pereira, M., Pinheiro, A., You, J., Zgank, A.: Factors influencing quality of experience. In: Quality of experience, pp. 55–72. Springer (2014)
41. Sabater-Mir, J., Vercouter, L.: Trust and reputation in multiagent systems. Multiagent Systems p. 381 (2013)
42. Samek, W., Wiegand, T., Müller, K.R.: Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. arXiv preprint arXiv:1708.08296 (2017)

43. Simon, J., High, D.R., Wilkinson, B., Mattingly, T., Cantrell, R., John, J.O.V., McHale, B., Jurich, J., et al.: Managing participation in a monitored system using blockchain technology (Aug 2 2018), uS Patent App. 15/881,715
44. Swanson, T.: Consensus-as-a-service: a brief report on the emergence of permissioned, distributed ledger systems (2015)
45. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., Fergus, R.: Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199 (2013)
46. Talia, D.: Clouds meet agents: Toward intelligent cloud services. IEEE Internet Computing **16**(2), 78–81 (2012)
47. Walmart Retail Company: Walmart wants blockchain to make shipping smarter. `https://mrtech.com/news/walmart-wants-blockchain-to-make-shipping-smarter/` (03 2018)
48. Wang, W., Niu, D., Li, B., Liang, B.: Dynamic cloud resource reservation via cloud brokerage. In: 2013 IEEE 33rd International Conference on Distributed Computing Systems. pp. 400–409. IEEE (2013)
49. Zhang, Q.s., Zhu, S.C.: Visual interpretability for deep learning: a survey. Frontiers of Information Technology & Electronic Engineering **19**(1), 27–39 (2018)