Causal Learning and Reasoning in Multi-Agent Systems: Goals, Issues, and Taxonomy

Abstract. *Causal reasoning* is a necessary prerequisite of *agency*, that is, the capability to act purposefully towards goals. Accordingly, *causal models* became a fundamental research topic for Artificial Intelligence (AI) agents and agentic AI systems, lately, as they provide for mathematically sound approaches to support correctness, adaptability, robustness, transparency, trustworthiness, and accountability of agents' decision making. However, most of the research agenda is focussed on *single-agent* systems, where the causal model either captures the agent inner reasoning, or its interactions with the operational environment—while assuming the single agent to be the only source of actions. Instead, the role of causal reasoning and learning in *Multi-Agent Systems* (MAS), i.e. to support both situated (agents-to-environment) and social interactions (agents-to-agents), is still under-explored. In this paper, we motivate that causal reasoning and learning in MAS is not a simple extension of the single-agent case, shed light on the many reasons why causal models – and especially *explicit* structural causal models – are necessary in MAS, propose a taxonomy to classify existing (and future) approaches to learn and exploit causal models for reasoning in MAS, and propose a MAS-centred research agenda in terms of open challenges.

Keywords: Causal learning · Causal reasoning · AI agents · Multi-agent system.

1 Introduction

In many scenarios, from robotics to IoT, the capability of developing agent systems capable of autonomously deciding how to act in the environment in a purposeful way (e.g., to achieve specific goals) is becoming increasingly important. *Causal reasoning* is a necessary prerequisite of *agency*, that is, the capability to act purposefully towards goals [2,59]. This need is irrespective of whether agents are manually *programmed* to adapt to changing environmental conditions (such as for approaches rooted in software engineering and distributed AI, aka "good old-fashioned AI") or autonomously *learn* how to act (e.g. with reinforcement learning). It is likewise present in both special-purpose agents and generalist ones, as those recently built on top of LLMs [5]. In all these cases equally, to act purposefully, agents need to know, or learn, *causal models* of their effects on the environment and other agents as well, with respect to their goal [15,53].

Accordingly, causal models became a fundamental research topic for AI agents and agentic AI systems, lately, as they provide for mathematically sound approaches to support correctness, adaptability, robustness, transparency, trust-

worthiness, and accountability of agents' decision making [56,39,33]. This is particularly evident in those approaches where agents are built through learning, such as with Reinforcement Learning (RL) (so much that causal RL is now an established field of research [3]), but is also surfacing in "good old-fashioned AI" (GOFAI) [15].

However, most of the research agenda is focussed on *single-agent* systems, where the causal model either captures the agent inner reasoning, or its interactions with the operational environment while assuming the single agent to be the only source of actions. Instead, the role of causal learning and reasoning in *Multi-Agent Systems* (MAS), i.e. to support both situated (agents-to-environment) and social interactions (agents-to-agents), is still under-explored, and literature about "causal MAS" still in its infancy [10,39,20]. There, having access specifically to *explicit*, structural causal models of how the environment responds to actions, and how other agents are influenced by actions [19], is particularly relevant. In fact, only then the agents sharing an operating environment would be able to truly achieve a correct, shared understanding of what's going on (as in distributed perception), and may hope to coordinate effectively towards the achievement of a shared or individual goal (for distributed decision making) [37]. Even in the case of non-cooperative interactions, an explicit causal model would help agents to track down opponents' influence on their own perceptions and plans [22]. Furthermore, implicit representations are more suited to System 1 type of thinking, that is mostly devoted to fast and efficient, albeit inaccurate, intuition, whereas explicit causal reasoning is usually attributed to System 2 type of thinking, that is a slower, effortful, more accurate process [6]. Also, when agents need to communicate or more broadly interact, having shared, explicit representations facilitates the task [51].

In this paper, after recalling the basic notions underpinning causal learning and reasoning in AI, and overviewing the state of the art through selected papers showcasing complementary approaches (Section 2), we motivate why causal modelling – and especially *explicit* structural causal models – is necessary in MAS (Section 3.1), shed light on the many goals it enables to achieve in MAS 3.2, propose a taxonomy to classify existing and future approaches to exploit causal models for reasoning in MAS (Section 3.3), and propose a MAS-centred research agenda in terms of open challenges (Section 4).

2 Causal Models and Software Agents

A Causal Model (CM) is a formalisation of *cause-effect relationships* between variables in a domain. It aims to capture and quantify how changes in one variable influence other variables, with the goal of enabling qualitative and quantitative *causal reasoning* over such relationships (Figure 1). Hence, a causal model can be easily understood as a Structured World Model [53]. A Structural Causal Model (SCM) is a specific type of CM that combines a Directed Acyclic Graph (DAG) with Structural Equations (SEs) [48]. In the DAG of an SCM, nodes correspond to system variables, while directed edges *qualitatively* indicate that the

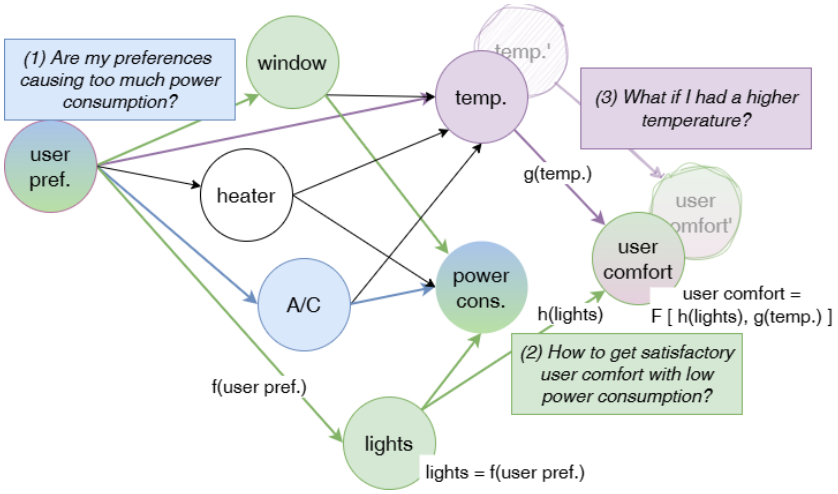

Fig. 1. Ladder of causation: *associative* thinking (in blue) only requires observation of phenomena; *interventional* thinking (in green) requires the opportunity to control (act on) variables; *counterfactual* thinking (in purple) requires imagination (i.e. planning in unobserved scenarios).

source of the edge is a cause for the sink of the edge (which becomes its effect). SEs, instead, *quantify* how each variable value is functionally determined by its "parent" variables' values. An alternative mathematical framework for causality is the *potential outcomes* (PO) framework [18]. SCMs already are a staple in the field of RL [66,56], mostly due to their operational representation of cause-effect relationships centred around the notion of *intervention* [49], i.e. a deliberate manipulation of variables that naturally maps to the agency of autonomous agents while acting in an environment. The PO framework, instead, provides an observational view meant to measure how the change in one observed variable's value affects the distribution of others'. In this paper, we stick to the SCM framework as the vast majority of the literature in AI.

