Strategic Responsibility Under Imperfect Information

Vahid Yazdanpanah
Department of Industrial Engineering and Business Information Systems,
University of Twente
Enschede, The Netherlands
v.yazdanpanah@utwente.nl

Mehdi Dastani
Department of Information and Computing Sciences,
Utrecht University
Utrecht, The Netherlands
m.m.dastani@uu.nl

Wojciech Jamroga
Institute of Computer Science,
Polish Academy of Sciences
Warsaw, Poland
w.jamroga@ipipan.waw.pl

Natasha Alechina
School of Computer Science,
University of Nottingham
Nottingham, UK
nza@cs.nott.ac.uk

Brian Logan
School of Computer Science,
University of Nottingham
Nottingham, UK
bsl@cs.nott.ac.uk

ABSTRACT

A central issue in the specification and verification of autonomous agents and multiagent systems is the ascription of responsibility to individual agents and groups of agents. When designing a (multi)agent system, we must specify which agents or groups of agents are responsible for bringing about a particular state of affairs. Similarly, when verifying a multiagent system, we may wish to determine the responsibility of agents or groups of agents for a particular state of affairs, and the contribution of each agent to bringing about that state of affairs. In this paper, we discuss several aspects of responsibility, including strategic ability of agents, their epistemic properties, and their relationship to the evolution of the system behavior. We introduce a formal framework for reasoning about the responsibility of individual agents and agent groups in terms of the agents' strategies and epistemic properties, and state some properties of the framework.

KEYWORDS

Responsibility in Agent Systems; Strategic Reasoning; Concurrent Game Structures; Temporal and Modal Logic.

ACM Reference Format:
Vahid Yazdanpanah, Mehdi Dastani, Wojciech Jamroga, Natasha Alechina, and Brian Logan. 2019. Strategic Responsibility Under Imperfect Information. In *Proc. of the 18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), Montreal, Canada, May 13–17, 2019,* IFAAMAS, 9 pages.

1 INTRODUCTION

A central concept in the specification and verification of autonomous systems and multiagent systems is the notion of responsibility. From a design perspective, ascribing responsibility to individual agents and groups of agent can guide the allocation of necessary and sufficient capabilities and resources to agents. This forward looking perspective is particularly useful for determining whether the assignment of responsibilities to agents and groups of agents by a multiagent organization is consistent with the agents' capabilities and resources. Similarly, when verifying a multiagent system we may wish to determine the responsibility of agents or groups of agents for a particular state of affairs, and the contribution of each agent to bringing about that state of affairs. This backward looking perspective is useful in determining which agents or groups of agents are responsible or even to blame for an undesirable state of affairs.

The concept of responsibility has been studied in philosophical literature (e.g., [7, 32]), where responsibility is analyzed in terms of, e.g., agents' abilities, knowledge, and intentions, and classified along different dimensions such as group vs. individual responsibility, forward vs. backward responsibility, and action-oriented vs. state-oriented responsibility. To capture such notions, recent contributions in artificial intelligence and multiagent systems have proposed reasoning frameworks, operational semantics, and decision support tools. For example, [8] focuses on the strategic dimension of responsibility, and defines notions for reasoning about responsibility with respect to agents' coalitional abilities; [11] focuses on the normative aspect, and argues that autonomous systems require value-aware methods to ensure that their behavior is aligned with social preferences; [5] focuses on the problem of autonomous vehicles, and shows that ensuring the security of autonomous agent systems may result in social dilemmas; and [26] focuses on the interrelation between *group* and *individual* responsibility, and highlights the complexities involved in responsibility-sharing among agents in collective decision-making scenarios.

In a multiagent setting, reasoning about (degrees of) responsibility for a (desirable or undesirable) state of affairs involves determining if (to what extent) a group of agents is or was (physically and epistemically) able to use its strategic power to influence (i.e., to ensure, avoid, or control) the occurrence of the state of affairs. Reasoning about responsibility may take place either before the occurrence of a situation (prospectively) or after it (retrospectively). The prospective form is known as *forward (looking)* responsibility and the retrospective one is called *backward (looking)* responsibility [30]. Forward responsibility is relevant when planning in multiagent systems—e.g., to ensure fault tolerance or to guarantee the feasibility of a task allocation profile. Backward responsibility is is relevant when analyzing system behavior, and can be a justification

Proc. of the 18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), N. Agmon, M. E. Taylor, E. Elkind, M. Veloso (eds.), May 13–17, 2019, Montreal, Canada.

for ascribing liability in legal systems and (in a more general sense) for sanction allocation in normative multiagent systems.

We deem a group of agents responsible for a situation if the group is or was able to avoid the situation using the strategies available to it. The distinction between a group that *is* able and one that *was* able underlies the distinction between *forward* and *backward* responsibility. Moreover, any analysis of responsibility of agents must take into account their epistemic uncertainty, as the ability of the agents to execute a strategy depends on their knowledge of the environment.

In this paper, we present a novel approach to modeling and reasoning about the responsibility of groups of agents in an imperfect information setting. Our approach uses *Concurrent Epistemic Game Structures (CEGS)* [1] as the underlying semantic machinery. CEGS allow us to integrate the *strategic* and *epistemic* dimensions of responsibility which is a fundamental requirement for ascribing responsibility [7] to agents. We show how our approach can be used to model and analyze various issues and dimensions related to responsibility, including the relation between forward and backward responsibility and the analysis of responsibility under imperfect information.

Our analysis of responsibility is compatible with the causal analysis proposed in [10], where responsibility is defined in terms the role of an agent (or group of agents) in bringing about a state of affairs. However, while most causal frameworks for modeling responsibility (e.g., in [10]) do not include an explicit notion of time, our notion of responsibility is explicitly temporal in being defined using ATL.

2 ANALYSIS AND FORMAL FRAMEWORK

In this section, we present the intuition behind our work using a running example and recall key notions from concurrent epistemic game structures that form the basis of our formal framework.

2.1 Conceptual Analysis

Imagine a scenario in which the leaders of four communities, N, E, W, and S, are invited to meet the King K. Leaders N and E have pills containing 1 and 3 grams of poison p, respectively; while W and S have pills containing 3 and 5 grams of theriacal t. 4 grams of either p or t is sufficient to kill a person, and one gram of t neutralizes the toxic effects of one gram of p (and vice versa). The effects of p and t, and their possession of the pills is common knowledge among the four community leaders. In the meeting with K, first N and E (independently) and then W and S (again independently) have access to K's cup of wine. After the meeting, the king's chamberlain discovers K dead, and, after investigating, learns that all the community leaders dropped their pills into K's cup of wine. The question is: "*who is responsible for K's death and to what extent?*" It is clear that none of the leaders acting individually has a strategy to either ensure the death of K or to avoid it. Moreover, they were each unaware of the others' actions (i.e., they were all acting under epistemic uncertainty). Hence, the responsibility for K's death should be distributed among N, E, W, and S, albeit in a reasonable manner.

