# Perspectives on the Explanation of Organizations in Multi-Agent Systems
## (Blue Sky Ideas)

Elena Yan[1 [0009−0000−6660−9378]], Luis G. Nardin[1 [0000−0002−4506−2745]], Alessandro Ricci[2 [0000−0002−9222−5092]], Samuele Burattini[2 [0009−0009−4853−7783]], Guillaume Muller[3 [0000−0003−3740−9232]], Jomi Fred Hübner[4 [0000−0001−9355−822X]], Jaime S. Sichman[5 [0000−0001−8924−9643]], and Olivier Boissier[1 [0000−0002−2956−0533]]

[1] Mines Saint-Etienne, Univ Clermont Auvergne, INP Clermont Auvergne, CNRS, UMR 6158 LIMOS, F-42023 Saint-Etienne France
{elena.yan,gnardin,olivier.boissier}@emse.fr
[2] Alma Mater Studiorum, University of Bologna - Cesena Campus, 47521 Cesena, Italy {a.ricci, samuele.burattini}@unibo.it
[3] Mines Saint-Etienne, Institut Henri Fayol, F-42023 Saint-Etienne France
guillaume.muller@emse.fr
[4] Federal University of Santa Catarina, Florianópolis, Brazil jomi.hubner@ufsc.br
[5] Laboratório de Técnicas Inteligentes (LTI), Escola Politécnica (EP), Universidade de São Paulo (USP), São Paulo, Brazil jaime.sichman@usp.br

**Abstract.** Explainability in Multi-Agent Systems (MAS) has emphasized the explanation of the behavior of individual autonomous agents. Since agents in MAS operate under an implicit or explicit organization, explainability must go beyond the individual agents and address the behavior of collectives of agents within an organization. We argue that explainability in MAS should account the different facets constituting the organization, which has not yet been explored in the literature. We discuss perspectives, possibilities, arguments, and visions for explaining organizations in the context of MAS.

**Keywords:** Explainability of Organizations · Explainability in Multi-Agent Systems · Multi-Agent Systems

## 1 Introduction

Explainability has emerged as a desirable non-functional requirement for Artificial Intelligence (AI) based systems [25,56]. Although investigated since the 1980s [49,104], eXplainable AI (XAI) [107] has lately focused on making Machine Learning (ML) models interpretable by explaining their opaque algorithms [2,21,45,48,70]. Despite increasing predictability, ML models typically remain steeped in methodological individualism [34,82], while many applications demand social understanding and cooperative intelligence [30].

Multi-Agent Systems (MAS) [105] capture these social and cooperative features by allowing multiple autonomous agents to interact in a shared environment to achieve their delegated goals, eventually within one or more *organizations* [97]. Organizations both shape and enforce the agents' behavior in MAS. They facilitate agents in MAS to act collaboratively to achieve system goals that would otherwise not be achievable (or not as easily) [42,1]. Organizations have been investigated in multiple fields of study, such as sociology [93,38,31,17], ethology [36], and organizational management [18,95,67,92,99]. These studies and their theories served as the basis for various organizational models in MAS (e.g., [52,77,37,74,5,41]).

Recent XAI efforts have been directed toward the explanation of the behavior of individual agents (e.g., [20,32,89,103,102,108]) under the label of *explainable agents*, i.e., agents that have the ability to explain their decisions and the reasoning that produced their choices [59]. Although Winikoff [102] argues that explainability in MAS should be designed from a multi-component perspective in which each component tackles different aspects of the system, few studies have been conducted to date to include components other than agents to enhance explainability. This narrow focus limits the explainability power of the existing approaches, as explaining the behavior of collectives of agents within an organization cannot be reduced to explaining individually the behavior of each agent [75]. For example, consider a decrease in overall productivity in an organization: while individual actions (e.g., operators postponing a task, a manager following a new procedure) are performed with local intentions, one cannot say that these individuals had the intention of lowering the organization's performance; the reduced productivity is rather an emergent effect of the behavior of many individual agents operating in the organization.

Therefore, we advocate the importance of investigating the explanation of organizations in MAS. This paper discusses the explanation of organizations from different viewpoints on organizations in MAS [16] based on two orthogonal dimensions: from agent-centered to organization-centered view, on one side, and from organization unawareness to awareness, on the other side.

Along these view points, we discuss explanations that reflect (i) the behavior of an individual agent whose behavior is shaped by an organization; (ii) the behavior of collectives of agents whose coordination is shaped by an organization; (iii) the organization components that process and structure the global behavior of the MAS (no details in terms of agents); (iv) the decisions of the designer of the organization; and (v) the state and dynamics of the organization. We then discuss what can be explained in an organization and how these explanations can be exchanged among entities in an MAS.

The paper is structured as follows. Section 2 presents the related work. Section 3 discusses the different perspectives on organizations with respect to explainability. Section 4 discusses the possible ways to explain the organization. Section 5 discusses the challenges and considerations in designing and engineering explanations of organizations. Finally, Section 6 concludes with our findings and visions.

## 2 Related Work

In general terms, explanation is the act of someone (the *explainer*) making someone else (the *explainee*) understand something by providing *explanations*, i.e., reasons or context unfolding the meaning surrounding a concept or information [27,68]. This broad definition has been investigated in various ways. Because we focus on explaining organizations in MAS, we start by reviewing the literature on explainability in MAS (Section 2.1). Next, we examine how the social sciences have looked at the explainability of organizations (Section 2.2). Finally, we summarize the key points that help us structure the analysis on explainability of organizations in MAS (Section 2.3).

### 2.1 Explainability in MAS

Explainable agent is an emerging field that focuses on agents explaining their own behavior and actions by referring to internal mental states and local perceptions [59]. Anjomshoae et al. [7] identify three main phases of explanation: *explanation generation*, *explanation communication*, and *explanation reception*. It has been recognized in [11,32,26] that the interpretation and reaction of the agent to the explanation are relevant, transforming communication and reception into rich dialogue patterns.

Many studies have focused on the first phase, emphasizing the type of information used to generate explanations. In Belief-Desire-Intention (BDI) [19,85] agents, the set of explicit concepts of the agent's mental state (e.g., belief, desire, intention, and plan) can use as types of information to generate explanations, supporting *explainability by design* [98,88]. In this context, Rodriguez et al. [89] propose explainability design patterns to engineer BDI agents explainable by design.

Winikoff [102] states the need for engineering explanation from a single component to multi-component systems. He proposes an architecture that enables users to ask questions and receive explanations from different explainer agents. Harber et al. [46] design different explanation algorithms in which, for instance, BDI agents' actions are explained based on the agent's current beliefs and goals, or the next action or goal. Winikoff et al. [103] build on these algorithms the concept of valuings (i.e., positive or negative effect toward the action) to explain the preferences of the agent's options. Broekens et al. [20] evaluate explanation generation algorithms, concluding that different agents actions need different explanation algorithms.

Other studies considered the two other phases from the perspective of the explainee. Yan et al. [108] use multiple levels of abstraction to generate explanations that are suitable for developers, designers, and end-users. The explanation for end-users is also discussed in [66]. Expanding these two phases, a subset of studies has investigated dialogue or argumentation in BDI agents to communicate the explanations. Dennis et al. [32] present explanations as dialogue, where agents or human participants can interact with and interrogate another agent to understand the divergences between what has happened and what is expected.

Panisson et al. [78] use an argumentation approach, where agents can generalize beliefs by evaluating accepted arguments and then explaining those beliefs in terms of arguments that support them. Similarly, in [71], the reason for the agent's actions is explained by arguments that led to that decision. Explanations can also be exchanged among agents themselves [26,76] referring to *inter-agent explainability* [11].