SCMs fully support the three rungs of the *ladder of causation* [49]:

1. *association*, answering questions like "Is *observing* X more likely to make Y appear?", typical of *purely statistical* machine learning (that, however, is limited to this);
2. *intervention*, adding questions of the kind "If we *do* set X to x, regardless of its observed distribution, what happens to Y?", that is acting in an environment to achieve the desired goal;
3. *counterfactuals*, the highest rung, asking "Had X been x' instead of the x we observed, how would have Y changed?", that is akin to imagination, as it considers alternative observations that may have never occurred.

The notion of *intervention* is naturally aligned with the RL paradigm. In RL, an agent interacts with the environment by selecting actions that actively modify

the state of the system. In this sense, actions can be interpreted as interventions on the environment, allowing the agent to learn through active experimentation rather than mere observation. This form of learning is therefore more informative than purely associational learning based on passive observation.

SCMs can be manually designed by programmers and domain experts. However, they can also be *learnt* from data (completely, or partially). *Identifiability* conditions have been formulated [63], that define under which circumstances cause-effect relationships can be discovered (qualitatively [21]) and measured (quantitatively [64]). In addition, manual design and autonomous learning can be seamlessly *combined*. In fact, research on causal discovery and inference makes available learning algorithms for many complementary use cases. One may ignore both the DAG and its SEs, and thus aim at learning the whole SCM from scratch, by sequencing causal discovery for learning the DAG, and causal inference for learning the SEs. Or, one may know the DAG, and be interested in learning only the SEs from data. Finally, a SCM may also be partially specified and then expanded (new nodes and causal links added) by learning.

Causal models are of course *implicit* in many AI agent models, programming paradigms, and learning techniques. For instance, the deep neural networks used to learn behavioural policies in RL essentially map observations to actions by recognising the causal effects of actions on the environment [55,9]. Similar considerations may apply to the latent semantic embeddings learnt by an LLM—although whether LLMs can perform causal reasoning is still under debate [28,16]. Even a GOFAI agent designed around the BDI model [7] can be said to embed a causal model (not learnt, but programmed), relating beliefs to plans and goals to plans as cause-effect relationships. Also, literature about "causal agency" in single-agent settings is abundant, and becoming established especially in fields such as RL (Figure 2)—where causal RL research is extremely active and prolific [3,4,67,31].

However, the idea of having an *explicit* causal model available for agent exploitation, and/or inspection and manipulation, is only recently gaining increasing attention in the context of MAS. Although early attempts to merge

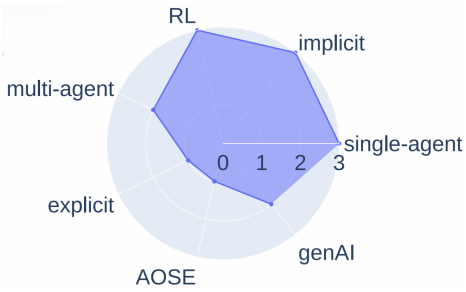

Fig. 2. Literature coverage of causal reasoning in agents and MAS according to different "settings" (AOSE stands for Agent-Oriented Software Engineering, and denotes non-learning approaches to build AI agents).

causal models and MAS date back to 2005 [38,34], when researchers proposed techniques to learn the global causal model of an environment shared amongst cooperating agents in a distributed way, most of the literature about causal models creation and exploitation in MAS started appearing in 2021.

In that year, Grimbly summarised the state of the art in causal RL and motivated the need to adopt causal modelling in multi-agent RL, too [20], also proposing research directions. However, not many approaches were available to that date, and his focus was only on agents built with RL techniques, thus did not cover traditionally programmed software agents nor the more recent wave of LLM-based ones. Then, Mariani et al. improved on the mentioned works of 2005, by proposing protocols to let agents in a MAS cooperatively discover minimal causal models of their environment to be then used for planning joint actions, where minimal means "sufficient for each agent to act on its own partial observability area" [35,36]. Hammond at al. instead focussed on studying rational outcomes of the grand coalition by translating a causal model to a multi-agent influence diagram through a game-theoretic approach that includes utility variables [22]. Parunak took quite a peculiar perspective by turning a MAS into a graphical causal model, where agents function as the update mechanism of the causal models, enhancing the representation and handling of causal influences [45]. Kerkhove at al. proposed Causal Concurrent Game Structures to focus on modelling the concept of responsibility in realising a certain collective outcome by a group of agents in a MAS [27]. Meyer-Vitali instead focused on multi-agent RL (MARL) by exploiting the synergy between interventions and counterfactuals in causal models with agents' actions and rewards in reinforcement learning. This approach can be used for collaborative planning, based on a shared causal understanding [39].

Causal MARL is indeed the most active research field trying to combine causal learning and reasoning with a multi-agent setting However, the causal models used therein are not always explicit, but are learnt implicitly in a deep neural network architecture [29,50,17,61,20,34,24]. But RL is not the only technique available to build AI autonomous agents, and optimisation of a behavioural policy not the only aspect to consider in a MAS. Hence, in the following we motivate why causal learning and reasoning in MAS deserves further attention, and propose a taxonomy to organise the existing literature and the contributions yet to come.

3 Causal Models in Multi-Agent Systems

In this section we argue why causal learning and reasoning, specifically with *explicit* structural causal models (SCM), is particularly relevant in MAS, and what goals one could achieve with it. Then we organise selected literature[1] about

[1] For instance, we filter out many papers about causal MARL not because they are not relevant, but because they would simply belong to the same category: our focus here is on shedding light on the different "angles" from which causal MAS can be studied, not providing an exhaustive review.

causal MAS in a forward-looking taxonomy aimed at providing both a reference conceptual framework for future research efforts, and a research agenda to stimulate work on open challenges.

3.1 Why

We articulate the argument for adopting explicit SCM in MAS along 4 main motivations: *partial interventions*, *interaction entanglement*, *goal-alignment*, and *human-compatible AI* (Figure 3).