We argue that the ascription of responsibility in multiagent settings should take into account the *strategic*, *temporal*, *normative*, and *epistemic* aspects of the notion of responsibility. That is, to ascribe responsibility for a state of affairs $\mathcal{S}$ (intuitively, the set of states satisfying some 'bad' property φ, such as violations of a norm) to a group of agents A, A must both be strategically and epistemically *able* to avoid $\mathcal{S}$ at a moment of time *prior* to the occurrence of $\mathcal{S}$. In other words, we say:

> *a group A is responsible for a state of affairs $\mathcal{S}$ iff $\mathcal{S}$ occurs and A was able (had a strategy) to preclude it given their knowledge.*

The first step in ascribing responsibility is to identify those groups of agents that have a strategy to avoid the death of K. For example, when W and S have an opportunity to put their theriacal t in K's wine, if W drops his 3 gram pill in the wine and S does not drop his 5 gram pill, they can avoid the death of K. They are therefore a responsible group. In contrast, the group NW is not responsible due to its lack of knowledge, and hence strategic ability, as NW's strategy to avoid K's death depends on knowing how E and S acted or will act. This is, only if it is known to them that E drops his pill in the wine, can NW avoid the death of K by not dropping their pills in the wine, as this strategy neutralizes the effects of any potential act of S. Note that E's 3 grams of p either remains non-lethal (if S does nothing) or mixes with 5 grams of t and leads to 2 non-lethal grams of t in the wine (if S drops his pill in the wine).

We further require that a responsible group contains no "excess" individuals, that is, agents whose presence/absence does not affect the preclusive ability of the group. For example, removing agent N from the the group NWS, results in the group WS which still has preclusive power with respect to king's death. Below, we present a formalization of the example and give a list of responsible groups.

In summary, to assign responsibility for a state of affairs $\mathcal{S}$ to a group of agents A, the following three conditions are necessary:

1. *$\mathcal{S}$-relevant history:* the history, i.e., a sequence of states h ending in a state in $\mathcal{S}$;
2. *A's ability to preclude $\mathcal{S}$:* that A had the potential to avoid $\mathcal{S}$ in some state on h;[1]
3. *Minimality of A:* there is no subgroup $B \subset A$ that was able to preclude $\mathcal{S}$ in any state on h.

In subsequent sections, we provide a formal account of these conditions and characterize our notion of responsibility in multiagent systems. We then show how standard game theoretic methods can be applied to distribute responsibility among members of a group according to their contribution. In the remainder of this section, we recall the formal machinery that forms the basis of our framework.

2.2 Concurrent Epistemic Game Structures

To model agent systems and analyze their strategic behavior under imperfect information, we use *Concurrent Epistemic Game Structures (CEGS)* [1] as an epistemic extension of *Concurrent Game Structures* (CGS) [3].

Concurrent Epistemic Game Structures: Formally, a *concurrent epistemic game structure* (CEGS) is a tuple $\mathcal{M} = \langle \Sigma, Q, Act, \sim_1, \ldots, \sim_n, d, o \rangle$ where: $\Sigma = \{a_1, \ldots, a_n\}$ is a finite, non-empty set

[1] Note that here, ability to preclude captures both the strategic ability of a group and their epistemic uncertainty.

of *agents*; Q is a finite, non-empty set of *states*; *Act* is a finite set of atomic *actions*; $\sim_a \subseteq Q \times Q$ is an *epistemic indistinguishability relation* for each agent $a \in \Sigma$ (we assume that $\sim_a$ is an equivalence relation, where $q \sim_a q'$ indicates that states q and q' are indistinguishable to a); function $d : \Sigma \times Q \mapsto \mathcal{P}(Act)$ specifies the sets of actions available to agents at each state (we require that the same actions be available to an agent in indistinguishable states, i.e., $d(a, q) = d(a, q')$ whenever $q \sim_a q'$); and o is a deterministic transition function that assigns the outcome state $q' = o(q, \alpha_1, \ldots, \alpha_n)$ to state q and a tuple of actions $\alpha_i \in d(i, q)$ that can be executed by Σ in q.

To represent and and reason about *strategies* and *outcomes* in agent systems with imperfect information we make use of the following auxiliary notions. (References to elements of $\mathcal{M}$ are to elements of a CEGS $\mathcal{M}$ modeling a given multiagent system, e.g., we write Q instead of Q in $\mathcal{M}$.)

Successors and Computations: For two states q and q', we say q' is a *successor* of q if there exist actions $\alpha_i \in d(i, q)$ for $i \in \{1, \ldots, n\}$ in q such that $q' = o(q, \alpha_1, \ldots, \alpha_n)$, i.e., agents in Σ can collectively guarantee in q that q' will be the next system state. A *computation* of a CEGS $\mathcal{M}$ is an infinite sequence of states $\lambda = q_0, q_1, \ldots$ such that, for all $i > 0$, we have that q_i is a successor of q_{i-1}. We refer to a computation that starts in q as a *q-computation*. For $i \in \{0, 1, \ldots\}$, we denote the i'th state in λ by $\lambda[i]$, and $\lambda[0, i]$ and $\lambda[i, \infty]$ respectively denote the finite prefix $q_0, \ldots, q_i$ and infinite suffix $q_i, q_{i+1}, \ldots$ of λ. We refer to any two arbitrary states q_i and q_{i+1} as two *consecutive* states in $\lambda[i, \infty]$. Finally, we say a finite sequence of states $q_0, \ldots, q_n$ is a *q-history* if $q_n = q$, $n \geq 1$, and for all $0 \leq i < n$ we have that q_{i+1} is a successor of q_i. We denote a q-history that starts in q_i and has n steps with $\lambda[q_i, n]$.

Strategies and Outcomes: A *memoryless imperfect information strategy*[2] for an agent $a \in \Sigma$ is a function $\zeta_a : Q \mapsto Act$ such that, for all $q \in Q$: (1) $\zeta_a(q) \in d(q, a)$, and (2) $q \sim_a q'$ implies $\zeta_a(q) = \zeta_a(q')$. For a group of agents $\Gamma \subseteq \Sigma$, a *collective strategy* $Z_\Gamma = \{\zeta_a \mid a \in \Gamma\}$ is an indexed set of strategies, one for every $a \in \Gamma$. Then, $out(q, Z_\Gamma)$ is defined as the set of potential q-computations that agents in Γ can enforce by following their corresponding strategies in Z_Γ. We extend the notion to sets of states $\omega \subseteq Q$ in the straightforward way: $out(\omega, Z_\Gamma) = \bigcup_{q' \in \omega} out(q', Z_\Gamma)$.

Uniform Strategies: A uniform strategy is one in which agents select the same actions in all states where they have the same information available to them. In particular, if agent $a \in \Sigma$ is uncertain whether the current state is q or q', then a should select the same action in q and in q'. A strategy ζ_a for agent $a \in \Sigma$ is called *uniform* if for any pair of states q, q' such that $q \sim_a q'$, $\zeta_a(q) = \zeta_a(q')$. A strategy Z_Γ is uniform if it is uniform for every $a \in \Gamma \subseteq \Sigma$. Realistic modeling of strategic ability under imperfect information requires restricting attention to uniform strategies only.

3 MODELING RESPONSIBILITY

Reasoning about responsibility may take place either before the occurrence of a situation (prospectively) or after it (retrospectively). The prospective form is known as *forward (looking)* responsibility and the retrospective one is called *backward (looking)* responsibility. Forward responsibility is relevant when planning in multiagent systems—e.g., to ensure fault tolerance or to guarantee the feasibility of a task allocation profile. Backward responsibility is is relevant when analyzing system behavior, and can be a justification for ascribing liability in legal systems and (in a more general sense) for sanction allocation in normative multiagent systems.

3.1 Forward Responsibility

The notion of forward responsibility—in the sense of [30]—has been formalized in organizational settings [20] and under perfect information [36]. Here we give a definition in our framework:

Definition 3.1 (Forward Group Responsibility). Let $\mathcal{M}$ be a CEGS, $\mathcal{S} \subseteq Q$ be a set of states, $\overline{\mathcal{S}} = Q \setminus \mathcal{S}$, and $q \in Q$ be a state. We say that $\Gamma \subseteq \Sigma$ is *forward responsible for $\mathcal{S}$ in q* iff:

(1) there is a uniform strategy for Γ, Z_Γ, such that for all states on all computations in $out(q, Z_\Gamma)$ belong to $\overline{\mathcal{S}}$, and
(2) Γ is minimal, that is, there is no $\Gamma' \subset \Gamma$ with the property formulated in clause (1).