We have found few studies that have connections to the explanation in the context of organizations. In [3], agents exchange explanations for collaborative workflows. In [47], the agents' explanations are provided in the context of collaboration in human-agent teamwork. Langley [58] uses the concept of an explainable agent, referred to as the *justified agent*, and defined as having the ability to explain one's activities in terms of norms. Other studies have explored the justification for norm compliance or deviation [12,62]. A justification explains why a decision is good, but it may not give an explanation of the actual decision-making process [14]. Organization is also cited not for its explainability but as a means to regulate the explanation process in MAS by the agents. For example, accountability in organizations [9] enables agents to provide an account of what happened. Explainability supports accountability in explaining and identifying who is accountable in complex situations or where no one is (or feels) accountable for a mistake [57]. Explainability combined with accountability supports the recovery from failure at the system level [8].

Although rich, the literature about XAI in MAS does not provide any study that directly addresses the explainability *of* organizations in MAS.

### 2.2 Explainability of Organizations in Social Sciences

Here, we provide insights from the social sciences on organizations. We focus on the aspects that are relevant to the explainability of organizations.

When studying how individuals interact in a society, sociology focuses on informal and implicit organizations that emphasize how the collective and social groups behave. These studies have identified that people explain the behavior of the group differently depending on how the group is constituted [75]. For aggregate groups of unrelated individuals (e.g., shoppers in a department store), explanations tend to appeal to causal history reasons, such as situational factors (e.g., a sale). In contrast, for groups of jointly acting individuals (e.g., friends shopping together), explanations tend to appeal to intentional reasons that refer to shared desires and goals (e.g., desire to meet together). This aligns with earlier work by Kass and Leake [55], who distinguish intentional explanations from social explanations that account for behavior without attributing explicit intentions. Following Searle's analysis of social reality [93], collective intentionality cannot be reduced to individual intentionality, and when constituting an institution, institutional phenomena cannot be fully explained by reference to explicit rules, norms, or decision procedures alone. Rather, humans evolve a set of dispositions that are sensitive to the rule structure. The dispositions explain the behavior, and the dispositions are explained by the system of rules.

When studying formal and explicit organizations [24], we can highlight works in organizational theory [67,92] that deals with the structures and operations of formal social organizations. Two global explanations of the organization can be considered [67]. One that explains its emergence as a necessary response to changes in the environment of the organization, while the other explains "strategical" changes when managers follow new ideas and objectives about how to manage. From members of the organization, we can distinguish other complementary types of explanations [95]: (i) personal motivation (e.g., explanation of "why do you keep (or take) this job?"), and (ii) organizational goals (e.g., explanation of "why do you make this particular investment decision?"). The organization can be understood by the organization knowledge transfer between people in the same organization or from different organizations [99].

Note that explanations in settings with an explicit organization tend to refer to organizational knowledge in terms of goals, decisions, roles, and other elements, while explanations in settings with an implicit organization tend to refer to the causal history or evolution of the organization.

### 2.3 Synthesis

From the literature, we identify that explainability has various aspects. More broadly, explainability is a process of generating, communicating, and receiving explanations about something by providing reasons or context. Explanation algorithms can generate different types of explanations, aimed at explainees of different types (e.g., humans or agents) or playing different roles (e.g., stakeholder, regulator, designer). These explanations range from individual to collective and organizational, and demand various levels of abstractions. Explanations can often be generated by composing partial explanations originating from different entities and aspects. Explainability can be supported by elements incorporated at the design time, i.e., explainability by design.

In the context of explaining organizations, systems can be engineered so that organizational concepts and organizational events are explicitly represented to facilitate explanation. This approach resonates with proposals to make accountability a first-class design abstraction that can make organizational structures and processes traceable and justified.

## 3 Perspectives on Organizations Through the Lens of Explainability

In this section, we first revisit the classification of organizations in the literature from the point of view of explainability (Section 3.1). Next, we detail what can be explained in an organization (Section 3.2) and then the different views on explaining organizations (Section 3.3).

### 3.1 Analysis of the Organizations Grid

Here, we use the classification of organizations (i.e., from agent- to organization-centered and organization unawareness to awareness) proposed in MAS [16], to present, in the following subsections, different views on explaining organizations in light of the literature on explainability. The possible organizations in the MAS domain that ground our perspectives on explaining organizations are:

- *Emergent Organization*: Organizations are the result of the collective emergent behavior of agents interacting in a shared and dynamic environment. These agents are unaware of the organization that they indirectly build from their individual behaviors. Examples are studies in swarm-based and self-organization systems, e.g., ant colonies [36].
- *Agent-Centered Organization*: Organizations are built from patterns of cooperation in a bottom-up process in which agents are aware of some collective behavior from a local and maybe inconsistent point of view. Examples are studies on social mechanisms within agents, e.g., social reasoning [94], social commitments [54,96]).
- *Designed Organization*: Organizations are specified by an external designer through an organization specification that is hard-coded into the agents, i.e., agents are not able to represent and reason about the organization (organizations are design models used to develop agents). Examples are studies done in the field of Agent-Oriented Software Engineering [13], e.g., GAIA [106], INGENIAS [80].
- *Organization-Centered Organization*: Organizations are designed by an external designer or the participating agents of the system themselves through an organization specification that is also accessible to the agents. Being accessible enables agents to reason about the organization, to use the specification to decide how to cooperate with other agents, or to adapt and design the organization specification at runtime (e.g., [35,53]). Examples are studies done in the field of MAS Organization-Oriented Programming [84], e.g., MOISE [52], Opera [5].

In the next section, we examine in detail the elements that can be explained in the organization.

### 3.2 Organization Facets

Having the global analysis of organizations in MAS, let us now consider what can explain and be explained in an organization. The *objects* of explanation can be classified according to two *organization facets*: *organization abstractions* denote concepts that define an organization and *development constructs* denote artifacts produced along the development lifecycle phases of an organization.

***Organization Abstractions*** The organization abstractions are used to describe the collective activity of the organization. Given the organization concepts studied in the literature (e.g., [31,42,61,52,87]), here we structure them into four main concepts: purpose, structure, function, and regulation.

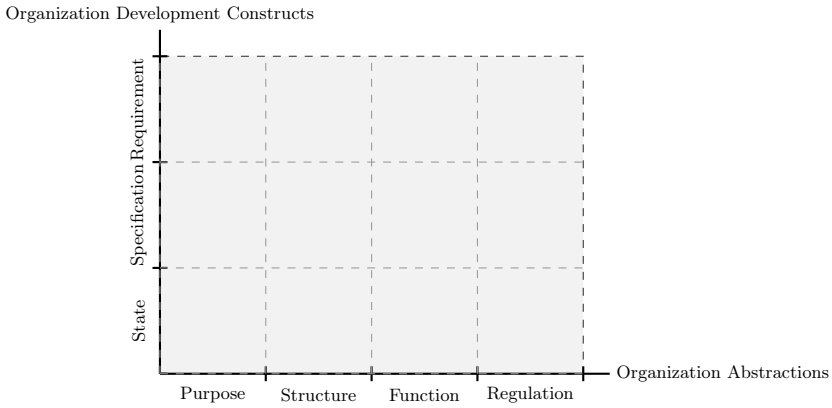


**Fig. 1.** Explainable aspects in an MAS organization. Each concept of organization abstractions (i.e., purpose, structure, function, and regulation) participates in each definition of development constructs (i.e., requirement, specification, and state).

- The *organization purpose* brings agents to work together and guides them toward achieving this purpose. Organization purposes may include common objectives, organizational goals, and organizational values [18] that the organization wants to achieve.
- The *organization structure* corresponds to structures similar to those in human organizations, such as departments, positions, and their interrelations. The organization structure defines a structured pattern of behavior that enhances the coordination of agent activities [100] may assume different forms. Generally, the organization's structure is influenced by the type of coordination in the system (i.e., market, network, or hierarchy) [33]. The structure can be subdivided into groups and subgroups. Each group can be composed of roles that denote the position that individual agents can occupy in the structure of an organization. Typically, roles are related to each other through relationships that allow building the social and interaction network in an organization.
- The *organization function* defines the functioning of the organization, for instance, the specification of global plans, policies to allocate tasks to agents, the coordination to execute a plan, and the quality (time consumption, resources usage, etc.) of a plan.
- The *organization regulation* comprises all constraints, norms, duties, sanctions, etc., that are used to govern the autonomy of the agents with respect to their structuring and coordination of activities. Three fundamental types of regulations are usually defined in MAS [15,51,65]. Constitutive norms make the link from the environment and agents to construct the social reality [93] in the organization. Regulative norms govern the behavior of agents by defining the obligations, permissions, and prohibitions in performing an action or achieving a state [81]. When a regulative norm is unfulfilled, it is

possible to sanction by specifying the consequence for unfulfillment [73] or to provide an account of what happened [9].