Partial interventions. A first reason that makes *explicit* SCM necessary in MAS, and that calls for a dedicated conceptualisation, is that an individual agent's actions no longer constitute a "complete" intervention (as defined by Pearl [49]) on the environment, but only a *partial intervention* whose effects also depend on the actions (or lack thereof) of other agents. As a consequence, agents need to have a (at least, partially) shared understanding about cause-effect relations at play to cooperate in designing interventions for *distributed decision making* (such as for joint planning). In this sense, causal reasoning in MAS is not hypothetically beneficial, but structurally necessary. The same is true for *distributed perception*: when agents are not the only cause of change in an environment, to make sense of the situation, they need to (at least, partially) cooperate to co-build or exchange a shared causal model capturing which agent can be the cause of which effect.

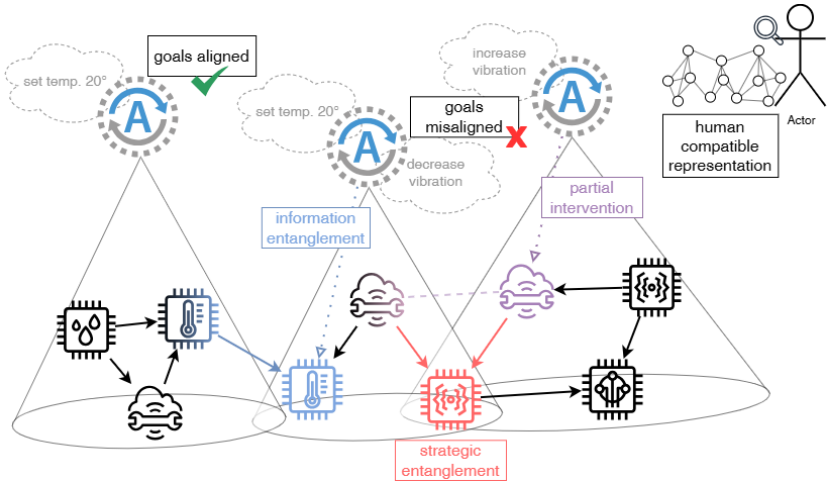

Fig. 3. Main motivations to adopt causal modeling in MAS: partial interventions (in purple) make coordination necessary, as multiple agents can control the same variables or conflicting ones. The same is true for interaction entanglement (in blue and in red) and goal alignment: for the former, agents may lack complete information to make sense of environment dynamics; for the latter, they may try to achieve conflicting goals. Explicit SCM also offer greater human compatibility compared to implicit ones.

Interaction entanglement. The word "partially" brings us to the second reason why explicit causal modelling is necessary in MAS: due to *partial observability*. Agents in MAS are typically limited in their perception and influence sphere due to geographical, network, administrative, or logical distribution—otherwise, one could argue that a MAS is not needed in the first place. Hence, agents can only build *local, individual causal models* of the portion of the shared environment under their observation range and control, and of the agents in their "social neighbourhood". However, due to partial interventions as described above, agents may need to share or exchange their partial models to effectively act and make sense of the situation. But to exchange models, such models need to be explicitly represented and to have a common semantic grounding. In summary, in a MAS causal modelling is complicated by the fact that (i) information relevant to the causal state is fragmented across agents and/or (ii) other agents constitute endogenous, strategic response mechanisms—already recognised in [62], see Chapter 10.1. We capture both effects through a unified notion, which we call *interaction entanglement*. Let us fix an agent i: the interaction is said to be entangled for agent i if the causal effect of its actions *cannot* be characterised without explicitly modelling either (i) hidden or private information (*informational entanglement*), or (ii) the response mechanisms of other agents (*strategic entanglement*).

Goal-alignment. Another reason that makes explicit SCM compelling in MAS concerns the fact that, when agents share an operating environment, the decisions of each of them depend on *goal alignment*. That is, on whether agents' goals are the same, different but compatible, mutually exclusive, and so on. In other words, depending on whether the MAS is fully cooperative (fully aligned goals), opportunistic (partially aligned), competitive (partially misaligned), adversarial (fully misaligned), or mixed, only an agent capable of modelling its interactions with others causally would be able to adapt to different alignment settings in a robust way. In fact, the notion of *robustness* in MAS – compared to a single-agent setting – naturally shifts from being mostly about uncertainty over the situated aspect (e.g. noise in perception of environmental properties and actions execution), to uncertainty over the social aspect (i.e. strategic responses of other agents)—similarly to what is done in [1] for a single agent setting. In this context, an agent equipped with an explicit SCM modelling dependencies with peers and/or opponent, in terms of how agents are affected by each other decisions, would enable the agent to seamlessly adapt to different goal alignment settings and changes thereof.

Human-compatible AI. AI systems that are compatible with human values and needs [54] consider aspects broadly related to interpretability, accountability, reliability, and more. Causal models support the necessary understanding of cause-effect relationships to make transparent and robust decisions. In a distributed setting, the causal models that are shared between agents improve their coordination and collaboration due to common world view and context. As a consequence, the trustworthiness of such systems is considerably increased. Many modern approaches to the design of autonomous AI agents and MASs feature

learning techniques (e.g. reinforcement learning, and the new breed of "agentic AI" systems) to successfully automate at least part of the complex design space. However, this creates concerns about AAs' *accountability*, *transparency*, *interpretability*, and *trustworthiness* [54]. In this regard, SCMs can greatly help, especially given that causal reasoning is common practice amongst human beings, and hence easily understandable. On the technical side, the fact that SCMs have a graphically appealing representation in their DAG strengthens this benefit.

3.2 Functional and Non-functional Goals

Similarly to the previous section, here we discuss the main functional and non-functional goals that causal reasoning and learning in MAS enable to achieve.

Coordination. From the arguments articulated in the previous section naturally arises the first functional achievement of SCM in MAS: *coordination*. The partial nature of interventions, the reciprocal dependencies amongst agents captured by the notion of interaction entanglement, and the presence (or not) of alignment between agents' goals make some form of coordination necessary [35,50]. Which form depends on a number of factors, such as whether explicit communication (e.g. message passing over a network) is possible or not (observation-based [14]), which communication means are available (e.g. visual, network, stigmergic, etc.), how agents can affect their environment, and crucially on goal alignment. In fact, depending on the cooperative or competitive (or mixed) setting, agents may want to share their causal models, or not. In competitive (especially, adversarial) settings, for instance, agents may even have incentives to share "fake" causal models, to trick others into sub-optimal behaviours. It is worth noting that with the term "coordination" we broadly mean any endeavour that involves joint activities, conflicts resolution, information exchange, so that traditional AI agents tasks such as joint planning, distributed perception, negotiation, etc. are considered.