3.2 Backward Responsibility

In this section we give a definition of backward responsibility in our framework. A group of agents A is backward responsible for the occurrence of a state of affairs $\mathcal{S}$ in a given history, if A was able to prevent the occurrence of $\mathcal{S}$ somewhere in the history. This is, we (1) reason about agent groups with *preclusive power* (in the sense of [25]) over a state of affairs and (2) reason in a *backward-looking* manner through the history (in the sense of [30]). We also take the agents' epistemic limitations into account by considering only uniform strategies.

Let $\mathcal{S}$ be a state of affairs in a multiagent system represented by CEGS $\mathcal{M}$. If the system is currently in a state $q \in \mathcal{S}$ and within the available q-history $q_0, \ldots, q_k$, there exists a state q_i (i < k) from which coalition $\Gamma \subseteq \Sigma$ has a uniform strategy to avoid $\mathcal{S}$, we say that it has backward-looking responsibility for $\mathcal{S}$. This is because the occurrence of $\mathcal{S}$ counterfactually depends on Γ's choice.[3]

Definition 3.2 (Backward Group Responsibility). Let $\mathcal{M}$ be a CEGS, $\mathcal{S}$ be a set of states, $q \in \mathcal{S}$ a state, and $\lambda[q_i, k]$ an arbitrary q-history. We say that $\Gamma \subseteq \Sigma$ is *backward responsible for $\mathcal{S}$ based on $\lambda[q_i, k]$* iff:

(1) there is a state q_j in $\lambda[q_i, k]$ such that for some uniform strategy for Γ, Z_Γ, all states on all computations in $out(q_j, Z_\Gamma)$ belong to $\overline{\mathcal{S}}$, and
(2) Γ is minimal, that is, there is no $\Gamma' \subset \Gamma$ with the property formulated in clause (1).

In the sequel, we use $\mathfrak{R}^{\mathcal{S}}_{q, \lambda[q_i, k]}$ to denote the set of all backward responsible groups for $\mathcal{S}$ based on the q-history $\lambda[q_i, k]$.

The following theorem links our two notions of backward and forward responsibility.

[2] We focus on memoryless strategies in the imperfect information setting and avoid other forms of strategy that assume the ability of agents to recall the evolution of the multiagent system, e.g., perfect recall strategies (see [9]).

[3] This approach in modeling power-based responsibility based on counterfactual dependence is related to causality-based models presented in [10]. We discuss causality-based models in Section 6.

Theorem 3.3 (Equivalence of the two notions of responsibility). *$\Gamma \subseteq \Sigma$ is backward responsible for $\mathcal{S}$ based on $\lambda[q_i, k]$ if and only if in one of the states on $\lambda[q_i, k]$, Γ is forward responsible for $\mathcal{S}$.*

Example 3.4 (Responsible Groups for K's Death). Our running example can be modeled as $\mathcal{M} = \langle \Sigma, Q, Act, \sim_N, \sim_E, \sim_W, \sim_S, d, o\rangle$ where: $\Sigma = \{N, E, W, S\}$;[4] $Q = \{q_0, \ldots, q_6\}$; $Act = \{0, 1, 3, 5\}$ where 0 represents not-releasing a pill and other non-zero integers (denoting the weight of the pill) represent releasing a pill; $d(N, q_0) = \{0, 1\}$, $d(E, q_0) = \{0, 3\}$, $d(W, q_0) = d(S, q_0) = \{0\}$, and for $i \in \{1, 2, 3, 4\}$: $d(N, q_i) = d(E, q_i) = \{0\}$, $d(W, q_i) = \{0, 3\}$, $d(S, q_i) = \{0, 5\}$; transition function o is as illustrated in Figure 1. E.g., in q_0, action profile $\langle 1, 3, 0, 0\rangle$ (respectively for $\langle N, E, W, S\rangle$) results in state q_4. Moreover, we represent the indistinguishability relations by dashed lines, e.g., states q_1 and q_2 are indistinguishable for N, W, and S.

Given the q-history $h = q_0, q_4, q_6$, we can use our notion of backward responsibility to reason about the groups responsible for $\mathcal{S} = \{q_6\}$ where the undesired property φ = "K is dead" holds. In h all agents dropped their pills in the wine (i.e., we are in q_6). In q_4, the only group with a uniform strategy to avoid q_6 is WS. This is because WS can guarantee that drinking the wine will not be fatal by playing $\langle 3, 0\rangle$. In q_0, NS, ES, and NEW are all able to avoid K's death. In the groups NS and ES, the agents should not drop their pills in the wine (i.e., N and E in q_0 and S in q_4 and all the states that he cannot distinguish from q_4). In the group NEW, agents N and W should not put their pills in the wine while E should. In total, the four groups NS, ES, WS, and NEW are responsible for $\mathcal{S}$ under the given history, as they had the strategic ability to avoid $\mathcal{S}$.

We can also ask the question: "to what extent should each agent be held individually responsible for $\mathcal{S}$?" One possible answer is that "they are *equally responsible* as they are all members of a responsible group." But it is notable that they possess pills that contain different amounts of p and t. This results in a *power imbalance* among the agents which we argue should be reflected in their "*individual*" responsibility. We address these concerns—all referring to the so called *responsibility gap* [23][5]—in Section 4.

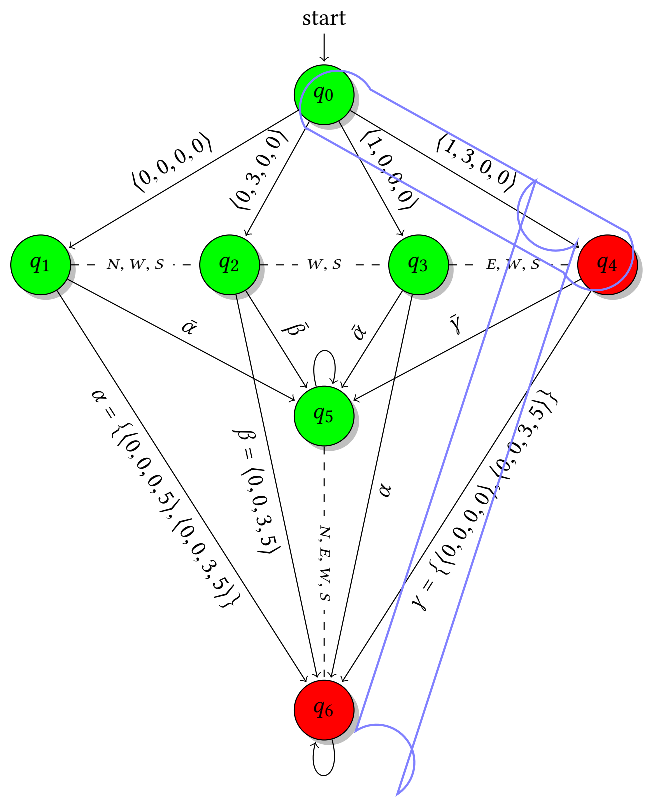

Figure 1: In q_0, agents N and E have an opportunity to drop (represented by the weight of their pill) or not drop (represented by 0) their pills in K's wine. As a result, when agents W and S have an opportunity to drop their pills, we either have a fatal glass of wine in q_4 or a non-fatal one in q_1, q_2, and q_3. Following the leaders' actions, K drinks and either survives in q_5 or is fatally poisoned in q_6. Dashed lines denote the indistinguishability relations, states in which drinking the wine is fatal are colored red, and the path outlined in blue denotes the history. To simplify the figure, we also group some action profiles and represent them as a set, e.g., in state q_4, the execution of either of the action profiles $\langle 0, 0, 0, 0\rangle$ or $\langle 0, 0, 3, 5\rangle$ denoted by γ results in state q_6, while all other action profiles denoted by $\bar{\gamma}$ lead to q_5.