***Organization Development Constructs*** The organization development constructs relate to the phases of the organization development life cycle.

- The *organization requirement* is the result of the analysis phase, defining a set of functional requirements (e.g., the organization must allocate at least one agent to monitor critical system alerts) and non-functional requirements (e.g., the organization must comply with the sustainability regulations).
- The *organization specification* is the result of the design phase, defining explicitly what is expected from and what is enforced on the agents in an organization by considering the organization requirements. The organization specification is given as a top-down approach to coordinate and govern the behavior of the agents in an organization.
- The *organization state* is the instance of the organization specification produced in the execution phase, given an organization specification and behavior of the agents.

### 3.3 Views on Explaining Organizations

Having presented the classifications of the organizations and the different elements that can be explained in an organization, we now complete the discussion about the different views on explaining organizations that are presented in Figure 2. There are different explainer entities that can explain the organization: they can be *designer* or *observer* (depending on the explainer's role) and have an *external* or *internal* point of view.

***Explanation from an Emergent Organization View*** In this view (Figure 2(a)), organizations only exist as observable emergent bottom-up phenomena that can be explained by some external observer entity. Observers can explain the observed organization state from the patterns of behavior of the agents, in terms of the purpose, structure, function, and regulation.

***Explanation from an Agent-Centered Organization View*** In this view (Figure 2(b)), explanations can be produced by one or more external or internal observer entities. External observer entities may explain the organization state based on their observation of the global state of the collective behavior, from which they can infer some organization (e.g., the purpose from the common goals or the function and regulation from the regular cooperation patterns). Internal entities in the organization are able to provide explanations of their internal and local representation of the collective behavior, based on their internal observation and participation in the organization. For instance, explanations can refer to the agents' social mechanisms to infer the local representation of the organization (e.g., explain the agent's internal organization structure in terms of the agent's social network dependencies). Explanations are subjective to the entities, possibly inconsistent with what the other entities provide, due to the locality and absence of a global and unique representation.

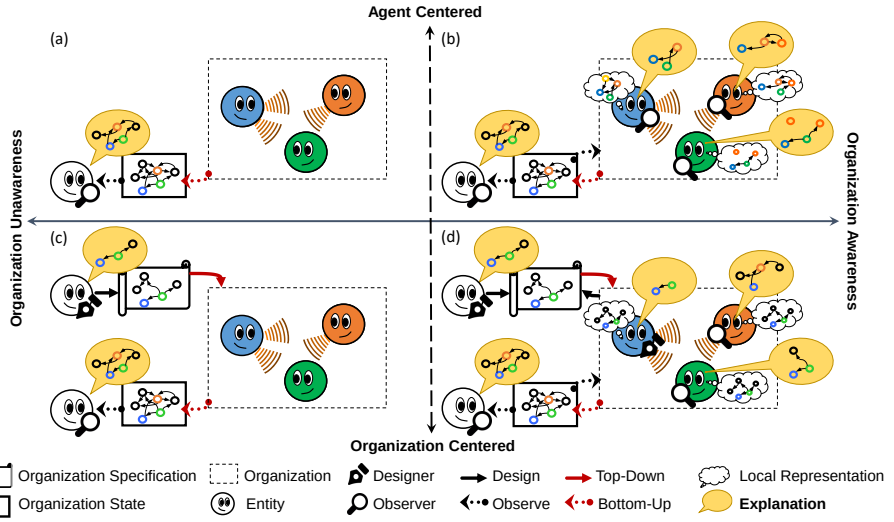


**Fig. 2.** Views on explaining organizations based on the classification of organizations, adapted from [16].

***Explanation from a Designed-Organization View*** In this view (Figure 2(c)), explanations can be produced by an external observer entity or an external designer entity. Besides the current state of the observed organization, the external designer can provide explanations of the organization requirements that led to the design of the organization specification. In the case in which a designer and an observer are the same entity, they can compare the organization specification and state to explain what is designed and what is observed.

***Explanation from an Organization-Centered Organization View*** In this view (Figure 2(d)), explanations are produced by an external observer entity, the designer of the organization specification, and by the participating entities in the organization that can reason about the organization. This view offers the most complete possible set of explanations about the organization including the largest possible types of explainer entities. These explanations may differ from one another since they may emphasize distinct aspects of the organization. External designer and observer entities can provide explanations about the specification and state of the organization, respectively. Internal designer and observer entities can also provide explanations about the specification and state of the organization, but they are constrained to their observability. The designer and observer may be the same entity, who can use both the organization specification and state to explain. Participants of the organization may have access to the organization specification and state of the organization, explaining the organization from an *internal* point of view. They can thus play the role of *observer* or *designer* in addition to participating in the organization.

## 4 Explaining Organizations

Given the different perspectives on explaining organizations, we go into the details of how we can share those explanations in MAS. We discuss the possible explanation questions that can be asked (Section 4.1) in explaining organizations from possible explainers (Section 4.2) to possible explainees (Section 4.3).

### 4.1 Explanation

A central aspect in the explanation is the ability to provide reasons and context to the questions that the explainee has asked [64]. Questions that have been studied, e.g., offer justifications [25] (i.e., why questions), provide transparency [83] (i.e., what and how questions), or are contrastive [64,69] (i.e., what if and why not). Here, we illustrate some of these questions by explaining the organization, which can cover the organization facets defined in Section 3.2.

- *What*: questions about the aspects of the organization, such as the function defined in the organization specification (e.g., "What is the role allocation? What are the obligations of this task?"), the organization state or evolution (e.g., "What happened in the organization? What is the current role allocation?").
- *What If*: counterfactual questions about alternatives, such as alternatives in changing the organization specification (e.g., "What if we reassign this role? What if this norm were relaxed?").
- *Why*: causal or justificative questions about the aspects of the organization (e.g., "Why did this group make that decision? Why was this agent sanctioned?").
- *Why Not*: questions about unexpected states (e.g., "Why did the group not follow the norms?") or misalignment with the specification and state of the organization (e.g., "Why was this plan not adopted?").
- *How*: procedural or prescriptive questions about, for instance, the function of the organization (e.g., "How to achieve this organization goal? How can one join this organization?").

Note that there are different possible ways and algorithms to generate the explanations to these questions. For instance, explanations may refer to the aspects in the organization requirement, specification, or state, on the temporal evolution and history of these aspects, the entities that operate in the organization, or the knowledge and personal motivations of the explainer (in case of humans or agents). Explanations also depend on the role and expectations of the explainees.

### 4.2 Explainer

The *explainer* is the entity that provides explanations. In Section 3, we stated that the explainer can be an observer or designer, and an external or internal entity of the organization. Considering the MAS as a Socio-Technical System, we can generalize the explainer as humans, components, or agents that provide explanations of the organization.

***Humans*** In all views presented in Figure 2, humans playing different roles can be the explainer entity in the explanation of an organization.

In the one hand, *stakeholders* involved in the development of the organization in MAS can explain their choices. *Domain experts* may explain the functional and non-functional requirements in constraining the design and implementation of the organization. *Regulators* may explain the legal, moral, or ethical functional and non-functional requirements that led to define the organization requirements. *Designers* may explain the requirements and architectural choices that led to the design of the organization specification. *Developers* may explain the settings and implementation choices that led to the implementation of the organization specification (as per [22]).

In the other hand, *end-users* may observe the organization state or design the organization specification and explain their requirements, preferences, or constraints, thereby clarifying the expectations against which organizational behavior should be interpreted and accountable.