Distributed perception. Of course, to properly coordinate and in general to bring about their tasks, agents need to make sense of the situations they experience, and of the dynamics that govern their operating environment. Here, besides causal modeling such dynamics at design-time, with expert knowledge, a great opportunity is given by causal discovery and inference algorithms [21,64]. Such algorithms enable to learn, from observational data and, possibly, interventions, both the structure of a SCM (that is, the cause-effect links among variables) and its structural equations (i.e. the precise functions that govern how variables change in response to others). For instance, this means that an AI agent may be able to autonomously and from scratch learn how to control devices in a smart home by experimenting with actuators to observe changes in sensor readings [36]. This enables such agent to effectively plan actions towards achievement of goals (e.g. keep the temperature at a comfort level while minimising energy expenditure). Besides environment dynamics, causal discovery and inference may be used to learn agents' reciprocal influences on each other.

In fact, early proposals to model game theoretic formalisms under a causal perspective started to emerge [22,19,27,26].

Robustness. Pearl himself argued that "The main role of causal models is to facilitate the evaluation of the effect of *novel* actions and policies that were unanticipated during the construction of the model" (see [47], Ch. 4). This has a natural interpretation in terms of robustness of behaviour, intended as the ability to adapt own behaviour to novel situations or to maintain correct behaviour despite disturbances. In RL especially, but not limited to that, robustness is often characterised in a similar way to generalisation and transferability of learnt policies, that is, the capability of maintaining correct behaviour outside of the training distribution, and to apply learnt behaviour to unseen tasks or in unseen environments [41,40]. More generally, for an agent to be robust means understanding the underlying dynamics of the operating environment, and the effect of own and other actions on the goal to be achieved: once these mechanisms are captured, the agent can behave correctly beyond the situations it was programmed to deal with, or learnt from.

XAI & interpretability. Causal models are considered indispensable for truly explainable AI systems because they can explain why effects occur, rather than providing only statistical probabilities and correlations [13]. Hence, accountability and trust are greatly increased. Without a causal basis, explanations can only provide analogies for potential relationships between the input and output of black boxes. Such "explanations" can be somewhat helpful but do not actually explain the reasons for the decision-making of an AI system; they merely approximate observed behaviour without revealing underlying mechanisms. It may be useful to generate hypotheses for causal relationships using traditional XAI methods to verify them using causal discovery. However, ideally, causal discovery should be performed in the first place to generate the explanations based on the discovered causal relationships. In either case, XAI for black boxes is insufficient without the actual causal understanding of the reasoning inside such systems. When causal relationships are revealed, transparency and accountability can emerge from explaining *why* effects have been found and who is responsible.

3.3 Taxonomy: what, when, and how

Given the above motivations to nurture research in causal learning and reasoning in MAS, and the focus on *explicit, multi-agent* causal models, we propose now a 3d taxonomy to categorise research efforts in the field. The dimensions of analysis that we propose are:

- **What** to model causally. In principle, one (human designer or software agent) could focus on the agent inner reasoning loop, on the environment endogenous dynamics, on the agent-environment interactions (situated), and/or on the agent-agent interactions (social). In our categorisation that follows this taxonomy, we focus on the latter 2 as the former 2 are not peculiar of MAS. In practice, this dimension considers the actual variables that would be

part of the agent causal model: e.g. which sensors, actuators, other agents, are under the sphere of observation and influence of the agent. Deciding what to model (or learn) entails issues about:

- *partial* vs. *full* observability
- which variables to consider *endogenous* (part of the modelled distributions) vs. *exogenous* (part of unobserved "noise")

– **When** to do the modeling. The choice here is basically about whether to perform causal modelling at *design-time* (that is, before actual deployment) of the agent, or at *run-time* (that is, during operation). It is worth pointing out two considerations. First, the boundary between these two stages may be blurred. In RL, for instance, the design-time is usually the stage in which the AI agent is trained in a simulated environment; there, the agent may learn a causal model, apparently, from its own perspective, at run-time while operating in the environment ("online" RL). Debating whether this has to be considered design-time or run-time is out of the scope of this taxonomy, as it depends on what the agent designer considers to be the agent deployment. Second, while it is natural to intuitively associate run-time modelling with learning, and design-time modelling with programming approaches (GOFAI) to build AI agents, this is not necessarily the case. In fact, that is our third dimension of analysis (see next bullet point): design-time modelling does not necessarily imply *expert knowledge* (could use supervised learning on past data to bootstrap a model), and run-time modelling does not necessarily imply *learning* (see autonomic and self-organising computing paradigms [52]). Deciding when to model or learn causal relationships entail issues about:

- cold-start problem in the run-time case
- continual learning complexities
- adaptability of design-time solutions to run-time contingencies

– **How** to build the causal model. The choice here is between exploiting expert, a-priori knowledge (of a human domain expert for instance) or learning it with experience (through causal discovery and inference methods). Again, this is orthogonal to the *when* dimension, although when expert knowledge is available it is likely that the causal model is built at design-time, whereas when the model is learnt it is likely to happen at run-time so as to exploit actual interventions on system variables. However, this is not necessarily and always the case, as already discussed. Another consideration has to be made about the fact that also here the boundary between the two extremes can be blurred. As we said, causal discovery and inference algorithms need not necessarily to work *from scratch* (see Section 2). This dimension entails issues about:

- what to consider prior knowledge necessary to perform learning, and the acquired knowledge that can be gathered from experience
- how to integrate symbolic approaches, typical of designs based on expert knowledge (e.g. programming the AI agent), with the sub-symbolic knowledge implicitly represented by deep learning methods within neural networks

Paper	What	When	How
[15]	Agent-Env	Run-time	Prior + Learned
[38]	Agent-Env	Run-time	Learned via discovery
[34]	Agent-Env	Run-time	Learned via discovery
[35]	Agent-Env	Run-time	Learned via discovery
[36]	Agent-Env	Run-time	Learned via discovery
[11]	Agent-Env	Run-time	Learned via discovery
[22]	Agent-Agent	Design-time	Given as prior
[27]	Agent-Agent	Design-time	Given as prior
[61]	Agent-Agent	Design-time	Learned via discovery
[17]	Agent-Agent	Run-time	Learned via discovery

Table 1. Papers utilizing explicit Structural Causal Models (SCMs) in MAS.