4 DISTRIBUTING RESPONSIBILITY

Using our notion of responsibility, we can determine which groups of agents are responsible for a state of affairs, given a history. However, to solve the problem of many hands[6], we require a method for ascribing responsibility to individual agents. For instance, in the running example (given the history q_0, q_4, q_6), we have that NS, ES, WS, and NEW are responsible groups. But, *to what extent each community leader is responsible?* We argue that a reasonable approach to *distributing* responsibility is to ascribe responsibility to agents based on their marginal contributions to responsible groups. If strategic power is not uniformly distributed among agents, sharing the responsibility equally does not reflect the contribution of each agent. A standard approach to distributing values among agents with respect to their contribution, is the *Shapley-value* [29]. Simply stated, using the Shapley value, one can calculate the average value that the presence of an agent brings to groups that it contributes to. Here, we propose a Shapley-based method for distributing responsibility (for a state of affairs) among agents in a multiagent system.

Definition 4.1 (Responsibility Value). Let $\mathcal{M}$ be a CEGS, $\mathcal{S}$ a state of affairs, $q \in \mathcal{S}$ a state, and $\lambda[q_i, k]$ an arbitrary q-history. We define the responsibility game $\mathcal{G}^{\mathcal{S}}_{q,\lambda[q_i,k]} = (\Sigma, \varrho)$ as a cooperative game where for any coalition $\Gamma \subseteq \Sigma$, the game's characteristic function $\varrho(\Gamma) = 1$ iff a coalition $\Gamma' \subseteq \Gamma$ is q-responsible for $\mathcal{S}$ given $\lambda[q_i, k]$;

[4]Note that in our example, K is not an agent capable of performing actions within the multiagent system, but merely a part of the environment.

[5]Briefly, a responsibility gap occurs when one can ascribe responsibility at the group level but not at the level of individuals.

[6]Basically, this problem refers to situations where a group's collective action resulted in an outcome and realizing the contribution of individuals is not straightforward. See [31] for a comprehensive account of this problem.

otherwise $\varrho(\Gamma) = 0$. The q-responsibility value of agent $a \in \Sigma$ for $\mathcal{S}$ given $\lambda[q_i, k]$, denoted $\rho^{a,\mathcal{S}}_{q,\lambda[q_i,k]}$, is:

$$\rho^{a,\mathcal{S}}_{q,\lambda[q_i,k]} = \sum_{\Gamma \subseteq \Sigma \setminus \{a\}} \frac{|\Gamma|!(|\Sigma| - |\Gamma| - 1)!}{|\Sigma|!} \left(\varrho(\Gamma \cup \{a\}) - \varrho(\Gamma)\right).$$

We show below how the properties that characterize the Shapley value and the Shapley-based notion of *fairness* in economics [27, 29] are reflected in reasoning about responsibility value of (individual) agents. In principle, fairness in value distribution—in a cooperative game among a group of agents—is axiomatized by the satisfaction of: (1) *Efficiency*, that the summation of distributed values is equal to the value of the grand coalition, (2) *Symmetry*, that any two agents with symmetric contributions to the group, will receive equal individual shares, (3) *Dummy Player*, that is to give to agents who do not contribute to the group, the value that they can gain individually, and (4) *Additivity* that is, for two different cooperative games, the value distribution be such that for each individual, the summation of what she receives in each game be equal to her share in the aggregated game. We first elaborate on how these properties relate to responsibility reasoning though an example and then present the general results.

Example 4.2 (Distributing Responsibility Among Many Hands). We can formulate the responsibility game for our running example in which $\varrho(\Gamma) = 1$ for any agent group Γ that is either responsible or is a superset of a responsible group—i.e., NS, ES, WS, NEW, and their supersets. Then for agents N, E, W, and S, we have that the responsibility value is respectively equal to: 2/12, 2/12, 2/12, and 6/12. Observe that: (1) agents that possess more power (i.e., have more poison) have a larger responsibility value—*with more power comes more responsibility*, (2) the responsibility values of agents with symmetric power, i.e., E and W, are equal (Symmetry), and (3) the responsibility values of all agents sum up to 1 (Efficiency).

Observations (2) and (3) in Example 4.2 above (Symmetry and Efficiency) hold in general due to properties of the Shapley allocation.

Proposition 4.3 (Fairness Properties). *Let $\mathcal{M}$ be a CEGS, $\mathcal{S}$ a state of affairs, and $\lambda[q_i, k]$ an arbitrary q-history. If $\Re^{\mathcal{S}}_{q,\lambda[q_i,k]} \neq \emptyset$, we have:*

(1) $\sum_{a \in \Sigma} \rho^{a,\mathcal{S}}_{q,\lambda[q_i,k]} = 1$ (Efficiency);
(2) for $a_1, a_2 \in \Sigma$, $\rho^{a_1,\mathcal{S}}_{q,\lambda[q_i,k]} = \rho^{a_2,\mathcal{S}}_{q,\lambda[q_i,k]}$ if for all $\Gamma \subseteq \Sigma \setminus \{a_1, a_2\}$ we have that $\varrho(\Gamma \cup \{a_1\}) = \varrho(\Gamma \cup \{a_2\})$ (Symmetry);
(3) for $a \in \Sigma$, $\rho^{a,\mathcal{S}}_{q,\lambda[q_i,k]} = 0$ if for all $\Gamma \subseteq \Sigma \setminus \{a\}$ we have that $\varrho(\Gamma \cup \{a\}) = \varrho(\Gamma)$ (Dummy player).

Proof. All three properties directly follow from the properties of the Shapley value [22]. □

Next, we show that considering two distinguishable state of affairs, the responsibility value of any agent for the union of the two is equal to the summation of its responsibility values for each.

Proposition 4.4. *[Conditional Additivity] Let $\mathcal{M}$ be a CEGS, $\mathcal{S}$ and $\mathcal{S}'$ two states of affairs such that $\mathcal{S} \setminus \mathcal{S}' \neq \emptyset$ and $\mathcal{S}' \setminus \mathcal{S} \neq \emptyset$, $\lambda[q_i, k]$ an arbitrary q-history, and $a \in \Sigma$. If $\Re^{\mathcal{S}}_{q,\lambda[q_i,k]}, \Re^{\mathcal{S}'}_{q,\lambda[q_i,k]} \neq \emptyset$, then we have that $\rho^{a,\mathcal{S}\cup\mathcal{S}'}_{q,\lambda[q_i,k]} = \rho^{a,\mathcal{S}}_{q,\lambda[q_i,k]} + \rho^{a,\mathcal{S}'}_{q,\lambda[q_i,k]}$ (Additivity).*

Proof. (sketch) the two state of affairs $\mathcal{S}$ and $\mathcal{S}'$ correspond to two responsibility games. Then, relying on the additivity of the Shapley value, we have the additivity of our Shapley-based notion of responsibility value. Note that as a result of the non-emptiness of both $\Re^{\mathcal{S}}_{q,\lambda[q_i,k]}$ and $\Re^{\mathcal{S}'}_{q,\lambda[q_i,k]}$, we have that $q \in \mathcal{S} \cap \mathcal{S}'$. □

This property relates to ascribing responsibility for properties that are explainable in a disjunctive form. For instance, imagine a case where the reasoner is concerned with both a murder (represented by the truth of the proposition m) and a robbery (represented by the truth of the proposition r). Then we can label the set of states in which the murder took place with $\mathcal{S}$ and the set of states in which the robbery occurs with $\mathcal{S}'$. Then, according to our results, we have that the responsibility value for an agent a for $m \vee r$ is equal to the summation of her responsibilities for each. In the next section, we explain how our CEGS-based semantics can be linked to reasoning about the truth of propositions to allow systematic responsibility verification in multiagent systems.