***Agents*** Individual agents participating in the organization can also provide explanations. However, explanations from an individual agent are influenced by their intentionality and present a form of subjectivity. From an organizational perspective, explanations also vary depending on the explainer's role, position, or relationship in the structure. For instance, operator agents may have only partial access to the state of the organization. Their explanations may tend to emphasize particular aspects of sub-goals and sub-tasks (e.g., how is the process for this task) that contribute to the overall organizational goals. Conversely, director or manager agents may produce a comprehensive explanation that appeals to organizational objectives, structure, or functions (e.g., what are the plans to achieve that organization goal). Regulator agents may provide explanations referring to the regulations (e.g., why this norm is enforced) and to decisions in managing the regulations (e.g., why a sanction is applied). Agents may also produce explanations about the accountability of something in the organization. They may justify and identify the accountable agent. Agents external to the organization, on the other hand, do not have full access to the organization. They may justify their choice for the organization based on their external observation.

Beyond individual agents, one can also consider *collectives of agents* as the explainer of explanations. In this case, the explanation is social and can be produced from the result of explanations of individuals in the collective (e.g., by collaboration, argumentation, or aggregation), taking into account the collective goals, intentions, plans, etc. For example, a team or department might provide a joint explanation of a collective decision.

***Components*** We may also consider dedicated *components* (i.e., the technical means used to create and manage explanations) can be designed to produce explanations of the organization. The explanation depends on the function held and algorithm implemented on these components. They can function as a *designer component* by producing explanations about the algorithms for designing the organization, an *observer component* by producing explanations about the

algorithms for observing and producing explanations (e.g., they may encode, aggregate, or reinterpret the observation or explanations from the organization), or an *organization component* by producing explanations about the algorithms that contribute to the structure and function of the organization. This may raise questions about responsibility (e.g., who is accountable for the content of these mediated explanations), alignment (e.g., how faithfully they reflect the organizational rationale), and security (e.g., vulnerability to attacks affecting the integrity of the explanation).

### 4.3 Explainee

The *explainee* refers to the entity that receives the explanation. Usually, the explainee can also initiate the request for explanation and later receive the explanation from the explainer entity. The explainee should be considered in the generation of the explanation because the content and the vocabulary used may change. Here, we only consider two global categories of explainees out of the three considered in the previous section: humans and agents. Components are not taken into account here, as they are passive entities.

***Humans*** Different stakeholders have been explored in receiving explanations based on their needs [57], here we provide some examples.

Explanations can be produced to *end-users* who may want a better understanding of the organization to increase their trust before joining it. This is especially relevant in Human–AI Systems [72] and Socio-Technical Systems [10,97], where the human interacts within the system and needs to understand, for instance, why the organization behaves in a certain way and how certain decisions are taken.

Explanations can aim at stakeholders in the development of MAS from an engineering perspective [108]. *Developers* can exploit explanations to debug and identify runtime failures [4,101], as proposed in [50], where debugging refers to explaining why a behavior is produced. *Designers* and *software architects* can use explanations to validate the MAS behavior in accordance with the system requirements [6,23,90] and to check the requirements in an Agile Methodology [60]. *Domain experts* are interested in the system functionalities rather than technical details or design and architectural choices.

In addition, explanations can be produced to *regulators*, who are responsible for improving legal certainty and facilitating the assessments of: legal regulation compliance (e.g., to GDPR or AI Act and contributing to initiatives such as the EUSAiR project[6]), responsible and accountable parties [9], moral and ethical standards [40], the fairness of organizations' structure [91] and function [28], and the transparency and trustworthiness of the organization [44].

***Agents*** Explanations can also be addressed to agents themselves [76]. In the context of organizations, explanations can be seen as organizational knowledge

[6] EU Regulatory Sandboxes for AI (EUSAiR) project: https://eusair-project.eu/

transfer [99], but explanations go beyond the mere transmission of knowledge by providing reasons and causes, and not just facts.

We distinguish the cases in which agents are external or participants in the organization. For participant agents, the explanation is useful, for instance, to clarify the how and why of the organizational dynamics for the purpose of completing the tasks, better collaboration, identifying inconsistencies, etc. For instance, the explanation can be produced by operators to managers, explaining how they have completed or why they have not completed their task; or from managers to operators, explaining what are the plans to achieve certain organizational goals or why certain organizational values are important.

For agents external to the organization, explanations are useful, for instance, when agents are considering entering the organization in open MAS [31]. They may first require some explanations about the functioning of the organization before joining. For instance, external agents that the organization is offering the service to or governing their behavior can request explanations from the organization to better understand their functions (e.g., how the production flow of the product is defined). In all these cases, explanations may need to be selectively constrained by confidentiality, security, or strategic considerations.

Explanations can also be addressed to *collective groups*, either within (e.g., a department, project team, or committee) or outside (e.g., user communities or external stakeholders) the organization. Explanations can be communicated or broadcast through shared channels, such as official documents, reports, dashboards, or public announcements, so that multiple recipients can align their understanding of the organization's behavior and decisions.

### 4.4 Remarks

All the previous classifications of explainer and explainee produce a quite large set of types of exchange of explanations for an organization (e.g., human-agent, human-component, agent-agent explanations). When considering humans as the explainer, a challenge is to make the agent able to interpret and react to the explanation. This could be integrated with the new possibilities offered by Large Language Models (LLM)-based agents and MAS [63]. When considering the agent as explainer and explainee, agents should be able to generate, communicate, receive, interpret, and react to explanations. This is the direction pushed in the *inter-agent explainability* [11] research line. Organization is essential when considering multiple agents interacting and organizing their activities.

## 5 Discussion

Given the rich perspectives identified in this work, one important challenge consists of engineering explanations of organizations in practice, realizing *explainable organizations*. We argue that using only explainable agents or dedicated components is not enough. Depending on the perspective of the organization in MAS and the engineering requirements for explainability, it is necessary to design different modules to integrate the ability to explain the organization into agents or dedicated components.

An important point to be considered is selecting the appropriate *vocabulary* in generating the explanation of the organization. The vocabulary can directly refer to the organization facets in organization-centered perspectives. Having explicitly designed and represented the organization concepts (e.g., [39,29,79,43]) in the organization model can support the explanation of the organization *by design*. Without explicit organizational concepts, additional efforts should be made in generating *post-hoc* [86] explanations of the organization. From an agent-centered organization perspective, agents can explain their organized behavior in terms of social dependencies, social networks, and relationships. From an emergent organization perspective, the explanation is inferred from the observation. We can also envision that, since the work in explainability based on the BDI model has a solid foundation in the literature, one could exploit these foundations by mapping and providing explanations about the organization dynamics in terms of the beliefs, desires, and intentions of the organization.

It is also important that the explainee entity, after receiving the explanation, is able to understand and interpret the explanation in order to react accordingly. Complementary to the vocabulary is the *level of detail* of the explanation based on the explainee [108]. Agents require technical and machinery explanations that are easy to encode [11]. Developers require explanations that align with the technology used to develop the organization. Designers require explanations that can reflect the design principles and architecture adopted in developing the organization. Domain experts require higher-level explanations that can be compared with the system's functional and non-functional requirements. Finally, regulators require explanations about regulations that can then be compared to legal regulations.

Another challenge is to be able to identify and explain the *cause* of certain states of the organization. The explanation of the causes varies from the explainer entity. For instance, humans and agents do not give complete explanations that cover all factors, but they select relevant factors and present those [68]. In these cases, the explanation may not be accurate and may be influenced by their mental state. However, components can provide an objective explanation, but they are based on the explanation algorithms that have been deployed in the component. Because the cause is relative to many complex factors that shape the organization, sometimes it is difficult to design explanation algorithms that would cover all these factors.

## 6 Conclusion

We discussed different perspectives in explaining organizations in MAS. We analyzed the facets that can be explained in organizations in MAS and the different views in explaining the organization based on the organization's grid in MAS. Finally, we analyzed the possible ways of explaining organizations, considering the possible explanation questions, explainers, and explainees.