This way we have a 3d taxonomy simple enough to allow for an exhaustive analysis of every combination, yet expressive enough to allow for the classification of existing approaches (and, possibly, those yet to appear). Accordingly, as a conceptual validation of the taxonomy, and to highlight areas still under-explored, Tables 1 and 2 classify the relevant papers we mentioned throughout the paper (those proposing approaches to causal learning and reasoning in MAS). Table 1 restricts analysis to explicit SCM approaches in MAS, whereas Table 2 relaxes this focus.

Causal modelling of the *agent-environment* interactions at *run-time* by using *learning* approaches (hence causal discovery and/or inference) is the most widely studied setting [38,34,35,36,11]. If we relax the focus on MAS, also paper [53] can be placed in this category, as it proposes to learn the (single) agent-environment dynamics via causal discovery to adapt to distributional shifts after an RL training stage.

On the other end of the spectrum, causal modelling of the *agent-agent* interactions at *run-time* with *expert knowledge* is the only category completely missing from the literature. Following closely, is the category where *agent-environment* interactions are modelled at *design-time*, either with expert knowledge or through learning. Altogether, only 3 papers belong to the category, neither of which strictly respects the focus on explicit SCM and application within MAS. In fact, in [19] a single-agent setting is considered (although multi-agent extensions should be possible) and a causal structure is specified by a domain expert at design-time to evaluate safety and fairness incentives in the AI agent design. In [29] a generic complex system is considered, and a learnt causal structure is used at design-time to understand how it works. Finally, in [50], not SCM but Granger causality is used to capture how individual rewards relates to team rewards.

All others categories are more or less equally represented by only a handful of papers, most of which relax either the MAS or SCM focus. Here we explicitly describe only those that preserve (more or less) such focus.

Paper	What	When	How	Relaxation
[19]	Agent-Env	Design-time	Given as prior	MAS (Single agent)
[29]	System-Env	Design-time	Learned via discovery	MAS (General system)
[50]	Agent-Reward	Design-time	Learned via discovery	SCM (Granger Causal)
[45]	Agent-Env	Run-time	Given as prior	SCM (Generic GCM)
[32]	Agent-Env	Run-time	Prior + Learned	SCM and MAS
[53]	Agent-Env	Run-time	Learned via discovery	MAS (Single agent)
[26]	Agent-Agent	Design-time	Given as prior	SCM (Action Theory)
[23]	Agent-Agent	Run-time	Learned via inference	SCM (Causal metrics)

Table 2. Papers where the SCM or MAS constraints are relaxed.

The only paper modeling *agent-environment* interactions at *run-time* with a mix of prior and learned knowledge is [15], where an expertly-crafted ontology is integrated with (causal) Bayesian belief revision to refine a starting causal model with experience.

Two papers focus on *agent-agent* interactions to study them from a causal standpoint at *design-time* via SCM (or variations thereof) created with *expert knowledge*. In [22] structural causal games are introduced to model strategic agent-agent dynamics by adapting SCM. The modeler specifies the causal graph that is used at design-time as an analytical tool to study Nash equilibria. In [27] a methodology to generate a system model from an existing SCM is proposed by SCMs into Concurrent Game Structures. The goal is to analyze the concept of strategic responsibility. Only in [61] expert knowledge is replaced entirely with causal discovery that seeks to recover the unknown causal structure and underlying reward functions from data, using a Dynamic Bayesian Network to analyze how individual agent rewards causally contribute to team rewards.

Finally, in [17] *agent-agent* interactions are focussed and *learnt* at *run-time*. There, intervention sampling and conditional mutual information estimation are used to detect active causal edges, so that influence amongst agents can be learnt and exploited to steer the collective towards cooperation.

4 Challenges

The taxonomy proposed in the previous section, and in particular the coverage of existing literature of the categories defined by such a taxonomy, highlights areas of research where causal learning and reasoning in MAS are more explored (e.g. causal MARL), and others where the expected benefits discussed in Section 3 have not been fully investigated, yet. The motivations behind this may be accidental, but likely also stem from open challenges in the field. In this section, we provide clues about research directions worth pursuing to deal with these challenges and assess the extent to which causal learning and reasoning could bring benefits to MAS.

Integration with Legacy. SCMs have not being conceived as tools to design MAS in the first place. The first open issue is then about integrating with the *legacy*, intended as the models, methods, and tools already existing to design and develop MAS. For instance, AOSE methodologies [8,65,43,46], agent architectures [12], Agent-Oriented Programming (AOP) languages [57], Reinforcement Learning (RL) frameworks [60], and generative AI agents [32]. All of this tooling can be revised and adapted to support a "causality first" approach. For instance, methodologies could include the design of the required SCMs right in the analysis stage, AOP languages can be expanded with statements to build and manipulate SCMs, or some constructs be automatically mapped to SCMs equivalents.

Scaling. Currently, most of the research in causal MAS revolves around fairly limited models, with tenths of variables and causal connections at most. Especially when SCMs are learnt (fully or partially), the sheer size of the problem domain, in terms of the number of variables and their potential causal connections, as well as the number of agents involved in causal modelling, can make causal learning and reasoning computationally heavy [58,11]. Investigating ways to learn and exploit causal models in a distributed and decentralised way should be a topmost concern of causal MAS research [38,35].

Non-stationarity. The very fact that agents in a MAS have the ability of affect their shared operational environment makes non-stationary dynamics unavoidable [44]. This means that the SCM of an AI agent, modelling a specific situated or social aspect, is likely to need some periodic revision, to be continuously adapted to the ever-evolving context of operation. Conceiving efficient methods to carry out such periodic revisions is currently an open issue.

Open Systems. In a closed MAS where every agent is known since the beginning, manual design of SCMs or their learning is facilitated by such prior knowledge (although not necessarily easy). In open systems the difficulty of causal modelling is exacerbated, as any agent joining or leaving the system may impact any other, in principle. Broadly understanding the impact of having multiple agents join and leave a shared environment, possibly each with its own partial view over the global causal model of the system (often, unfeasible to centrally represent), is another relevant open issue.

Causal discovery policy. The fact that agents in a MAS can only partially intervene on system variables, generally speaking, raises the issue of coordination, as already discussed in Section 3.1. A consequence of this is that agents willing to cooperate to understand the causal model of the shared environment or of their social influence sphere must coordinate to carry out meaningful sets of interventions. This not only calls for coordination approaches to be integrated within causal discovery algorithms [36], but also for *experimental design* principles to plan highly informative interventions (i.e. actions) yielding relevant causal information, as proposed by [30]. In those cases where AI agents are built via RL, deciding whether to use the same RL policy under training to guide causal discovery, or another one, may be a though choice worth formal and experimental analysis.