Next, we have that if a singleton group Γ is the unique q-responsible group for $\mathcal{S}$, then such a "polarizing dictator" is the only agent with responsibility value equal to 1 while other agents receive 0.

Proposition 4.5 (Uniquely Responsible). *Let $\mathcal{M}$ be a CEGS, $\mathcal{S}$ a state of affairs, $q \in \mathcal{S}$ a state, and $\lambda[q_i, k]$ an arbitrary q-history. $\{a \in \Sigma\}$ is the unique q-responsible (singleton group) for $\mathcal{S}$ based on a q-history $\lambda[q_i, k]$ iff $\rho^{a,\mathcal{S}}_{q,\lambda[q_i,k]} = 1$ and for all $a' \in \Sigma \setminus \{a\}$, we have $\rho^{a',\mathcal{S}}_{q,\lambda[q_i,k]} = 0$.*

Proof. (sketch) except a, marginal contribution of all agents to any subgroup is zero while a's contribution to any group is equal to 1. □

5 VERIFYING RESPONSIBILITY

In this section we give a logical characterization for our notion of backward responsibility. We use a variant of Alternating Time Temporal Logic (ATL) [3] proposed in [17] that adds indistinguishably relations to explicitly specify the epistemic uncertainty of agents and hence allows reasoning about group responsibility under imperfect information.[7]

5.1 Preliminaries: ATL_{ir}

The language of ATL is built from the following components: $\Sigma = \{a_1, \ldots, a_n\}$ a set of n agents and Π a set of propositions. Formulas of the language $\mathcal{L}_{ATL}$ are defined by the following syntax, $\varphi, \psi ::=$

$$p \mid \neg\varphi \mid \varphi \wedge \psi \mid \langle\langle\Gamma\rangle\rangle \bigcirc \varphi \mid \langle\langle\Gamma\rangle\rangle \varphi \mathcal{U} \psi \mid \langle\langle A\rangle\rangle \Box \varphi$$

where $p \in \Pi$ is a proposition, and $\Gamma \subseteq \Sigma$ is a typical group of agents.

We consider the semantics of ATL under imperfect information and with memoryless strategies, which is usually denoted by

[7] To reason about backward responsibility with respect to a history, we could use ATL extended with linear past (ATL_{lp}) [14]. However ATL_{lp} does not allow modeling imperfect information settings and the model checking problem for ATL_{lp} is EXPTIME-complete, see [6] and our Theorem 5.1.

ATL_{ir} [9] (i for imperfect information, r for memoryless rather than perfect recall strategies). Recall the notions of CEGS and uniform memoryless strategies given in Section 2.2. We extend CEGS with a propositional labeling function $\pi : \Pi \to 2^Q$.

Informally, $\langle\!\langle\Gamma\rangle\!\rangle \bigcirc \varphi$ means that Γ has a collective strategy to ensure that the next state satisfies φ; $\langle\!\langle\Gamma\rangle\!\rangle \varphi \mathcal{U} \psi$ means that Γ has a collective strategy to ensure ψ while maintaining the truth of φ; and $\langle\!\langle\Gamma\rangle\!\rangle \Box \varphi$ means that Γ has a collective strategy to ensure that φ is always true. The semantics of ATL_{ir} is defined relative to a CEGS $\mathcal{M}$ and state q and is given below:

- $\mathcal{M}, q \models p$ iff $q \in \pi(p)$
- boolean cases are standard
- $\mathcal{M}, q \models \langle\!\langle\Gamma\rangle\!\rangle \bigcirc \varphi$ iff exists a strategy Z_Γ such that for all computations $\lambda \in out(q, Z_\Gamma)$, $\mathcal{M}, \lambda[1] \models \varphi$
- $\mathcal{M}, q \models \langle\!\langle\Gamma\rangle\!\rangle \varphi \mathcal{U} \psi$ iff exists a strategy Z_Γ such that for all computations $\lambda \in out(q, Z_\Gamma)$, for some i, $\mathcal{M}, \lambda[i] \models \psi$, and for all $j < i$, $\mathcal{M}, \lambda[j] \models \varphi$
- $\mathcal{M}, q \models \langle\!\langle\Gamma\rangle\!\rangle \Box \varphi$ iff exists a strategy Z_Γ such that for all computations $\lambda \in out(q, Z_\Gamma)$, for all i, $\mathcal{M}, \lambda[i] \models \varphi$.

5.2 Logical Characterization of Responsibility

We first provide a translation of our notion of state of affairs as a subset of Q in terms of a logically verifiable proposition. Given a formula φ, we denote by $[\![\varphi]\!]_{\mathcal{M}}$ the set of states in which φ holds. Then instead of writing "Γ *is responsible for* $\mathcal{S} = [\![\varphi]\!]_{\mathcal{M}}$", we simply say "$\Gamma$ *is responsible for* φ". Note that all the states in which φ holds corresponds to a fixed set of states $\mathcal{S} \subseteq Q$. In the following we show that our notions of responsibility can be formalized within ATL.[8]

Γ is responsible for φ given the q-history $\lambda[q_i, k]$ iff:

(1) $\mathcal{M}, q \models \varphi$ (*relevance of the history*) and
(2) there exists a $q' \in \lambda[q_i, k]$, s.t. $\mathcal{M}, q' \models \langle\!\langle\Gamma\rangle\!\rangle \Box \neg\varphi$ and for all $\Gamma' \subset \Gamma$, $\mathcal{M}, q' \not\models \langle\!\langle\Gamma'\rangle\!\rangle \Box \neg\varphi$ (minimality and preclusive power).

The following theorem establishes the complexity of verifying responsibility under imperfect information in a multiagent system.

Theorem 5.1 (Complexity). *Let $\mathcal{M}$ be a CEGS, q a state in $\mathcal{M}$, $\lambda[q_i, k]$ a q-history, and φ a formula of ATL_{ir}. The problem of checking whether a group Γ is responsible for φ given $\lambda[q_i, k]$ is Δ_2^P-complete w.r.t. the size of $\mathcal{M}$ and φ, and the length of $\lambda[q_i, k]$.*

Proof. We use the fact that model checking ATL_{ir} is Δ_2^P-complete [18, 28].

Upper bound: It suffices to check that (1) $\mathcal{M}, q \models \varphi$, and (2) there is $q' \in \lambda[q_i, k]$ such that $\mathcal{M}, q' \models \langle\!\langle\Gamma\rangle\!\rangle \Box \neg\varphi$ and $\mathcal{M}, q \models \neg\langle\!\langle\Gamma \setminus \{a\}\rangle\!\rangle \Box \neg\varphi$ for every agent $a \in \Gamma$. This requires $1 + k(1 + |\Gamma|)$ calls to ATL_{ir} model checking, which yields a procedure in Δ_2^P.

Lower bound: Take the reduction of SNSAT_2 in [18]. There, for every SNSAT_2 instance Θ with r assignments, one constructs a CEGS $\mathcal{M}_r$ and an ATL_{ir} formula Φ_r such that Θ returns *true* iff $\mathcal{M}_r, q_0^r \models \Phi_r$.[9] We extend the reduction as follows.