All these perspectives push into an intensive research direction on explaining organizations and aid in identifying the means to achieve *explainable organizations*. In systems without explanation, all the reasons and context underlying organizational decisions remain opaque; humans and agents are not able to request and receive explanations. By adding explanation capabilities, we not only support trust, understanding, traceability, transparency, and other desirable properties provided by explainability, but also enhance coordination, collaboration, and governance in the organizations. Moreover, explanations of the organization can be integrated within several research lines in MAS organization, supporting, for instance, accountability [9], reorganization [53], regulation management [111], and regulation adaptation [110].

As future work, we aim to propose and devise *explainability of MAS*, taking into account the explanation of individual *agents*, their *interactions*, the *environment*, and the *organization*, as we discussed in [109].

**Acknowledgments.** This study is partially funded by ANR-FAPESP NAIMAN project (ANR-22-CE23-0018-01, FAPESP 2022/03454-1) and "European Lighthouse to Manifest Trustworthy and Green AI" (ENFIELD) from the European Union's Horizon Europe research and innovation program under grant agreement No. 101120657. The authors are also members of the UNBIAS team, which is a component of the THUS pillar of the USP-CNRS International Research Center.

## References

1. Abbas, H.A., Shaheen, S.I., Amin, M.H.: Organization of multi-agent systems: An overview. International Journal of Intelligent Information Systems **4**(3), 46–57 (2015). https://doi.org/10.11648/j.ijiis.20150403.11
2. Abdul, A., Vermeulen, J., Wang, D., Lim, B.Y., Kankanhalli, M.: Trends and trajectories for explainable, accountable and intelligible systems: An HCI research agenda. In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems. p. 1–18. CHI'18, Association for Computing Machinery, New York, NY, USA (2018). https://doi.org/10.1145/3173574.3174156
3. Abiteboul, S., Bourhis, P., Vianu, V.: Explanations and transparency in collaborative workflows. In: Proceedings of the 37th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems. p. 409–424. PODS'18, Association for Computing Machinery, New York, NY, USA (2018). https://doi.org/10.1145/3196959.3196975
4. Ahlbrecht, T.: An algorithmic debugging approach for belief-desire-intention agents. Annals of Mathematics and Artificial Intelligence **92**(4), 797–814 (May 2023). https://doi.org/10.1007/s10472-023-09843-4

5. Aldewereld, H., Álvarez-Napagao, S., Dignum, V., Jiang, J., Vasconcelos, W., Vázquez-Salceda, J.: OperA/ALIVE/OperettA, pp. 173–196. Springer International Publishing, Cham (2016). https://doi.org/10.1007/978-3-319-33570-4_9
6. Alelaimat, A., Ghose, A., Dam, H.K.: Mining and validating belief-based agent explanations. In: Calvaresi, D., Najjar, A., Omicini, A., Aydogan, R., Carli, R., Ciatto, G., Mualla, Y., Främling, K. (eds.) Explainable and Transparent AI and Multi-Agent Systems. pp. 3–17. Springer Nature Switzerland, Cham (2023). https://doi.org/10.1007/978-3-031-40878-6_1
7. Anjomshoae, S., Najjar, A., Calvaresi, D., Främling, K.: Explainable agents and robots: Results from a systematic literature review. In: Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems. p. 1078–1088. AAMAS'19, International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC (2019)
8. Baldoni, M., Baroglio, C., Micalizio, R., Tedeschi, S.: Is explanation the real key factor for innovation? In: Musto, C., Magazzeni, D., Ruggieri, S., Semeraro, G. (eds.) Proceedings of the Italian Workshop on Explainable Artificial Intelligence co-located with 19th International Conference of the Italian Association for Artificial Intelligence, XAI.it@AIxIA 2020, Online Event, November 25-26, 2020. CEUR Workshop Proceedings, vol. 2742, pp. 87–95. CEUR-WS.org (2020), https://ceur-ws.org/Vol-2742/short2.pdf
9. Baldoni, M., Baroglio, C., Micalizio, R., Tedeschi, S.: Accountability in multi-agent organizations: from conceptual design to agent programming. Autonomous Agents and Multi-Agent Systems **37**(1) (Nov 2022). https://doi.org/10.1007/s10458-022-09590-6
10. Baxter, G., Sommerville, I.: Socio-technical systems: From design methods to systems engineering. Interacting with Computers **23**(1), 4–17 (08 2010). https://doi.org/10.1016/j.intcom.2010.07.003
11. Beaumont, K., Yan, E., Burattini, S., Collier, R.: Engineering inter-agent explainability in BDI agents. In: International Workshop on EXplainable, Trustworthy, and Responsible AI and Multi-Agent Systems - International Workshop, EXTRAAMAS 2025, Detroit, Michigan, USA, May 20, 2025 (2025)
12. Bench-Capon, T., Modgil, S.: Norms and value based reasoning: justifying compliance and violation. Artif. Intell. Law **25**(1), 29–64 (Mar 2017). https://doi.org/10.1007/s10506-017-9194-9
13. Bergenti, F., Gleizes, M.P., Zambonelli, F.: Methodologies and software engineering for agent systems: the agent-oriented software engineering handbook, vol. 11. Springer Science & Business Media (2006)
14. Biran, O., Cotton, C.: Explanation and justification in machine learning: A survey. In: IJCAI-17 workshop on explainable AI (XAI). vol. 8, pp. 8–13 (2017)
15. Boella, G., van der Torre, L.W.N.: Regulative and constitutive norms in normative multiagent systems. In: Dubois, D., Welty, C.A., Williams, M. (eds.) Principles of Knowledge Representation and Reasoning: Proceedings of the Ninth International Conference (KR2004), Whistler, Canada, June 2-5, 2004. pp. 255–266. AAAI Press (2004)
16. Boissier, O., Hübner, J.F., Sichman, J.S.: Organization oriented programming: From closed to open organizations. In: O'Hare, G.M.P., Ricci, A., O'Grady, M.J., Dikenelli, O. (eds.) Engineering Societies in the Agents World VII. pp. 86–105. Springer Berlin Heidelberg, Berlin, Heidelberg (2007). https://doi.org/10.1007/978-3-540-75524-1_5