Causal model assessment. Ideally, the causal models either designed for or learnt by agents in a MAS should be evaluated for correctness before being actually used during agents operations. However, validation of a learnt causal model is far from trivial, especially considering that ground truth may be unavailable (otherwise the agent would have been probably equipped with it since the beginning) [25]. The issue, while already critical by itself, is exacerbated in MAS, where agents may have different causal models even when they share the same environment and the same peers: different agents may model the same dynamics or task in different ways, depending on their goals. Thus, simply comparing models to find differences and similarities across agents, or seeking for a "majority voting" style consensus may be not enough. Investigating these forms of consensus, or ways for agent to collectively build a shared causal model and than find agreement on that, is an exciting path ahead.

Formal account. Besides experimental results, having a formal account about all aspects of causal learning and reasoning in MAS is necessary to establish clear boundaries around what can be expected and achieved, and shat not. For learning approaches, for instance, the work in [42], albeit considering a single-agent case, is an interesting first step: authors prove that, given assumptions on the "common dynamics" underlying a family of environments, perfect generalisation up to any desired error bound can be achieved via causal model-based RL. Another interesting step would be to formally define the notion of "minimal causal model" adopted in [35], informally described as the minimal causal model that captures all the variables directly controlled and observed by an agent, plus all the variables causally related to them (acquired via communication). Finally, having a rigorous definition of what are the causal dynamics "relevant" for any given scenario-task pair, especially when causal models are learnt, would be another step into a proper mathematical characterisation of causality in MAS.

Trust in agent behaviour. The shared understanding of causal models among groups of agents enables trust to emerge because they can explain the reasons for taking actions. Such transparency leads agents, as well as humans who interact with them, to experience truthful behaviour, which, in turn, results in more reliable systems and allows for improved governance. However, it is not evident how to guarantee truthfulness if individual agents might simply claim to adhere to a shared causal model, without actually complying with it in their reasoning. Conflicting goals may still lead to unfair behaviour.

5 Conclusions

In this paper, we argued that *explicit*, structural causal models are necessary for the advancement of multi-agent systems, particularly where agents must coordinate despite the availability of partial interventions and the lack of full observability. While causal learning and reasoning is maturing in single-agent AI, especially in reinforcement learning, its application to MAS requires a transition

from implicit to explicit representations to support distributed perception and decision-making.

The proposed 3D taxonomy provides a simple yet effective framework for classifying research based on the target of modeling (the *what*), the timing of the modeling process (the *when*), and the methodology used (the *how*). This analysis identifies significant research gaps, specifically in run-time agent-agent interaction modeling and the integration of expert knowledge with autonomous causal discovery. Advancing this agenda is a prerequisite also for achieving human-compatible AI in MAS. Explicit SCMs enhance transparency and accountability by providing interpretable accounts of agent behavior, which are essential for establishing trust.

Ultimately, adopting a "causality-first" approach will lead to more robust, adaptable, and trustworthy multi-agent systems.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Abate, A., Badings, T., Giacomo, G.D., Fabiano, F.: Best-effort policies for robust markov decision processes. CoRR **abs/2508.07790** (2025). https://doi.org/10.48550/ARXIV.2508.07790
2. Barandiaran, X.E., Di Paolo, E., Rohde, M.: Defining agency: Individuality, normativity, asymmetry, and spatio-temporality in action. Adaptive behavior **17**(5), 367–386 (2009)
3. Bareinboim, E., Zhang, J., Lee, S.: An Introduction to Causal Reinforcement Learning. Tech. Rep. R-65, Causal Artificial Intelligence Lab, Columbia University (Dec 2024), https://causalai.net/r65.pdf
4. Bareinboim, E., Forney, A., Pearl, J.: Bandits with unobserved confounders: A causal approach. Advances in neural information processing systems **28** (2015)
5. Belcak, P., Heinrich, G., Diao, S., Fu, Y., Dong, X., Muralidharan, S., Lin, Y.C., Molchanov, P.: Small Language Models are the Future of Agentic AI (Jun 2025). https://doi.org/10.48550/arXiv.2506.02153
6. Booch, G., Fabiano, F., Horesh, L., Kate, K., Lenchner, J., Linck, N., Loreggia, A., Murugesan, K., Mattei, N., Rossi, F., Srivastava, B.: Thinking fast and slow in AI. In: Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI. pp. 15042–15046. AAAI Press (2021). https://doi.org/10.1609/AAAI.V35I17.17765
7. Bordini, R.H., Hübner, J.F.: BDI agent programming in agentspeak using *Jason* (tutorial paper). In: Computational Logic in Multi-Agent Systems, 6th International Workshop CLIMA. Lecture Notes in Computer Science, vol. 3900, pp. 143–164. Springer (2005). https://doi.org/10.1007/11750734_9
8. Bresciani, P., Perini, A., Giorgini, P., Giunchiglia, F., Mylopoulos, J.: Tropos: An agent-oriented software development methodology. Autonomous Agents and Multi-Agent Systems **8**(3), 203–236 (2004). https://doi.org/10.1023/B:AGNT.0000018806.20944.EF
9. Briglia, G., Lippi, M., Mariani, S., Zambonelli, F.: Improving reinforcement learning-based autonomous agents with causal models. In: PRIMA: Principles and Practice of Multi-Agent Systems - 25th International Conference. Lecture