First, we construct model $\mathcal{M}'_r$ by adding a new agent i and an extra state q'_0 with two outgoing transitions fully controlled by i: one looping at q'_0, and the other proceeding to q_0^r. The new agent does not influence any transition in the rest of $\mathcal{M}'_r$. Moreover, no atomic proposition holds in the new state (in particular, propositions yes and neg do not hold there). Secondly, we observe that $\mathcal{M}'_r, q'_0 \not\models \Phi_r$, since the system can loop in q'_0, and never reach either yes or neg. Thus, we also have that $\mathcal{M}'_r, q'_0 \models \langle\!\langle i\rangle\!\rangle \Box \neg\Phi_r$. Thirdly, the only proper subset of $\{i\}$ is the empty coalition $\varnothing$, and we have that $\mathcal{M}'_r, q'_0 \models \langle\!\langle\varnothing\rangle\!\rangle \Box \neg\Phi_r$ iff $\mathcal{M}'_r, q_0^r \not\models \Phi_r$. Hence, also $\mathcal{M}'_r, q'_0 \models \neg\langle\!\langle\varnothing\rangle\!\rangle \Box \neg\Phi_r$ iff $\mathcal{M}'_r, q_0^r \models \Phi_r$. Now, take the history $h = (q'_0, q_0^r)$ in $\mathcal{M}$, and consider the responsibility of coalition $\{i\}$ for Φ_r based on h. Both conditions are now equivalent to checking if $\mathcal{M}'_r, q_0^r \models \Phi_r$. Thus, $\mathcal{M}_r, q_0^r \models \Phi_r$ iff $\{i\}$ is responsible for Φ_r based on h, which completes the reduction. $\square$

6 DISCUSSION

Following Chockler and Halpern [10] who argue that: "*We cannot say that a definition is right or wrong, but we can and should examine how useful a definition is, particularly the extent to which it captures our intuitions*", we focus on three well-known examples in the literature on responsibility, identify responsible agents and agent groups, show how our approach allows the ascription of (a degree of) responsibility to agents, and how the strategic, temporal, normative, and epistemic aspects that plat a role can be captured.

6.1 Responsibility and Blameworthiness

The Traveller and Two Enemies. This scenario (adapted from [24]) is about a traveller P who requires water to survive a trip across the desert. P has two enemies E_1 and E_2. The night before P's departure, E_1 adds poison to the water in P's canteen. Later, but before P departs, E_2 empties the (poisoned) water from the canteen. P dies of thirst in the middle of the desert. The question is: who is responsible for P's death? Given the history above, from our definition of responsibility, we have that neither E_1 nor E_2 are individually responsible for the death (see Figure 2 for the game structure).

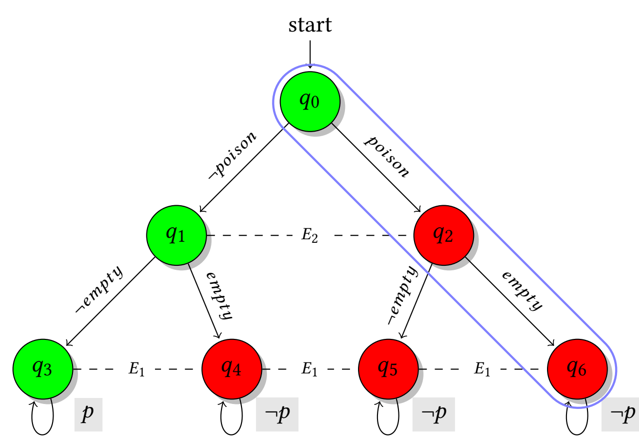

Figure 2: In q_0, E_1 may poison the water or not. In q_1 and q_2, E_2 can either empty the canteen or not. As a result, P is alive in q_3 (represented by proposition p) and dead in q_4, q_5, and q_6 (represented by $\neg p$). The path outlined in blue denotes the history.

[8] We misuse notation and say "$q \in \lambda$" to refer to a state q that *occurs* in a history λ.
[9] Due to lack of space, we must refer the reader to [18] for the details of the construction.

Given the history q_0, q_2, q_6, E_2's emptying of the canteen had no influence on P's death (and hence E_2 has no strategy to avoid it). On the other hand, E_1 could act differently in q_0 by not poisoning the water in P's canteen. However, by not poisoning the water, E_1 cannot ensure that P will not die (as E_2's may empty the canteen). E_1, E_2 is therefore the minimal group that has a uniform strategy to avoid the death of P, and due to their symmetric contributions, both E_1 and E_2 are $1/2$ responsible for the death.

In causal models (e.g. in [10, 15]), responsibility is modeled as a derived notion (basically as a degree of causality). In causality-based responsibility modeling, the direct cause of P dying of thirst is the empty canteen, which was the work of E_2. Hence, E_2 is causally responsible for P's death. (Note that the notion of responsibility in [10] applies to events rather than agents; this can be extended to agents as in [2].) E_1's degree of causal responsibility for P's death is 0 (E_1 is not part of the cause), and E_2's is 1. The notion of *blame* in [10] is closer to our notion of responsibility, since it involves an epistemic component and applies to agents. An agent may not know properties of the current state, the effects of actions, and causal relationships in general. In some of the epistemic alternatives the agent considers possible, the agent's actions have some degree of responsibility for the current outcome, in some not. The agent has a probability distribution over the alternatives. The agent's degree of blame is the expected value of the degree of responsibility. For example, if E_1 assigns probability 0 to the actual state of affairs (where P dies of thirst rather than of poison, and E_1's degree of responsibility is 0), and probability 1 to the state of affairs where P has died because of drinking poisoned water, where E_1's degree of responsibility for the death is 1, then E_1's degree of blame is 1. Chockler and Halpern [10] assign blame based on the agent's (possibly false) belief in having performed actions that caused the outcome. In contrast, our approach assigns responsibility based on having a uniform strategy to avoid the outcome. However, in both approaches E_1 can be assigned either a degree of responsibility or a degree of blame.

6.2 Agency and Strategic Responsibility

One of the widely discussed examples in the literature of responsibility is due to Frankfurt [12], and concerns scenarios in which agents perform actions that they did not intend to perform, e.g., under the influence of an implanted device in their brain, and where an undesired situation results as a consequence (see [34] for various versions of this example). For example, while under the influence/side-effects of medication agent A shoots agent B. If B dies as a result, is A responsible?

Our notion builds on agents' abilities and accounts for their preclusive power. In principle, it requires that the responsible agent should have an effective agency to influence the situation through a strategy to avoid the outcome. Hence, in Frankfurt's scenario and following our notion, A is not strategically responsible for the death of B—as A has no strategy to avoid the shooting B due to the effects of medications. This result corresponds with the perspective that sees agency as a crucial requirement for for ascribing responsibility, e.g., see [7]. Relating to the causality-based notion of Chockler and Halpern [10], the agent who performed the action that caused the death, i.e., A, is causally responsible. However, depending on A's probability distribution over effects of actions etc., A may not be to blame for the murder. Here we see that—as explained earlier—our notion of responsibility is close to what [10] defines as degree of blameworthiness.

6.3 Responsibility as Ability to Prevent vs Ability to Cause

In this scenario (taken from [10]), Billy and Suzy each throw a rock (accurately) at a bottle. Suzy throws harder, and her rock hits the bottle first so that the bottle shatters before Billy's rock reaches it. The question is: "*who is responsible for the shattered bottle?*" See Figure 3 for the CGS representation of this example.