17. Bond, A.H.: A computational model for organizations of cooperating intelligent agents. In: Proceedings of the ACM SIGOIS and IEEE CS TC-OA Conference on Office Information Systems. p. 21–30. COCS'90, Association for Computing Machinery, New York, NY, USA (1990). https://doi.org/10.1145/91474.91483
18. Bourne, H., Jenkins, M.: Organizational values: A dynamic perspective. Organization Studies **34**(4), 495–514 (2013). https://doi.org/10.1177/0170840612467155
19. Bratman, M.: Intention, Plans, and Practical Reason. Cambridge, MA: Harvard University Press, Cambridge (1987)
20. Broekens, J., Harbers, M., Hindriks, K., Bosch, K., Jonker, C., Meyer, J.j.: Do you get it? User-evaluated explainable BDI agents. In: Multiagent System Technologies: 8th German Conference, MATES 2010, Leipzig, Germany, September 27-29, 2010. Proceedings 8. vol. 6251, pp. 28–39. Springer, Berlin, Heidelberg (09 2010). https://doi.org/10.1007/978-3-642-16178-0_5
21. Burkart, N., Huber, M.F.: A survey on the explainability of supervised machine learning. Journal of Artificial Intelligence Research **70**, 245–317 (2021). https://doi.org/10.1613/jair.1.12228
22. Caglar, T., Sreedharan, S., Vered, M.: Who am I dealing with? Explaining the designer's hidden intentions. In: Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems. p. 436–444. AAMAS'25, International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC (2025)
23. Carrera, Á., Iglesias, C.A., Garijo, M.: Beast methodology: An agile testing methodology for multi-agent systems based on behaviour driven development. Information Systems Frontiers **16**(2), 169–182 (Jul 2013). https://doi.org/10.1007/s10796-013-9438-5
24. Castelfranchi, C.: Formalising the informal?: Dynamic social order, bottom-up social control, and spontaneous normative relations. Journal of Applied Logic **1**(1), 47–92 (2003). https://doi.org/10.1016/S1570-8683(03)00004-1
25. Chazette, L., Karras, O., Schneider, K.: Do end-users want explanations? Analyzing the role of explainability as an emerging aspect of non-functional requirements. In: 2019 IEEE 27th International Requirements Engineering Conference (RE). pp. 223–233 (2019). https://doi.org/10.1109/RE.2019.00032
26. Ciatto, G., Magnini, M., Buzcu, B., Aydoğan, R., Omicini, A.: A general-purpose protocol for multi-agent based explanations. In: Explainable and Transparent AI and Multi-Agent Systems: 5th International Workshop, EXTRAAMAS 2023, London, UK, May 29, 2023, Revised Selected Papers. p. 38–58. Springer-Verlag, Berlin, Heidelberg (2023). https://doi.org/10.1007/978-3-031-40878-6_3
27. Ciatto, G., Schumacher, M.I., Omicini, A., Calvaresi, D.: Agent-based explanations in AI: Towards an abstract framework. In: Calvaresi, D., Najjar, A., Winikoff, M., Främling, K. (eds.) Explainable, Transparent Autonomous Agents and Multi-Agent Systems. pp. 3–20. Springer International Publishing, Cham (2020). https://doi.org/10.1007/978-3-030-51924-7_1
28. Colquitt, J.A., Zipay, K.P.: Justice, fairness, and employee reactions. Annual Review of Organizational Psychology and Organizational Behavior **2**(1), 75–99 (Apr 2015). https://doi.org/10.1146/annurev-orgpsych-032414-111457
29. Corkill, D.D., Lesser, V.R.: The use of meta-level control for coordination in a distributed problem solving network. In: Proceedings of the Eighth International Joint Conference on Artificial Intelligence - Volume 2. p. 748–756. IJCAI'83, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (1983)

30. Dafoe, A., Bachrach, Y., Hadfield, G., Horvitz, E., Larson, K., Graepel, T.: Cooperative AI: machines must learn to find common ground. Nature **593**(7857), 33–36 (2021)
31. Demazeau, Y., Costa, A.R.: Populations and organizations in open multi-agent systems. In: Proceedings of the 1st National Symposium on Parallel and Distributed AI. pp. 1–13 (1996)
32. Dennis, L.A., Oren, N.: Explaining BDI agent behaviour through dialogue. In: Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems. p. 429–437. AAMAS'21, International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC (2021)
33. Dignum, M.V.: A model for organizational interaction: based on agents, founded in logic. Ph.D. thesis, Utrecht University (2004)
34. Dignum, V.: Relational artificial intelligence (2022), https://arxiv.org/abs/2202.07446
35. Dignum, V., Sonenberg, E., Dignum, F.: Towards dynamic organization of agent societies. In: Vouros, G. (ed.) Proceedings of Workshop on Coordination in Emergent Agent Societies (2004)
36. Drogoul, A., Corbara, B., Lalande, S.: Manta: New experimental results on the emergence of (artificial) ant societies. In: Gilbert, N., Conte, R. (eds.) Artificial societies, pp. 172–191. Routledge (2006). https://doi.org/10.4324/9780203993699-18
37. Esteva, M., Rodríguez-Aguilar, J.A., Sierra, C., Garcia, P., Arcos, J.L.: On the formal specifications of electronic institutions. In: Dignum, F., Sierra, C. (eds.) Agent Mediated Electronic Commerce, The European AgentLink Perspective. Lecture Notes in Computer Science, vol. 1991, pp. 126–147. Springer (2001). https://doi.org/10.1007/3-540-44682-6_8
38. Ferber, J., Gutknecht, O.: A meta-model for the analysis and design of organizations in multi-agent systems. In: Proceedings International Conference on Multi Agent Systems (Cat. No.98EX160). pp. 128–135 (1998). https://doi.org/10.1109/ICMAS.1998.699041
39. Fox, M.S.: An organizational view of distributed systems. In: Bond, A.H., Gasser, L. (eds.) Readings in Distributed Artificial Intelligence, pp. 140–150. Morgan Kaufmann (1988). https://doi.org/10.1016/B978-0-934613-63-7.50015-2
40. Fritz, J.M.H., Arnett, R.C., Conkel, M.: Organizational ethical standards and organizational commitment. Journal of Business Ethics **20**(4), 289–299 (Jul 1999). https://doi.org/10.1023/a:1005939325707
41. García, E., Valero, S., Giret, A.: ROMAS-Magentix2, pp. 153–171. Springer International Publishing, Cham (2016). https://doi.org/10.1007/978-3-319-33570-4_8
42. Gasser, L.: Perspectives on Organizations in Multi-agent Systems, pp. 1–16. Springer Berlin Heidelberg, Berlin, Heidelberg (2001). https://doi.org/10.1007/3-540-47745-4_1
43. Gasser, L., Rouquette, N.F., Hill, R.W., Lieb, J.: Representing and using organizational knowledge in distributed ai systems. In: Gasser, L., Huhns, M.N. (eds.) Distributed Artificial Intelligence, pp. 55–78. Morgan Kaufmann, San Francisco (CA) (1989). https://doi.org/10.1016/B978-1-55860-092-8.50007-1
44. Grimmelikhuijsen, S.G., Meijer, A.J.: Effects of transparency on the perceived trustworthiness of a government organization: Evidence from an online experiment. Journal of Public Administration Research and Theory **24**(1), 137–157 (11 2012). https://doi.org/10.1093/jopart/mus048

45. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., Pedreschi, D.: A survey of methods for explaining black box models. ACM Comput. Surv. **51**(5) (Aug 2018). https://doi.org/10.1145/3236009
46. Harbers, M., van den Bosch, K., Meyer, J.J.: Design and evaluation of explainable BDI agents. In: Huang, J.X., Ghorbani, A.A., Hacid, M., Yamaguchi, T. (eds.) Proceedings of the 2010 IEEE/WIC/ACM International Conference on Intelligent Agent Technology, IAT 2010, Toronto, Canada, August 31 - September 3, 2010. pp. 125–132. IEEE Computer Society Press, Toronto, Canada (2010). https://doi.org/10.1109/WI-IAT.2010.115
47. Harbers, M., Bradshaw, J.M., Johnson, M., Feltovich, P., van den Bosch, K., Meyer, J.J.: Explanation in human-agent teamwork. In: Cranefield, S., van Riemsdijk, M.B., Vázquez-Salceda, J., Noriega, P. (eds.) Coordination, Organizations, Institutions, and Norms in Agent System VII. pp. 21–37. Springer Berlin Heidelberg, Berlin, Heidelberg (2012). https://doi.org/10.1007/978-3-642-35545-5_2
48. Hayes-Roth, B.: BB1: An architecture for blackboard systems that control, explain, and learn about their own behavior. Stanford University (1984)
49. Héder, M.: Explainable AI: A brief history of the concept. ERCIM News pp. 9–10 (2023)
50. Hindriks, K.V.: Debugging is explaining. In: Rahwan, I., Wobcke, W., Sen, S., Sugawara, T. (eds.) PRIMA 2012: Principles and Practice of Multi-Agent Systems. pp. 31–45. Springer Berlin Heidelberg, Berlin, Heidelberg (2012). https://doi.org/10.1007/978-3-642-32729-2_3
51. Hollander, C.D., Wu, A.S.: The current state of normative agent-based systems. Journal of Artificial Societies and Social Simulation **14**(2) (2011). https://doi.org/10.18564/JASSS.1750
52. Hübner, J.F., Sichman, J.S., Boissier, O.: A model for the structural, functional, and deontic specification of organizations in multiagent systems. In: Bittencourt, G., Ramalho, G.L. (eds.) Advances in Artificial Intelligence, 16th Brazilian Symposium on Artificial Intelligence, SBIA 2002, Porto de Galinhas/Recife, Brazil, November 11-14, 2002, Proceedings. Lecture Notes in Computer Science, vol. 2507, pp. 118–128. Springer (2002). https://doi.org/10.1007/3-540-36127-8_12
53. Hübner, J.F., Sichman, J.S., Boissier, O.: Using the Moise+ for a cooperative framework of MAS reorganisation. In: Bazzan, A.L.C., Labidi, S. (eds.) Advances in Artificial Intelligence - SBIA 2004, 17th Brazilian Symposium on Artificial Intelligence, São Luis, Maranhão, Brazil, September 29 - October 1, 2004, Proceedings. Lecture Notes in Computer Science, vol. 3171, pp. 506–515. Springer (2004). https://doi.org/10.1007/978-3-540-28645-5_51
54. Jennings, N.R.: Commitments and conventions: The foundation of coordination in multi-agent systems. The Knowledge Engineering Review **8**(3), 223–250 (1993). https://doi.org/10.1017/S0269888900000205
55. Kass, A., Leake, D.: Types of explanations. Tech. rep. (1987)
56. Köhl, M.A., Baum, K., Langer, M., Oster, D., Speith, T., Bohlender, D.: Explainability as a non-functional requirement. In: 2019 IEEE 27th International Requirements Engineering Conference (RE). pp. 363–368 (2019). https://doi.org/10.1109/RE.2019.00046
57. Langer, M., Oster, D., Speith, T., Hermanns, H., Kästner, L., Schmidt, E., Sesing, A., Baum, K.: What do we want from explainable artificial intelligence (XAI)? – A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. Artificial Intelligence **296**, 103473 (2021). https://doi.org/10.1016/j.artint.2021.103473