Notes in Computer Science, vol. 15395, pp. 267–283. Springer (2024). https://doi.org/10.1007/978-3-031-77367-9_20
10. Briglia, G., Mariani, S., Zambonelli, F.: A roadmap towards improving multi-agent reinforcement learning with causal discovery and inference. CoRR **abs/2503.17803** (2025). https://doi.org/10.48550/ARXIV.2503.17803
11. Briglia, G., Mariani, S., Zambonelli, F.: Towards safe action policies in multi-robot systems with causal reinforcement learning. In: Agents and Robots for reliable Engineered Autonomy. pp. 51–71. Springer Nature Switzerland (2025)
12. Bryson, J.: Cross-paradigm analysis of autonomous agent architecture. Journal of Experimental and Theoretical Artificial Intelligence **12**(2), 165–189 (2000). https://doi.org/10.1080/095281300409829
13. Carloni, G., Berti, A., Colantonio, S.: The role of causality in explainable artificial intelligence (Sep 2023). https://doi.org/10.48550/arXiv.2309.09901
14. Castelfranchi, C., Pezzulo, G., Tummolini, L.: Behavioral implicit communication (BIC): communicating with smart environments. International Journal of Ambient Computing and Intelligence **2**(1), 1–12 (2010). https://doi.org/10.4018/JACI.2010010101
15. Ceballos, H.G., Cantu, F.J.: Modelling intelligent agents through causality theory. In: 2007 Sixth Mexican International Conference on Artificial Intelligence, Special Session (MICAI). pp. 201–210 (2007). https://doi.org/10.1109/MICAI.2007.25
16. Chi, H., Li, H., Yang, W., Liu, F., Lan, L., Ren, X., Liu, T., Han, B.: Unveiling causal reasoning in large language models: Reality or mirage? In: The Thirty-eighth Annual Conference on Neural Information Processing Systems (2024), https://openreview.net/forum?id=1IU3P8VDbn
17. Du, X., Ye, Y., Zhang, P., Yang, Y., Chen, M., Wang, T.: Situation-dependent causal influence-based cooperative multi-agent reinforcement learning. In: Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence. AAAI Press (2024). https://doi.org/10.1609/aaai.v38i16.29684
18. Espinosa, V., Dasgupta, T., Rubin, D.B.: A bayesian perspective on the analysis of unreplicated factorial experiments using potential outcomes. Technometrics **58**(1), 62–73 (2016). https://doi.org/10.1080/00401706.2015.1006337
19. Everitt, T., Carey, R., Langlois, E.D., Ortega, P.A., Legg, S.: Agent incentives: A causal perspective. In: Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI. pp. 11487–11495. AAAI Press (2021). https://doi.org/10.1609/AAAI.V35I13.17368
20. Grimbly, S.J., Shock, J.P., Pretorius, A.: Causal multi-agent reinforcement learning: Review and open problems. CoRR **abs/2111.06721** (2021), https://arxiv.org/abs/2111.06721
21. Guo, R., Cheng, L., Li, J., Hahn, P.R., Liu, H.: A survey of learning causality with data: Problems and methods. ACM Computing Surveys **53**(4) (Jul 2020). https://doi.org/10.1145/3397269
22. Hammond, L., Fox, J., Everitt, T., Carey, R., Abate, A., Wooldridge, M.J.: Reasoning about causality in games. Artificial Intelligence **320**, 103919 (2023). https://doi.org/10.1016/J.ARTINT.2023.103919
23. Jaques, N., Lazaridou, A., Hughes, E., Gulcehre, C., Ortega, P., Strouse, D., Leibo, J.Z., De Freitas, N.: Social influence as intrinsic motivation for multi-agent deep reinforcement learning. In: International conference on machine learning. pp. 3040–3049. PMLR (2019)

24. Jiao, L., Wang, Y., Liu, X., Li, L., Liu, F., Ma, W., Guo, Y., Chen, P., Yang, S., Hou, B.: Causal Inference Meets Deep Learning: A Comprehensive Survey. Research **7**, 0467 (Sep 2024). https://doi.org/10.34133/research.0467
25. Karimi-Mamaghan, A.M., Tigas, P., Johansson, K.H., Gal, Y., Annadani, Y., Bauer, S.: Challenges and considerations in the evaluation of bayesian causal discovery. In: Forty-first International Conference on Machine Learning. OpenReview.net (2024)
26. Karimian, M.H., Khan, S.M., Lespérance, Y.: On the semantics of actual causality in situation calculus concurrent game structures. In: 38th Canadian Conference on Artificial Intelligence, Canadian AI. Canadian Artificial Intelligence Association (2025)
27. Kerkhove, S.S., Alechina, N., Dastani, M.: Causes and strategies in multiagent systems. In: Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems, AAMAS. pp. 1098–1106. International Foundation for Autonomous Agents and Multiagent Systems / ACM (2025). https://doi.org/10.5555/3709347.3743630
28. Kiciman, E., Ness, R., Sharma, A., Tan, C.: Causal reasoning and large language models: Opening a new frontier for causality. Transactions on Machine Learning Research (2024), https://openreview.net/forum?id=mqoxLkX210, featured Certification
29. Kocaoglu, M., Jaber, A., Shanmugam, K., Bareinboim, E.: Characterization and learning of causal graphs with latent variables from soft interventions. In: 32nd Annual Conference on Neural Information Processing Systems. pp. 14346–14356 (2019)
30. Kocaoglu, M., Shanmugam, K., Bareinboim, E.: Experimental design for learning causal graphs with latent variables. In: 30th Annual Conference on Neural Information Processing Systems. pp. 7018–7028 (2017)
31. Li, M., Zhang, J., Bareinboim, E.: Confounding robust deep reinforcement learning: A causal approach. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems
32. Lin, B.Y., Fu, Y., Yang, K., Brahman, F., Huang, S., Bhagavatula, C., Ammanabrolu, P., Choi, Y., Ren, X.: Swiftsage: A generative agent with fast and slow thinking for complex interactive tasks. In: Annual Conference on Neural Information Processing Systems (NeurIPS) (2023)
33. Lippi, M., Mariani, S., Martinelli, M., Zambonelli, F.: Autonomous mental development at the individual and collective levels: Concept and challenges. IEEE Access **13**, 5907–5921 (2025). https://doi.org/10.1109/ACCESS.2024.3522362
34. Maes, S., Meganck, S., Manderick, B.: Inference in multi-agent causal models. International Journal of Approximate Reasoning **46**(2), 274–299 (2007). https://doi.org/https://doi.org/10.1016/j.ijar.2006.09.005, special Track on Uncertain Reasoning of the 18th International Florida Artificial Intelligence Research Symposium (FLAIRS 2005)
35. Mariani, S., Roseti, P., Zambonelli, F.: Multi-agent learning of causal networks in the internet of things. In: Advances in Practical Applications of Agents, Multi-Agent Systems, and Cognitive Mimetics. The PAAMS Collection - 21st International Conference, PAAMS. Lecture Notes in Computer Science, vol. 13955, pp. 163–174. Springer (2023). https://doi.org/10.1007/978-3-031-37616-0_14
36. Mariani, S., Zambonelli, F.: Distributed discovery of causal networks in pervasive environments. In: IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom). pp. 1–6. IEEE (2024). https://doi.org/10.1109/PERCOMWORKSHOPS59983.2024.10502971