Given the history q_0, q_4, the group bs is responsible for the shattered bottle—as it is a minimal group able to avoid it in q_0 by enforcing q_1 (in which $\neg Sh$ holds). Accordingly, the responsibility value of both Suzy (s) and Billy (b) is $1/2$ (as they have symmetric contributions to the group). Seeing the symmetric set of available actions and their potentials with respect to the bottle shattering, it is intuitive to share the responsibility equally.

In [10], where causal responsibility is considered, this formalization is referred to as a naive model. The actual cause of the bottle shattering is Suzy's rock hitting the bottle, because she threw harder. So Suzy's throw is the only cause, and it has degree of responsibility 1. If Suzy considers it equally possible that she threw harder or that Billy threw harder, then her degree of blame for the shattered bottle would be $1/2$ (because she has degree of responsibility 1 in the alternative where she threw harder, and degree of responsibility 0 in the alternative where Billy threw harder). In Figure 3 we model a situation where there is no epistemic uncertainty over action outcomes (both agents know that the result of them both throwing a rock is Suzy hitting the bottle) but still Suzy does not have a strategy to prevent the bottle shattering because if she does not throw the rock but Billy does, the bottle still shatters. In our approach, Suzy does not have degree of responsibility 1 even under perfect knowledge of her ability to throw harder.

7 RELATED WORK

Responsibility is a multidimensional concept that has been analyzed, modeled, and studied in different disciplines and from various perspectives, including *causal* responsibility [2, 10, 15], *power-based* responsibility [8, 35, 36], and *influence* responsibility (defined as an agent's ability to cause (indirect) harm or violation of social norms through other agents [21]). Moreover, responsibility can be seen as relating to a situation that has already occurred (*backward-looking* responsibility), or to whether a group of agents can bring about a particular situation (*forward-looking* responsibility) [30]. Below, we briefly compare our approach to related work in artificial intelligence, philosophy, and multiagent systems literature, and highlight the aspects of responsibility we aim to capture.

Our approach to modeling backward-looking responsibility in terms of strategic ability to influence the occurrence of a situation in a given history, is related to some approaches in modeling forward-looking coalitional responsibility, e.g., [8, 35, 36], based on the notion of preclusive power [25]. In comparison to [8, 35] where strategic ability is limited to single-step strategies, our view of

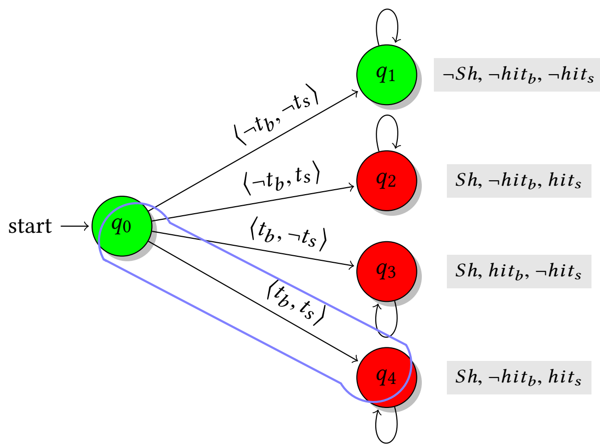

Figure 3: In q_0, Billy (b) and Suzy (s) may each throw a rock at a bottle. We denote the act of throwing by t_i for $i \in \{b, s\}$. In states q_1, q_2, q_3, q_4, the proposition Sh denotes that the bottle is shattered, and hit_i denotes that i's stone hit the bottle. The path outlined in blue denotes the history.

strategic ability is more general. Moreover, they assume perfect information, while we relax this assumption and model responsibility under imperfect information. In [36], so called *distant* responsibility is modeled using multi-step strategies, however they again conceptualize the responsibility of agents assuming perfect information.

Another strand of related work is [21], which aims to capture the "indirect" responsibility of agents for influencing the (undesired) choices of other agents. They focus on modeling the responsibility of individual agents and abstract from epistemic aspect of agents' ability to influence others. As they model the notion of influence in agent systems, their work can be seen as a basis for modeling *coalitional* responsibility in *semi-autonomous* settings—i.e., where agents are not fully autonomous in choosing among their legal actions and can be influenced by others. In contrast, in our approach, we focus on strategic ability of groups of agents to preclude the occurrence of a situation, and abstract from agents' interactions or communicative potentials.

8 CONCLUSIONS AND FUTURE WORK

In this work, we proposed a notion for reasoning about responsibility in multiagent systems under imperfect information. We used the semantic machinery of ATL_{ir}, which enables a systematic responsibility verification process, and applied value allocation methods from cooperative games, which resulted in the individual notion of responsibility value. Based on our notion of responsibility value, one can verify whether a group of agents is strategically responsible for the occurrence of a situation. In (normative) legal systems—where the legal text (e.g., the established criminal or civil law) assigns a sanction value $Sanc(S)$ to an undesirable situation S—one can apply our responsibility notion to verify the "blameworthiness" of any agent group A for S and then employ our distribution method to determine how A's members should collectively pay $Sanc(S)$.

In future work, we plan to investigate the dynamics of responsibility in richer settings, including: (1) in normative multiagent organizations by taking into account the set of organizational obligations and the nuances of verifying responsibility under (potentially) conflicting obligations, and (2) under preferences, e.g., in voting/election settings where responsibility for a *social* choice can be formulated in terms of *individual* preferences. We are also interested in studying the applicability of our responsibility notions for reasoning about related concepts such as *accountability* [4, 13] and the so called *task-/role-responsibility* [16, 30].

In this work, to verify responsibility, we reasoned from the stand point of an "omniscient" judge who is external to the system and is able to verify if a group is responsible based on her perfect knowledge (of the system states and their evolution). We aim to extend this by modeling a "non-omniscient" judge as a distinguished agent, who is internal to the system and subject to epistemic uncertainty. Specifying such a judge agent—and an indistinguishability relation for her—models the situations in which the process of responsibility verification (in addition to the execution of actions/strategies) may take place under imperfect information. Such a non-omniscient judge may reason in a *cautious* manner[10] by seeing a group Γ responsible only if in *all* the possible histories (observable by the judge), group Γ is responsible. We aim to also investigate *credulous* reasoning about responsibility—by seeing a group Γ responsible if under *some* (and not necessarily all) of the potential histories, Γ is responsible—and then study the relations among all the three modes of responsibility verification (i.e., from the point of view of an omniscient, a credulous, and a cautious judge). Different notions of responsibility may be applicable in different domains. For example, a judge who is looking for a killer may ascribe responsibility for death under cautious reasoning, whereas a physician who is looking for responsible viruses (for a death) may opt to reason credulously and start sanctioning/treating viruses in a more extensive manner (in comparison to the more narrow approach of a cautious judge).

Finally, we would like to extend our notion of responsibility value such that it can capture the "difficulty" of strategies. To see this, we highlight that in the current work, we see all the q-responsible groups for $\mathcal{S}$, equally responsible for it. For instance, we do not distinguish if the state (within the history) in which a group is able to avoid $\mathcal{S}$ is immediately before the current state q (near *past*) or is multiple states far from q (far *past*). This is to assume that a group that is able to avoid $\mathcal{S}$ is responsible for it regardless of the complexity of its available strategy. We aim to relax this assumption in future work using the so called "hardness" of strategies in [19] or the notion of responsibility distance in [36]. Then, we can distribute responsibility among agents using a responsibility index that considers the length/difficulty of available strategies.

ACKNOWLEDGMENTS

We thank the anonymous reviewers for their insightful comments. Wojciech Jamroga acknowledges the support of the National Centre

[10]Here, we use the terminology of [33] by referring to *cautious* and *credulous* reasoning. Corresponding to their classification on reasoning modes in argumentation games, we refer to responsibility ascription under all the potentially observable histories as *cautious* and under some histories as *credulous*..

for Research and Development (NCBR), Poland, under the PolLux project VoteVerif (POLLUX-IV/1/2016).