58. Langley, P.: Explainable, normative, and justified agency. AAAI'19/IAAI'19/EAAI'19, AAAI Press (2019). https://doi.org/10.1609/aaai.v33i01.33019775
59. Langley, P., Meadows, B., Sridharan, M., Choi, D.: Explainable Agency for Intelligent Autonomous Systems. In: Singh, S., Markovitch, S. (eds.) Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017. pp. 4762–4764. AAAI Press, San Francisco, California, USA (2017). https://doi.org/10.1609/aaai.v31i2.19108
60. Larman, C.: Agile and iterative development: a manager's guide. Addison-Wesley Professional (2004)
61. Lemaître, C., Excelente, C.B.: Multi-agent organization approach. In: Proceedings of II Iberoamerican Workshop on DAI and MAS. pp. 7–16. Toledo (1998)
62. Letia, I.A., Goron, A.: Towards justifying norm compliance. In: Cranefield, S., van Riemsdijk, M.B., Vázquez-Salceda, J., Noriega, P. (eds.) Coordination, Organizations, Institutions, and Norms in Agent System VII. pp. 110–128. Springer Berlin Heidelberg, Berlin, Heidelberg (2012). https://doi.org/10.1007/978-3-642-35545-5_7
63. Li, X., Wang, S., Zeng, S., Wu, Y., Yang, Y.: A survey on LLM-based multi-agent systems: workflow, infrastructure, and challenges. Vicinagearth **1**(1) (Oct 2024). https://doi.org/10.1007/s44336-024-00009-2
64. Lim, B.Y., Dey, A.K., Avrahami, D.: Why and why not explanations improve the intelligibility of context-aware intelligent systems. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. p. 2119–2128. CHI'09, Association for Computing Machinery, New York, NY, USA (2009). https://doi.org/10.1145/1518701.1519023
65. Mahmoud, M., Ahmad, M., Yusoff, M., Mustapha, A.: A review of norms and normative multiagent systems. The Scientific World Journal **2014**, 684587 (07 2014). https://doi.org/10.1155/2014/684587
66. Mauri, M., Minor, M.: Towards explainable BDI agents for end-users. In: Engineering Multi-Agent Systems - 13th International Workshop, EMAS 2025, Detroit, Michigan, USA, May 20-21, 2025 (2025)
67. McAuley, J., Duberley, J., Johnson, P.: Organization theory: Challenges and perspectives (2007)
68. Miller, T.: Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence **267**, 1–38 (2019). https://doi.org/10.1016/j.artint.2018.07.007
69. Mittelstadt, B., Russell, C., Wachter, S.: Explaining explanations in ai. In: Proceedings of the Conference on Fairness, Accountability, and Transparency. p. 279–288. FAT*'19, Association for Computing Machinery, New York, NY, USA (2019). https://doi.org/10.1145/3287560.3287574
70. Molnar, C.: Interpretable machine learning. Lulu. com (2020)
71. Morveli-Espinoza, M., Nieves, J.C., Tacla, C.A., Jasinski, H.M.R.: An argumentation-based approach for goal reasoning and explanations generation. Journal of Logic and Computation **33**(5), 984–1021 (09 2022). https://doi.org/10.1093/logcom/exac052
72. Mueller, S.T., Hoffman, R.R., Clancey, W.J., Emrey, A., Klein, G.: Explanation in Human-AI Systems: A literature meta-review, synopsis of key ideas and publications, and bibliography for Explainable AI. ArXiv **abs/1902.01876** (2019)
73. Nardin, L.G., Balke-Visser, T., Ajmeri, N., Kalia, A.K., Sichman, J.S., Singh, M.P.: Classifying sanctions and designing a conceptual sanctioning process model

for socio-technical systems. The Knowledge Engineering Review **31**(2), 142–166 (2016). https://doi.org/10.1017/S0269888916000023
74. Noriega, P., de Jonge, D.: Electronic Institutions: The EI/EIDE Framework, pp. 47–76. Springer International Publishing, Cham (2016). https://doi.org/10.1007/978-3-319-33570-4_4
75. O'Laughlin, M.J., Malle, B.F.: How people explain actions performed by groups and individuals. Journal of personality and social psychology **82**(1), 33 (2002). https://doi.org/10.1037//0022-3514.82.1.33
76. Omicini, A.: Not just for humans: explanation for agent-to-agent communication. In: CEUR Workshop Proceedings. vol. 2776, pp. 1–11. Sun SITE Central Europe, RWTH Aachen University (2020)
77. Padget, J., ElDeen Elakehal, E., Li, T., De Vos, M.: InstAL: An Institutional Action Language, pp. 101–124. Springer International Publishing, Cham (2016). https://doi.org/10.1007/978-3-319-33570-4_6
78. Panisson, A.R., Engelmann, D.C., Bordini, R.H.: Engineering explainable agents: An argumentation-based approach. In: Alechina, N., Baldoni, M., Logan, B. (eds.) Engineering Multi-Agent Systems. pp. 273–291. Springer International Publishing, Cham (2022). https://doi.org/10.1007/978-3-030-97457-2_16
79. Pattison, H.E., Corkill, D.D., Lesser, V.R.: Instantiating descriptions of organizational structures. Distributed artificial intelligence **1**, 59–96 (1987)
80. Pavón, J., Gómez-Sanz, J.J.: Agent oriented software engineering with INGENIAS. In: Marík, V., Müller, J.P., Pechoucek, M. (eds.) Multi-Agent Systems and Applications III, 3rd International Central and Eastern European Conference on Multi-Agent Systems, CEEMAS 2003, Prague, Czech Republic, June 16-18, 2003, Proceedings. Lecture Notes in Computer Science, vol. 2691, pp. 394–403. Springer (2003). https://doi.org/10.1007/3-540-45023-8_38
81. Peczenik, A.: On Law and Reason. Springer Verlag, Dordrecht, Netherland (1989)
82. Pedreschi, D., Pappalardo, L., Ferragina, E., Baeza-Yates, R., Barabási, A.L., Dignum, F., Dignum, V., Eliassi-Rad, T., Giannotti, F., Kertész, J., Knott, A., Ioannidis, Y., Lukowicz, P., Passarella, A., Pentland, A.S., Shawe-Taylor, J., Vespignani, A.: Human-ai coevolution. Artificial Intelligence **339**, 104244 (2025). https://doi.org/10.1016/j.artint.2024.104244
83. Pieters, W.: Explanation and trust: what to tell the user in security and AI? Ethics and Inf. Technol. **13**(1), 53–64 (Mar 2011). https://doi.org/10.1007/s10676-010-9253-3
84. Pynadath, D.V., Tambe, M., Chauvat, N., Cavedon, L.: Toward team-oriented programming. In: Jennings, N.R., Lespérance, Y. (eds.) Intelligent Agents VI, Agent Theories, Architectures, and Languages (ATAL), 6th International Workshop, ATAL'99, Orlando, Florida, USA, July 15-17, 1999, Proceedings. Lecture Notes in Computer Science, vol. 1757, pp. 233–247. Springer (1999). https://doi.org/10.1007/10719619_17
85. Rao, A.S., Georgeff, M.P.: BDI agents: From theory to practice. In: Lesser, V.R., Gasser, L. (eds.) Proceedings of the First International Conference on Multiagent Systems, June 12-14, 1995, San Francisco, California, USA. pp. 312–319. The MIT Press (1995)
86. Retzlaff, C.O., Angerschmid, A., Saranti, A., Schneeberger, D., Röttger, R., Müller, H., Holzinger, A.: Post-hoc vs ante-hoc explanations: XAI design guidelines for data scientists. Cognitive Systems Research **86**, 101243 (2024). https://doi.org/10.1016/j.cogsys.2024.101243