37. Mariani, S., Zambonelli, F.: On the role of causal reasoning in autonomous agents and multi-agent systems. In: PRIMA 2025: Principles and Practice of Multi-Agent Systems - 26th International Conference. Lecture Notes in Computer Science, vol. 16366, pp. 524–531. Springer (2025). https://doi.org/10.1007/978-3-032-13562-9_40
38. Meganck, S., Maes, S., Manderick, B., Leray, P.: Distributed learning of multi-agent causal models. In: Proceedings of the 2005 IEEE/WIC/ACM International Conference on Intelligent Agent Technology. pp. 285–288. IEEE Computer Society (2005). https://doi.org/10.1109/IAT.2005.66
39. Meyer-Vitali, A.: Multi-Agent Causal Reinforcement Learning. In: Proceedings of the 13th International Conference on Model-Based Software and Systems Engineering - MBSE-AI Integration. pp. 435–442. SciTePress (2025). https://doi.org/10.5220/0013400100003896
40. Mutti, M., De Santi, R., Restelli, M., Marx, A., Ramponi, G.: Exploiting causal graph priors with posterior sampling for reinforcement learning. In: The Twelfth International Conference on Learning Representations
41. Mutti, M., De Santi, R., Rossi, E., Calderon, J.F., Bronstein, M., Restelli, M.: Provably efficient causal model-based reinforcement learning for systematic generalization. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 9251–9259 (2023)
42. Mutti, M., Santi, R.D., Rossi, E., Calderón, J.F., Bronstein, M.M., Restelli, M.: Provably efficient causal model-based reinforcement learning for systematic generalization. In: Thirty-Seventh Conference on Artificial Intelligence. pp. 9251–9259. AAAI Press (2023). https://doi.org/10.1609/AAAI.V37I8.26109
43. Omicini, A., Ricci, A., Viroli, M.: Artifacts in the a&a meta-model for multi-agent systems. Autonomous Agents and Multi-Agent Systems **17**(3), 432–456 (2008). https://doi.org/10.1007/S10458-008-9053-X
44. Papoudakis, G., Christianos, F., Rahman, A., Albrecht, S.V.: Dealing with non-stationarity in multi-agent deep reinforcement learning. CoRR **abs/1906.04737** (2019)
45. Parunak, H.V.D.: How to turn an MAS into a graphical causal model. In: Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems, AAMAS. pp. 1845–1847. ACM (2023). https://doi.org/10.5555/3545946.3598849
46. Pavón, J., Gómez-Sanz, J.J.: Agent oriented software engineering with INGENIAS. In: 3rd International Central and Eastern European Conference on Multi-Agent Systems (CEEMAS). Lecture Notes in Computer Science, vol. 2691, pp. 394–403. Springer (2003). https://doi.org/10.1007/3-540-45023-8_38
47. Pearl, J.: Causality. Cambridge university press (2009)
48. Pearl, J.: Causal inference. In: Causality: Objectives and Assessment. JMLR Proceedings, vol. 6, pp. 39–58. JMLR.org (2010)
49. Pearl, J.: The seven tools of causal inference, with reflections on machine learning. Communications of ACM **62**(3), 54–60 (2019). https://doi.org/10.1145/3241036
50. Pina, R., De Silva, V., Artaud, C.: Discovering causality for efficient cooperation in multi-agent environments. Neurocomputing **642**, 130358 (2025). https://doi.org/https://doi.org/10.1016/j.neucom.2025.130358
51. Polanyi, M.: The tacit dimension. In: Knowledge in organisations, pp. 135–146. Routledge (2009)
52. Reyd, S., Diaconescu, A., Dessalles, J., Esterle, L.: A roadmap for causality research in complex adaptive systems. In: IEEE International Conference on Autonomic

Computing and Self-Organizing Systems, ACSOS. pp. 35–40. IEEE (2024). https://doi.org/10.1109/ACSOS-C63493.2024.00027
53. Richens, J., Everitt, T.: Robust agents learn causal world models. In: The Twelfth International Conference on Learning Representations (2024), https://openreview.net/forum?id=pOoKI3ouv1
54. Russell, S.: Human-compatible artificial intelligence. In: Human-Like Machine Intelligence, pp. 3–23. Oxford University Press (2022). https://doi.org/10.1093/OSO/9780198862536.003.0001
55. Schulte, O., Poupart, P.: Why online reinforcement learning is causal. CoRR **abs/2403.04221** (2024). https://doi.org/10.48550/ARXIV.2403.04221
56. Schulte, O., Poupart, P.: When should reinforcement learning use causal reasoning? Transactions on Machine Learning Research **2025** (2025)
57. Shoham, Y.: Agent-oriented programming. Artificial Intelligence **60**(1), 51–92 (1993). https://doi.org/10.1016/0004-3702(93)90034-9
58. Tigas, P., Annadani, Y., Jesson, A., Schölkopf, B., Gal, Y., Bauer, S.: Interventions, where and how? experimental design for causal models at scale. In: Annual Conference on Neural Information Processing Systems (NeurIPS) (2022)
59. Tramacere, A.: A causal view of the sense of agency. Philosophical Psychology **35**(3), 442–465 (2022). https://doi.org/10.1080/09515089.2021.1985102
60. Tuyls, K., Weiss, G.: Multiagent learning: Basics, challenges, and prospects. AI Magazine **33**(3), 41–52 (2012). https://doi.org/10.1609/AIMAG.V33I3.2426
61. Wang, Z., Du, Y., Zhang, Y., Fang, M., Huang, B.: MACCA: offline multi-agent reinforcement learning with causal credit assignment. Transactions on Machine Learning Research **2025** (2025)
62. Weiss, G.: Multiagent systems: a modern approach to distributed artificial intelligence. MIT press (1999)
63. Xia, K., Lee, K., Bengio, Y., Bareinboim, E.: The causal-neural connection: Expressiveness, learnability, and inference. In: Advances in Neural Information Processing Systems 34. pp. 10823–10836 (2021)
64. Yao, L., Chu, Z., Li, S., Li, Y., Gao, J., Zhang, A.: A survey on causal inference. ACM Transactions on Knowlwdge Discovery from Data **15**(5) (May 2021). https://doi.org/10.1145/3444944
65. Zambonelli, F., Jennings, N.R., Wooldridge, M.J.: Developing multiagent systems: The gaia methodology. ACM Transactions on Software Engineering and Methodology **12**(3), 317–370 (2003). https://doi.org/10.1145/958961.958963
66. Zeng, Y., Cai, R., Sun, F., Huang, L., Hao, Z.: A survey on causal reinforcement learning. IEEE Transactions on Neural Networks and Learning Systems **36**(4), 5942–5962 (2025). https://doi.org/10.1109/TNNLS.2024.3403001
67. Zhang, J., Bareinboim, E.: Markov decision processes with unobserved confounders: A causal approach. Purdue AI Lab, West Lafayette, IN, USA, Tech. Rep (2016)