REFERENCES

[1] Thomas Ågotnes, Valentin Goranko, Wojciech Jamroga, and Michael Wooldridge. 2015. Knowledge and ability. In *Handbook of Epistemic Logic*, Hans van Ditmarsch, Joseph Halpern, Wiebe van der Hoek, and Barteld Kooi (Eds.). College Publications, 543–589.

[2] Natasha Alechina, Joseph Y. Halpern, and Brian Logan. 2017. Causality, Responsibility and Blame in Team Plans. In *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems, AAMAS 2017, São Paulo, Brazil, May 8-12, 2017*. ACM, 1091–1099.

[3] Rajeev Alur, Thomas A. Henzinger, and Orna Kupferman. 2002. Alternating-time temporal logic. *J. ACM* 49, 5 (2002), 672–713. https://doi.org/10.1145/585265.585270

[4] Matteo Baldoni, Cristina Baroglio, Olivier Boissier, Katherine Marie May, Roberto Micalizio, and Stefano Tedeschi. 2018. Accountability and Responsibility in Agent Organizations. In *International Conference on Principles and Practice of Multi-Agent Systems*. Springer, 261–278.

[5] Jean-François Bonnefon, Azim Shariff, and Iyad Rahwan. 2016. The social dilemma of autonomous vehicles. *Science* 352, 6293 (2016), 1573–1576.

[6] Laura Bozzelli, Aniello Murano, and Loredana Sorrentino. 2018. Results on alternating-time temporal logics with linear past. In *25th International Symposium on Temporal Representation and Reasoning (TIME 2018)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 6:1–6:22.

[7] Matthew Braham and Martin Van Hees. 2012. An anatomy of moral responsibility. *Mind* 121, 483 (2012), 601–634.

[8] Nils Bulling and Mehdi Dastani. 2013. Coalitional Responsibility in Strategic Settings. In *Computational Logic in Multi-Agent Systems - 14th International Workshop, CLIMA XIV, Corunna, Spain, September 16-18, 2013. Proceedings*. Springer, 172–189.

[9] N. Bulling and W. Jamroga. 2014. Comparing Variants of Strategic Ability: How Uncertainty and Memory Influence General Properties of Games. *Journal of Autonomous Agents and Multi-Agent Systems* 28, 3 (2014), 474–518.

[10] Hana Chockler and Joseph Y Halpern. 2004. Responsibility and blame: A structural-model approach. *Journal of Artificial Intelligence Research* 22 (2004), 93–115.

[11] Virginia Dignum. 2017. Responsible Autonomy. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*. ijcai.org, 4698–4704.

[12] Harry G Frankfurt. 1969. Alternate possibilities and moral responsibility. *The journal of philosophy* 66, 23 (1969), 829–839.

[13] Davide Grossi, Lambèr M. M. Royakkers, and Frank Dignum. 2007. Organizational structure and responsibility. *Artif. Intell. Law* 15, 3 (2007), 223–249.

[14] Dimitar P Guelev, Catalin Dima, and Constantin Enea. 2011. An alternating-time temporal logic with knowledge, perfect recall and past: axiomatisation and model-checking. *Journal of Applied Non-Classical Logics* 21, 1 (2011), 93–131.

[15] Joseph Y. Halpern. 2016. *Actual Causality*. The MIT Press.

[16] Herbert Lionel Adolphus Hart. 2008. *Punishment and responsibility: Essays in the philosophy of law*. Oxford University Press.

[17] Wojciech Jamroga. 2003. Some remarks on alternating temporal epistemic logic. In *Proceedings of formal approaches to multi-agent systems (FAMAS 2003)*. University of Warsaw, 133–140.

[18] Wojciech Jamroga and Jürgen Dix. 2006. Model checking abilities under incomplete information is indeed Δ_2^P-complete. *EUMAS* 6 (2006), 14–15.

[19] Wojciech Jamroga, Vadim Malvone, and Aniello Murano. 2017. Reasoning about Natural Strategic Ability. In *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems, AAMAS 2017, São Paulo, Brazil, May 8-12, 2017*. ACM, 714–722.

[20] Tiago De Lima, Lambèr M. M. Royakkers, and Frank Dignum. 2010. Modeling the problem of many hands in organisations. In *ECAI 2010 - 19th European Conference on Artificial Intelligence, Lisbon, Portugal, August 16-20, 2010, Proceedings*. IOS Press, 79–84.

[21] Emiliano Lorini and Giovanni Sartor. 2015. Influence and Responsibility: A Logical Analysis. In *Legal Knowledge and Information Systems - JURIX 2015: The Twenty-Eighth Annual Conference, Braga, Portual, December 10-11, 2015*. IOS Press, 51–60.

[22] Andreu Mas-Colell, Michael Dennis Whinston, Jerry R Green, et al. 1995. *Microeconomic theory*. Vol. 1. Oxford university press New York.

[23] Andreas Matthias. 2004. The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and information technology* 6, 3 (2004), 175–183.

[24] James Angell McLaughlin. 1925. Proximate cause. *Harvard law review* 39, 2 (1925), 149–199.

[25] Nicholas R Miller. 1981. Power in game forms. In *Power, voting, and voting power*. Springer, 33–51.

[26] Gillman Payette. 2017. Ramifications of Imposing Uniform Responsibility on Collective Action. *Logique et Analyse* 61, 243 (2017), 237–268.

[27] Alvin E. Roth and Robert E. Verrecchia. 1979. The Shapley Value As Applied to Cost Allocation: A Reinterpretation. *Journal of Accounting Research* 17, 1 (1979), 295–303.

[28] Pierre-Yves Schobbens. 2004. Alternating-time logic with imperfect recall. *Electronic Notes in Theoretical Computer Science* 85, 2 (2004), 82–93.

[29] Lloyd S Shapley. 1953. A value for n-person games. *Contributions to the Theory of Games* 2, 28 (1953), 307–317.

[30] Ibo Van de Poel. 2011. The relation between forward-looking and backward-looking responsibility. In *Moral responsibility*. Springer, 37–52.

[31] Ibo Van de Poel, Lambèr Royakkers, and Sjoerd D Zwart. 2015. *Moral responsibility and the problem of many hands*. Routledge.

[32] Nicole A Vincent, Ibo Van de Poel, and Jeroen Van Den Hoven. 2011. *Moral responsibility: beyond free will and determinism*. Vol. 27. Springer.

[33] Gerard AW Vreeswik and Henry Prakken. 2000. Credulous and sceptical argument games for preferred semantics. In *European Workshop on Logics in Artificial Intelligence*. Springer, 239–253.

[34] David Widerker. 2000. Frankfurt's attack on the principle of alternative possibilities: A further look. *Noûs* 34 (2000), 181–201.

[35] Vahid Yazdanpanah and Mehdi Dastani. 2015. Quantified Degrees of Group Responsibility. In *Coordination, Organizations, Institutions, and Normes in Agent Systems XI - COIN 2015 International Workshops, COIN@AAMAS, Istanbul, Turkey, May 4, 2015, COIN@IJCAI, Buenos Aires, Argentina, July 26, 2015, Revised Selected Papers*. Springer, 418–436.

[36] Vahid Yazdanpanah and Mehdi Dastani. 2016. Distant Group Responsibility in Multi-agent Systems. In *PRIMA 2016: Princiles and Practice of Multi-Agent Systems - 19th International Conference, Phuket, Thailand, August 22-26, 2016, Proceedings*. Springer, 261–278.