87. da Rocha Costa, A.C., Dimuro, G.P.: A minimal dynamical MAS organization model. In: Dignum, V. (ed.) Handbook of Research on Multi-Agent Systems - Semantics and Dynamics of Organizational Models, pp. 419–445. IGI Global (2009). https://doi.org/10.4018/978-1-60566-256-5.CH017, https://doi.org/10.4018/978-1-60566-256-5.ch017
88. Rodriguez, S., Thangarajah, J.: Explainable agents (XAg) by design. In: Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems. pp. 2712–2716 (2024)
89. Rodriguez, S., Thangarajah, J., Davey, A.: Design patterns for explainable agents (XAg). In: Dastani, M., Sichman, J.S., Alechina, N., Dignum, V. (eds.) Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2024, Auckland, New Zealand, May 6-10, 2024. pp. 1621–1629. ACM, Richland, SC (2024)
90. Rodriguez, S., Thangarajah, J., Winikoff, M.: A behaviour-driven approach for testing requirements via user and system stories in agent systems. In: Agmon, N., An, B., Ricci, A., Yeoh, W. (eds.) Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2023, London, United Kingdom, 29 May 2023 - 2 June 2023. pp. 1182–1190. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC (2023)
91. Schminke, M., Cropanzano, R., Rupp, D.E.: Organization structure and fairness perceptions: The moderating effects of organizational level. Organizational Behavior and Human Decision Processes **89**(1), 881–905 (2002). https://doi.org/10.1016/S0749-5978(02)00034-1
92. Scott, W.G.: Organization theory: An overview and an appraisal. The Journal of the Academy of Management **4**(1), 7–26 (1961). https://doi.org/10.2307/254584
93. Searle, J.: The Construction of Social Reality. Free Press (1995)
94. Sichman, J.S.a., Conte, R., Castelfranchi, C., Demazeau, Y.: A social reasoning mechanism based on dependence networks. In: Proceedings of the 11th European Conference on Artificial Intelligence. p. 188–192. ECAI'94, John Wiley & Sons, Inc., USA (1994)
95. Simon, H.A.: On the concept of organizational goal. Administrative Science Quarterly **9**(1), 1–22 (1964). https://doi.org/10.2307/2391519
96. Singh, M.P.: A social semantics for agent communication languages. In: Dignum, F., Greaves, M. (eds.) Issues in Agent Communication. Lecture Notes in Computer Science, vol. 1916, pp. 31–45. Springer (2000). https://doi.org/10.1007/10722777_3
97. Singh, M.P.: Norms as a basis for governing sociotechnical systems. ACM Transactions on Intelligent Systems and Technology **5**(1), 21:1–21:23 (2013). https://doi.org/10.1145/2542182.2542203
98. Tsakalakis, N., Stalla-Bourdillon, S., Huynh, D., Moreau, L.: A typology of explanations to support explainability-by-design. ACM J. Responsib. Comput. **2**(1) (Feb 2025). https://doi.org/10.1145/3708504
99. Van Wijk, R., Jansen, J.J., Lyles, M.A.: Inter-and intra-organizational knowledge transfer: a meta-analytic review and assessment of its antecedents and consequences. Journal of management studies **45**(4), 830–853 (2008)
100. Vázquez-Salceda, J.: The role of norms and electronic institutions in multi-agent systems applied to complex domains. the HARMONIA framework. AI Communications **16**(3), 209–212 (2003)
101. Winikoff, M.: Debugging agent programs with why? questions. In: Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems. p. 251–259.

AAMAS'17, International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC (2017)
102. Winikoff, M.: Towards Engineering Explainable Autonomous Systems. In: Briola, D., Cardoso, R.C., Logan, B. (eds.) Engineering Multi-Agent Systems - 12th International Workshop, EMAS 2024, Auckland, New Zealand, May 6-7, 2024, Revised Selected Papers. Lecture Notes in Computer Science, vol. 15152, pp. 144–155. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-71152-7_9
103. Winikoff, M., Sidorenko, G., Dignum, V., Dignum, F.: Why bad coffee? Explaining BDI agent behaviour with valuings. Artificial Intelligence **300**, 103554 (2021). https://doi.org/10.1016/j.artint.2021.103554
104. Winograd, T., Flores, F.: Understanding computers and cognition: A new foundation for design. Intellect Ltd (1986)
105. Wooldridge, M.: An introduction to multiagent systems. John wiley & sons, USA (2009)
106. Wooldridge, M.J., Jennings, N.R., Kinny, D.: The Gaia methodology for agent-oriented analysis and design. Autonomous Agents and Multiagent Systems. **3**(3), 285–312 (2000). https://doi.org/10.1023/A:1010071910869
107. Xu, F., Uszkoreit, H., Du, Y., Fan, W., Zhao, D., Zhu, J.: Explainable AI: A brief survey on history, research areas, approaches and challenges. In: Tang, J., Kan, M.Y., Zhao, D., Li, S., Zan, H. (eds.) Natural Language Processing and Chinese Computing. pp. 563–574. Springer International Publishing, Cham (2019). https://doi.org/10.1007/978-3-030-32236-6_51
108. Yan, E., Burattini, S., Hübner, J.F., Ricci, A.: A multi-level explainability framework for engineering and understanding BDI agents. Autonomous Agents and Multiagent Systems. **39**(1), 9 (2025). https://doi.org/10.1007/S10458-025-09689-6
109. Yan, E., Burattini, S., Nardin, L.G., Fred Hübner, J., Boissier, O., Sichman, J.S., Ricci, A.: Exploiting the MAOP approach for multi-level explainability of multi-agent systems. In: 14th International Workshop on Engineering Multi-Agent Systems - EMAS 2026. Paphos, Cyprus (2026)
110. Yan, E., Nardin, L., Boissier, O., Sichman, J.: A regulation adaptation model for multi-agent systems. In: 28th European Conference on Artificial Intelligence (ECAI 2025). vol. 413, pp. 3671–3678. IOS Press (2025). https://doi.org/10.3233/FAIA251245
111. Yan, E., Nardin, L.G., Boissier, O., Sichman, J.S.: A unified view on regulation management in multi-agent systems. In: Tzeng, S.T., Dell'Anna, D., Sichman, J.S. (eds.) Coordination, Organizations, Institutions, Norms, and Ethics for Governance of Multi-Agent Systems XVIII. pp. 55–74. Springer, Cham (2026). https://doi.org/10.1007/978-3-032-17542-7_4